Received: 30 January 2018 | Revised: 18 April 2020 | Accepted: 23 April 2020

DOI: 10.1002/jrsm.1414

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Funnel plots may show asymmetry in the absence of publication bias with continuous outcomes dependent on baseline risk: presentation of a new publication bias test

Brett Doleman[1] | Suzanne C Freeman[2] | Jonathan N Lund[1] | John P Williams[1] | Alex J Sutton[2]

[1]Department of Surgery and Anaesthesia, Royal Derby Hospital, University of Nottingham, Derby, UK

[2]Department of Health Sciences, College of Medicine, University of Leicester, Leicester, UK

Correspondence
Brett Doleman, Division of Medical Sciences and Graduate Entry Medicine, University of Nottingham, Royal Derby Hospital, Derby, DE22 3DT, UK.
Email: dr.doleman@gmail.com

Funding information
Department of Health; National Institute for Health Research

This study aimed to determine for continuous outcomes dependent on baseline risk, whether funnel plot asymmetry may be due to statistical artefact rather than publication bias and evaluate a novel test to resolve this. Firstly, we conducted assessment for publication bias in nine meta-analyses of postoperative analgesics (344 trials with 25 348 participants). Secondly, we attempted to resolve the observed asymmetry by considering meta-regression residuals as outcome (rather than mean difference) and (inverse) sample size as the exploratory variable (rather than SE). Since the approach resolved the asymmetry, we evaluated it, and related approaches, using a simulation study considering four scenarios comprised of every combination of baseline interactions and absolute selective publication bias being present or not (10 000 simulated meta-analyses per scenario with no residual between-study heterogeneity). The test based on meta-regression residuals and inverse sample size performed as well as conventional tests (Egger's test) when no baseline risk was present and reduced type I errors when baseline risk was present. It also had modest power to detect publication bias in the presence of baseline risk. We demonstrated that correlation between effect estimates and SEs produces funnel plot asymmetry in the presence of no publication bias for continuous outcomes dependent on baseline risk. Our novel approach of assessing funnel plot asymmetry using a modified funnel plot and test based on residuals and inverse sample size may have improved performance when carrying out publication bias assessments for unstandardized mean differences where treatment effects are dependent on baseline risk.

KEYWORDS

funnel plot, publication bias

1 | INTRODUCTION

Publication bias can affect the validity of results and reduce the quality of evidence derived from meta-analyses.[1] Studies with positive findings are both more likely to be published and are published more quickly than studies with negative findings.[2] Many methods exist to help identify possible publication bias (small study effects). In fact, when referring to publication bias from this point, we are actually referring to small study effects as, strictly speaking, it is larger effects in the smaller studies—a possible symptom of publication bias—that is being considered including funnel plots and quantitative tests such as Egger's linear regression test.[3,4] Indeed, selective publication of these smaller studies with larger treatment effects and larger standard errors (as $\mathrm{SE} = \mathrm{SD}/\sqrt{N}$) is what contributes to the funnel plot asymmetry observed. Of note, other factors can contribute to small study effects such as selective outcome reporting, clinical heterogeneity or statistical artefact.

Despite a number of research studies evaluating these methods in meta-analyses with binary outcomes, little research has been conducted in other types of outcomes. In particular, little work has been conducted assessing funnel plot asymmetry for unstandardized mean differences[5] and current recommendations advise the use of traditional tests.[6] Concerns have recently been raised when funnel plots are used to help identify publication bias in meta-analyses with proportion outcomes[7] although it is as yet unknown whether similar concerns exist when using continuous outcomes.

In meta-analyses of postoperative pain where morphine consumption is used as the outcome measure, we have recently demonstrated that reductions in consumption are dependent on baseline risk, that is, the degree of pain experienced (and thus morphine consumed) over the first 24 hours of the trial.[8,9] However, these findings have implications for the use of funnel plots and regression tests when assessing funnel plots for asymmetry.[10] The underlying issue relates to the fact that, on average, studies with higher baseline risk will have larger SDs, and, if effect estimates are also dependent on baseline risk (ie, on average, if trials in patients with higher pain levels offer more potential for larger absolute reductions in pain) then this may cause correlation between mean differences (x-axis) and SEs (y-axis). Such correlation could result in funnel plot asymmetry even in the absence of publication bias, which has important implications for the interpretation of the results derived from these analyses. Further, as postoperative pain studies in general have a tendency to recruit a small number of participants this further exacerbates this issue.

Consequently, our study aimed to establish whether this correlation can influence funnel plot asymmetry in practice using data from postoperative pain trials, and then proposes and evaluates methods intended to overcome this issue. In Section 2, we present the motivating funnel plots from nine postoperative pain meta-analyses and examine whether baseline risk interactions with treatment effects could be the cause of funnel plot asymmetry. In Section 3, we outline our proposed method for assessing publication bias in mean difference outcomes when baseline risk interactions with treatment are a possibility. In Section 4, we evaluate the statistical performance of this approach and several related alternatives via a simulation study. Section 5, the discussion, concludes the paper.

2 | MOTIVATING DATASETS

2.1 | Meta-analyses of postoperative pain trials

We identified randomised controlled trials from a search strategy we have described previously.[10] We performed meta-analyses for 10 different postoperative analgesics: paracetamol, non-steroidal anti-inflammatory drugs (NSAIDS) and cyclooxygenase 2 (COX-2) inhibitors, tramadol, intravenous ketamine, alpha-2 agonists (clonidine and dexmedetomidine), gabapentin, pregabalin, lidocaine, magnesium and dexamethasone (Table 1). We extracted study data onto an electronic database including: study name, type of analgesic used and data used to calculate effect estimates. In order to minimise selective outcome reporting, where standard SDs were not reported, we estimated these from similar studies in the analysis. This is due to statistically non-significant results being less likely to be fully reported than significant results. If multiple sub-groups were reported within a study (such as different doses), we used data from the most statistically significant subgroup, as we assumed one statistically significant sub-group would increase the chances of that study being published.

2.2 | Publication bias assessment

The outcome of interest was 24-hour morphine consumption (in milligrams), a common outcome in postoperative pain trials. If alternative opioids to morphine were used, we converted these to morphine-equivalents as described previously.[10] We constructed funnel plots for each meta-analysis and allocated each trial to a sub-group depending on the different levels of baseline risk

TABLE 1 Information on included meta-analyses partly reproduced from reference 10. Nefopam data not shown due to the low number of studies (5 with 394 participants)

Analgesic meta-analysis	Studies (participants)	I^2	R^2 (control mean vs SE)	Mean difference (95% CI)
Paracetamol	25 (1812)	99%	$R^2 = 52\%$; $P < .001$	−8.18 mg (−10.57 to −6.73 mg)
NSAIDS and COX-2 inhibitors	86 (6937)	92%	$R^2 = 45\%$; $P < .001$	−11.09 mg (−12.73 to −9.45 mg)
Tramadol	11 (889)	86%	$R^2 = 23\%$; $P = .14$	−8.48 mg (−11.88 to −4.89 mg)
Ketamine	62 (4309)	95%	$R^2 = 63\%$; $P < .001$	−8.13 mg (−10.23 to −6.03 mg)
Alpha-2 agonists	33 (1930)	96%	$R^2 = 55\%$; $P < .001$	−10.7 mg (−12.38 to −9.01 mg)
Gabapentin	67 (5082)	97%	$R^2 = 60\%$; $P < .001$	−8.6 mg (−9.73 to −7.46 mg)
Pregabalin	34 (3201)	94%	$R^2 = 48\%$; $P < .001$	−8.18 mg (−9.6 to −6.76 mg)
Lidocaine	22 (1319)	80%	$R^2 = 31\%$; $P = .007$	−5.04 mg (−7.42 to −2.66 mg)
Magnesium	22 (1194)	91%	$R^2 = 5\%$; $P = .34$	−6.77 mg (−8.39 to −5.15 mg)
Dexamethasone	16 (2163)	88%	$R^2 = 18\%$; $P = .1$	−4.23 mg (−5.79 to −2.67 mg)

Abbreviations: CI, confidence interval; I^2, measure of variability in results due to between-study differences compared to sampling variance; R^2, proportion of between-study variance explained by model.

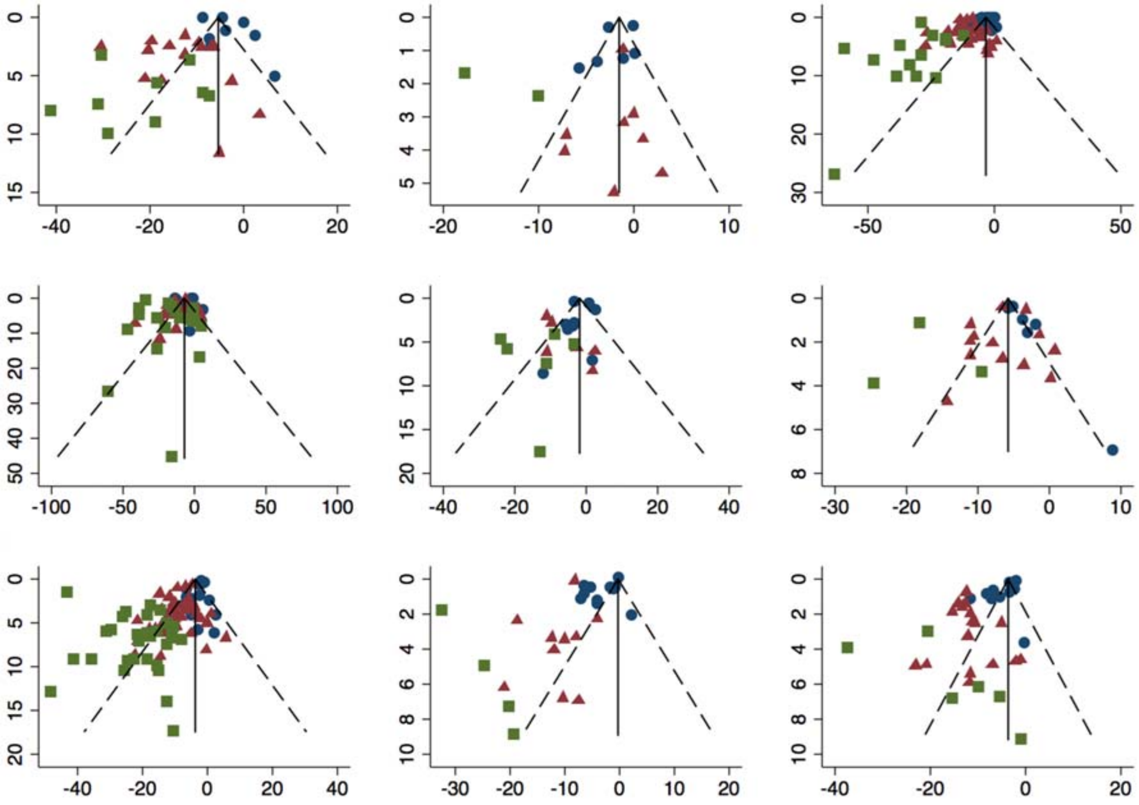

FIGURE 1 Funnel plots of postoperative analgesics. Analgesics are as follows from top left: alpha-2 agonists, dexamethasone, gabapentin, ketamine, lidocaine, magnesium, NSAIDS and COX-2 inhibitors, paracetamol and pregabalin. Tramadol not included due to the low number of studies. *X*-axis is the mean difference in morphine consumption and the *y*-axis is the SE on a reverse scale. Studies are labelled as blue dots (low baseline risk; <20 mg), red triangles (medium baseline risk; 20-50 mg) and green squares (high baseline risk; >50 mg). It can be observed that there is a relation between baseline risk with both mean differences and SEs (larger baseline risk nearer the bottom left of the plot) [Colour figure can be viewed at wileyonlinelibrary.com]

(milligrams of mean control group consumption). Groups included consumptions of <20 mg (low), 20 to 50 mg (medium) and >50 mg (high) (Figure 1). These categories were based on clinical experience rather than any empirical data and are used to visually highlight the impact of baseline risk (N.B. Baseline risk is treated as a continuous variable in all subsequent statistical analyses to ensure no loss of information or power compared to treating it as a categorical variable). We used Egger's linear regression test to assess funnel plots for asymmetry with a $P < .05$ as evidence of possible publication bias. To quantify the relation between baseline risk and SEs of the effect sizes, we also performed (unweighted) linear regression analysis using SEs as the outcome variable and baseline risk as the predictor variable. All analyses were conducted using Stata Version 14.2.[11]

2.3 | Results of publication bias assessment on postoperative analgesic trials meta-analyses

We included 344 randomised controlled trials with 25 348 participants (Table 1) although only 339 trials were included in the final analysis due to the low number of studies in one meta-analysis of nefopam.[10] Observation of funnel plots and quantitative analysis using Egger's linear regression test demonstrated asymmetric study effects for 6 out of 10 analyses (60%). These included: alpha-2 agonists ($P = .02$), gabapentin ($P < .001$), lidocaine ($P = .02$), NSAIDS and COX-2 inhibitors ($P < .001$), paracetamol ($P = .02$) and pregabalin ($P < .001$). There was less evidence of asymmetric study effects with ketamine ($P = .17$), magnesium ($P = .21$), dexamethasone ($P = .09$) and tramadol ($P = .23$).

When studies were assigned a subgroup on the basis of baseline risk (mean control group morphine consumption), this appeared to explain some or all of the asymmetry (Figure 1). That is, those studies that included patients with higher pain (and therefore higher morphine consumption) had larger reductions in morphine consumption with the intervention and also larger SEs. On linear regression analysis, baseline risk predicted SEs for alpha-2 agonists ($R^2 = 55\%$; $P < .001$), gabapentin ($R^2 = 60\%$; $P < .001$), ketamine ($R^2 = 63\%$; $P < .001$), lidocaine ($R^2 = 31\%$; $P = .007$), NSAIDS and COX-2 inhibitors ($R^2 = 45\%$; $P < .001$), paracetamol ($R^2 = 52\%$; $P < .001$) and pregabalin ($R^2 = 48\%$; $P < .001$). There was no significant relation for dexamethasone ($R^2 = 18\%$; $P = .1$), magnesium ($R^2 = 5\%$; $P = .34$) and tramadol ($R^2 = 23\%$; $P = .14$) (of note, we have previously excluded the issue of regression to the mean with frequentist meta-regression when our postoperative pain data were re-analysed using Bayesian meta-regression[10]).

The above findings support the notion that while there would appear to be concern regarding publication bias in the literature due to asymmetry of funnel plots, the asymmetry could be induced by treatment effects interacting with baseline risk and not publication bias, as explained in detail above. In the next section, we outline our novel approach to assessment of publication bias for continuous outcome meta-analysis and apply it to the motivating datasets presented in this section.

3 | MODIFIED ASSESSMENT OF FUNNEL ASYMMETRY ADJUSTING FOR BASELINE RISK USING REGRESSION RESIDUALS

3.1 | Outline of approach

Instead of plotting the observed treatment effects on the funnel plot, residuals from a meta-regression model[10] including baseline risk as a study-level covariate are instead plotted on the *x*-axis ($x_i = \text{residual}_i$). The residuals are generated from fitting a meta-regression model of the form:

$$\hat{MD}_i = \alpha + \beta\mu.\hat{c}_i + u_i + \hat{\sigma}_i\varepsilon_i, \qquad \varepsilon_i\tilde{N}(0,1); u_i\tilde{N}\left(0\tau^2\right) \quad (1)$$

where $i = 1,...., I$ indexes studies, $\hat{MD}_i$ is the observed mean difference treatment effect in the *i*th study, α and β are the intercept and slope of the regression, respectively, $\mu.\hat{c}_i$ is the observed mean baseline rate, u_i is a random effect term (assumed normally distributed with mean 0 and variance τ^2) which accounts for any remaining between study heterogeneity not explained by the regression, and $\hat{\sigma}_i\varepsilon_i$ is the random error term where $\hat{\sigma}_i^2$ is the sample estimate of $\text{Var}\left(\hat{MD}_i\right)$.[12]

In this way, the effect of baseline risk interactions with treatment effects is adjusted for prior to publication bias assessment in the belief that it will remove any artefactual asymmetry. SE of the residuals was considered for plotting on the *y*-axis (on a reverse scale) ($y_i = s.e.$ (residual_i)). The SE of the *i*th study residual is estimated by the square root of the *i*th diagonal element of the matrix defined by (I-H) × M × transpose(I-H), where M is a matrix with diagonal equal to the variance estimates of each study and 0 otherwise, H is the hat matrix from the regression model and I is the identity matrix. However, due to the concerns regarding problems associated with the correlation between baseline risk and the SE of the residuals, study sample size and inverse of the study sample size were also considered as the y-axis scale (y_i = sample. size$_i$ and y_i = 1/sample. size$_i$, respectively). As well as producing this modified funnel plot, a formal regression test based on Egger's test but using the new axes scales (ie, residuals and their SE or study sample size/inverse sample size) can be conducted for each of the three competing y-axis options:

$$\text{residual}_i = \alpha + \beta(s.e.(\text{residual}_i)) + \varepsilon_i,$$
$$\text{where } \varepsilon_i \tilde{N}\left(0, s.e.(\text{residual}_i)^2 \varphi\right), \tag{2a}$$

$$\text{residual}_i = \alpha + \beta(\text{sample.size}_i) + \varepsilon_i,$$
$$\text{where } \varepsilon_i \tilde{N}\left(0, s.e.(\text{residual}_i)^2 \varphi\right), \tag{2b}$$

$$\text{residual}_i = \alpha + \beta(1/\text{sample.size}_i) + \varepsilon_i,$$
$$\text{where } \varepsilon_i \tilde{N}\left(0, s.e.(\text{residual}_i)^2 \varphi\right), \tag{2c}$$

where φ is a multiplicative dispersion parameter estimated from the data which allows for heterogeneity inflation. Stata code used to perform these analyses is available in Data S1.

3.2 | Results of applying modified assessment to postoperative analgesic trials meta-analyses

Figure 2 presents pairs of funnel plots for the six of the nine meta-analyses described in Section 2 that demonstrated funnel plot asymmetry. The first of each pair, using blue plotting symbols, presents the residuals of the meta-regression model adjusting for baseline risk morphine consumption (x-axis) vs inverse sample size of the individual studies (y-axis) (as proposed in Section 3.1). The second funnel of each pair, using red plotting

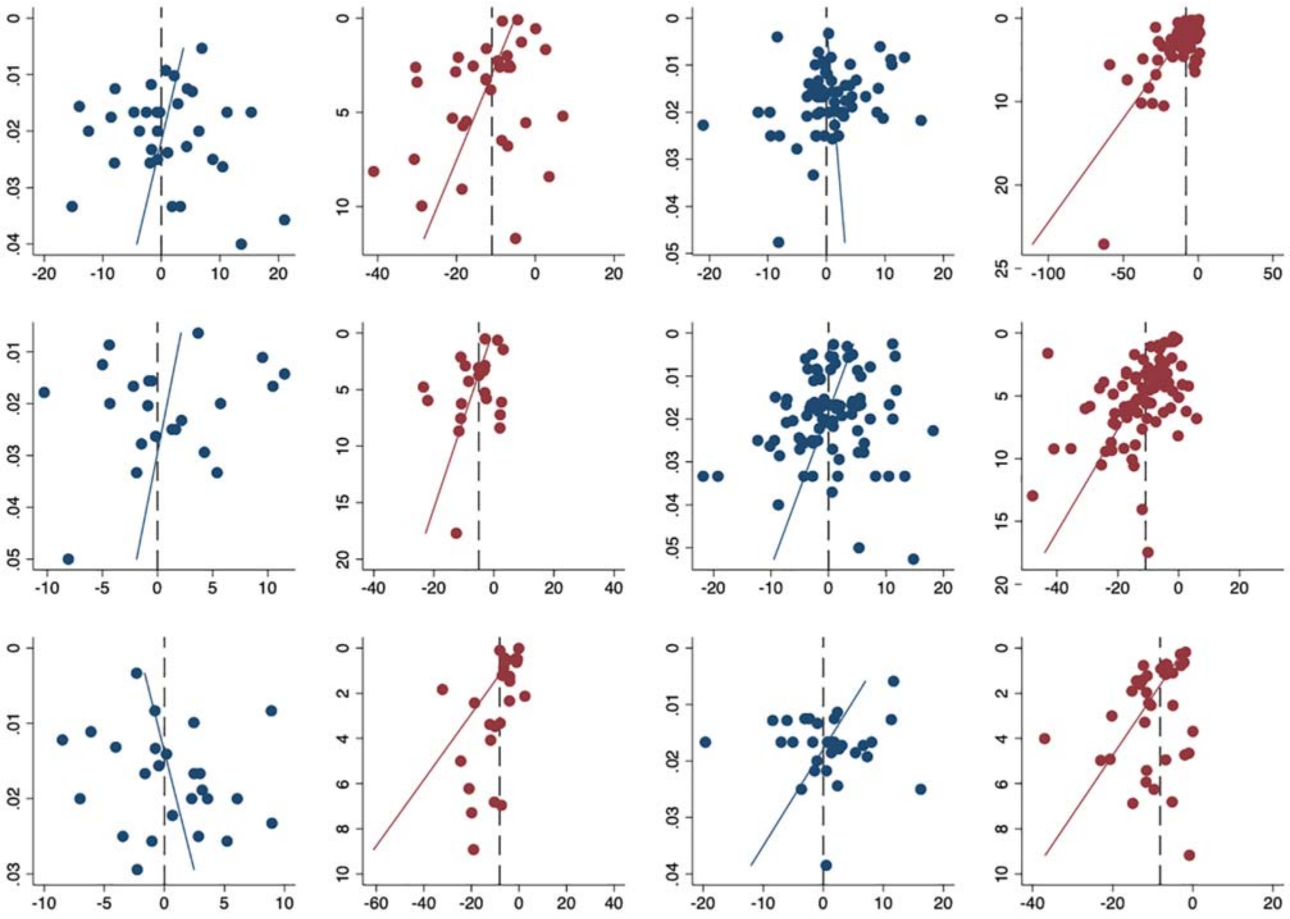

FIGURE 2 Funnel plots of postoperative pain meta-analyses with residuals (x-axis) vs inverse sample size (on a reverse scale, blue plots, y-axis) and traditional plots with mean difference (x-axis) vs SE (on a reverse scale, red plots, y-axis). Plots from top left with p values for new publication bias test in parentheses: alpha-2 agonists (P = .17), gabapentin (P = .55), lidocaine (P = .60), NSAIDs (P = .002), paracetamol (P = .20) and pregabalin (P = .01). It can be observed that the new method results in more plots demonstrating funnel plot symmetry. Coloured lines are from regression asymmetry test [Colour figure can be viewed at wileyonlinelibrary.com]

symbols, is a "standard" funnel plot of the observed data, that is, the mean difference vs SE (these are essentially the same plots as in Figure 1 without the symbol coding for baseline risk). Visual examination suggests that asymmetry has been greatly reduced by adjusting for baseline risk × treatment interactions and plotting inverse sample size on the *y*-axis as described above. The associated regression test (Equation 2c) *P*-values are also given in the Figure 2 legend and support the visual inspection in that four of the six funnels have *P*-values that would be considered non-significant at either 5% or 10% levels.

Given the promising results of the assessment described above, we decided to formally evaluate the modified regression test to establish its performance, both compared to the standard approach and also in absolute terms. In addition, since three measures of study "size" had been considered (SE, sample size, inverse sample size) we wished to explore how these performed compared to one another.

4 | SIMULATION STUDY TO EVALUATE RESIDUAL BASED REGRESSION TEST

4.1 | Simulation of meta-analysis data

Data from the postoperative pain trials (Section 2) were examined in order to simulate data with similar characteristics. The approach to data simulation for each trial in each meta-analysis dataset is provided below. In order to simulate data from an individual two-arm trial with a continuous outcome, individual patient responses in the control arm were assumed to be normally distributed. That is,

$$c_i\tilde{N}\left(mu.csigma^2\right), \tag{3a}$$

where c_i is the outcome response for the *i*th patient in the control arm of the trial, *mu.c* is the underlying average response in the control group of the trial (baseline risk), and *sigma* is the SD of responses in a trial. In scenarios where the SD was assumed to depend on the mean of the response in the arm (ie, baseline risk) this expression was extended to:

$$c_i\tilde{N}\left(mu.c, \left(sigma + (0.5\times mu.c)\right)^2\right), \tag{3b}$$

Individual patient responses in the treatment arm are also assumed to be normally distributed, with the same variance as for the control arm, but with a treatment effect added:

$$t_i\tilde{N}\left(mu.c + trt.diff sigma^2\right), \tag{4a}$$

where t_i is the outcome response for the *i*th patient in the treatment arm of the trial, *trt.diff* is the intervention effect and all other variables are as defined in Equation (3a). In scenarios where both the SD and treatment effect were assumed to depend on the mean of the response in the arm (ie, baseline risk) this expression was extended to:

$$\begin{aligned} t_i\tilde{N}(mu.c + trt.diff - (b.interaction \times mu.c),\\ \left(sigma + (0.5\times mu.c)\right)^2\right), \end{aligned} \tag{4b}$$

with all terms being defined previously except *b.interaction* which represents the strength of the baseline risk × treatment effect interaction. Thus the equations indicate that the variance in the treatment arm was specified to be the same as in the control arm (and thus was assumed to be influenced by baseline risk in the same way). Note that *trt.diff* was always held constant across trials within any meta-analysis which implies that, other than the effect of baseline risk on treatment effect, homogeneity of treatment effects over all studies in each meta-analysis was assumed. The Normal distributions in Equations (3) and (4) were truncated at 0 to ensure response could not go negative (which makes no sense in the context of the motivating morphine consumption outcome).

For scenarios in which publication bias was simulated, this was achieved by excluding any trials that generated a *P*-value >.05 for the effect size and generating further trials until the meta-analysis consisted of the prespecified number of trials, all with *P*-values ≤.05.

The summary statistics from each study, required for meta-analysis, were the observed mean responses in both arms, the SDs of observed responses in both arms and the sample sizes in both arms. From these the estimated mean difference and associated variance could be calculated for use in the meta-analysis. These could be derived in a straightforward manner once trial data had been simulated using the approach described above.

Eight scenarios were considered, consisting of all permutations of whether baseline risk interactions with treatment and publication bias were present or not and whether a small (15) or large (30) number of trials were available. For all scenarios, within a simulated meta-analysis dataset for each trial, "baseline" (*mu. c*) took one of the values 20, 25, 30, 35, 40, 45 or 50 and trial arm size took one of the values 15, 25 or 50 (patients); both evenly distributed across trials. The differences across the eight scenarios are explicitly outlined below (N.B. in interpreting test results we consider asymmetry in the funnel plot [$P < .05$] to be an indication of publication bias):

1 30 trials per meta-analysis, no baseline risk interaction and no publication bias (*treat.diff* = −5, *sigma* =10)
2 30 trials per meta-analysis, baseline risk interaction (*b. interaction* = 0.7) and no publication bias (*treat. diff* = −5, *sigma* =10)
3 30 trials per meta-analysis, no baseline risk interaction and publication bias (*treat.diff* = −3, *sigma* =6)
4 30 trials per meta-analysis, baseline risk interaction (*b. interaction* = 0.5) and publication bias (*treat.diff* = −5, *sigma* =10)
5 15 trials per meta-analysis, no baseline risk interaction, no publication bias (*treat.diff* = −5, *sigma* =10)
6 15 trials per meta-analysis, baseline risk interaction (*b. interaction* = 0.7) and no publication bias (*treat. diff* = −5, *sigma* =10)
7 15 trials per meta-analysis, no baseline risk interaction and publication bias (*treat.diff* = −3, *sigma* =6)
8 15 trials per meta-analysis, baseline risk interaction (*b. interaction* = 0.5) and publication bias (*treat.diff* = −5, sigma =10)

Note, that the strength of treatment effect, magnitude of participant-level variance and magnitude of interaction with baseline risk change between the eight scenarios above. This was done to ensure that the underlying effects of publication bias were neither too large (ie, most studies suppressed) or too small (ie, no or virtually no studies suppressed) while also keeping the simulations broadly representative of the motivating datasets (eg, preventing any negative outcomes for individuals). For all scenarios we simulated 10 000 meta-analyses. Each simulated meta-analysis dataset was analysed in six ways. Where adjustment for baseline risk is conducted, this was achieved using the mixed-effect model given by Equation (1).

1. Egger's original regression test (with s.e.[mean difference] as the predictor) for funnel plot asymmetry on observed data (conventional test).

2. Regression test (with s.e.[residuals] as the predictor) for funnel plot asymmetry on residuals following adjustment for baseline risk (Equation (2a)).

3. Regression test with sample size as the predictor for funnel plot asymmetry on observed data.

4. Regression test with sample size as the predictor for funnel plot asymmetry on residuals following adjustment for baseline risk (Equation (2b)).

5. Regression test with inverse sample size as the predictor for funnel plot asymmetry on observed data.

6. Regression test with inverse sample size as the predictor for funnel plot asymmetry on residuals following adjustment for baseline risk (Equation (2c)).

The estimand of interest is the proportion of times the *P*-value for the funnel plot asymmetry test is less than or equal to .05. We calculated its SE using the following formula:

$$\mathrm{SE} = \sqrt{\frac{pq}{n}},$$

where p is the probability of an event, $q = 1 - p$ and n is the sample size (ie, number of simulations which always equalled 10 000).

Data were simulated and analysed using R (version 3.6.0) in RStudio (version 1.0.143.0). Meta-regression and all regression tests[3] were implemented using the **metafor** package in R.[13] The code used to generate and analyse the simulated data can be found in Data S2.

4.2 | Results of simulation study

A summary of the performance of each of the six testing procedures across all eight scenarios is shown in Table 2. When there is no baseline risk interaction or publication bias present (scenarios 1 and 5), all six approaches to testing for funnel asymmetry produced a significant result approximately 5% of the time (ie, the nominal rate expected by chance alone).

When a baseline risk × treatment interaction (but not publication bias) is present (scenarios 2 and 6), the funnel plot asymmetry test using SE as the predictor (conventional Egger's test) incorrectly identifies evidence of statistically significant funnel plot asymmetry 60% of the time based on the observed data when considering meta-analyses including 30 studies (scenario 2). If the number of trials in a meta-analysis is reduced to 15, statistically significant funnel asymmetry is identified 39% of the time (scenario 6). When sample size or inverse sample size are used to test for asymmetry, using the observed effect sizes, significant results are obtained in approximately 1% to 2% of simulations. Once a regression adjustment for baseline risk is conducted prior to testing the test based on SE identifies significant asymmetry 8% of the time, while the test based on sample size or its inverse identifies asymmetry approximately 5% to 6% of the time that is, close to the nominal 5%.

When publication bias is present but there is no effect of baseline risk on treatment, all approaches to testing correctly identify evidence of statistically significant publication bias approximately the same number of times (approximately 60% of the time when meta-analyses

TABLE 2 Comparisons of naïve/conventional (Egger's) tests for asymmetry of funnel plot vs tests based on meta-regression residuals having adjusted for the effect of baseline risk using SE, sample size and inverse sample size as predictors in the asymmetry tests. 10 000 simulations were used for every scenario

	Proportion of times $P \leq .05$ for asymmetry test					
	Publication bias test using SE		Publication bias test using sample size		Publication bias test using inverse sample size	
Scenario:	Naïve/conventional on observed data (Egger's test)	Two-stage (Test meta-regression residuals adjusting for baseline risk)	Naïve/Conventional on observed data	Two-stage (Test meta-regression residuals adjusting for baseline risk)	Naïve/conventional on observed data	Two-stage (Test meta-regression residuals adjusting for baseline risk)
30 trials per simulated meta-analysis						
1. No baseline risk, no publication bias	0.0444 (SE = 0.00206)	0.0438 (SE = 0.00205)	0.0532 (SE = 0.00224)	0.0568 (SE = 0.00231)	0.0558 (SE = 0.00230)	0.0585 (SE = 0.00235)
2. Baseline risk, no publication bias	0.6026 (SE = 0.00489)	0.0780 (SE = 0.00268)	0.0089 (SE = 0.00094)	0.0563 (SE = 0.00231)	0.0116 (SE = 0.00107)	0.0602 (SE = 0.00238)
3. No baseline risk, publication bias	0.6324 (SE = 0.00482)	0.6249 (SE = 0.00484)	0.5892 (SE = 0.00492)	0.5806 (SE = 0.00493)	0.6267 (SE = 0.00484)	0.6252 (SE = 0.00484)
4. Baseline risk, publication bias	0.9860 (SE = 0.00117)	0.0067 (SE = 0.00082)	0.3547 (SE = 0.00478)	0.3595 (SE = 0.00480)	0.3319 (SE = 0.00471)	0.4018 (SE = 0.00490)
15 trials per simulated meta-analysis						
5. No baseline risk, no publication bias	0.0466 (SE = 0.00211)	0.0541 (SE = 0.00226)	0.0551 (SE = 0.00228)	0.0624 (SE = 0.00242)	0.0551 (SE = 0.00228)	0.0624 (SE = 0.00242)
6. Baseline risk, no publication bias	0.3865 (SE = 0.00487)	0.0587 (SE = 0.00235)	0.0178 (SE = 0.00132)	0.0624 (SE = 0.00242)	0.0178 (SE = 0.00132)	0.0624 (SE = 0.00242)
7. No baseline risk, publication bias	0.3109 (SE = 0.00463)	0.3137 (SE = 0.00464)	0.3012 (SE = 0.00459)	0.2974 (SE = 0.00457)	0.3149 (SE = 0.00464)	0.3212 (SE = 0.00467)
8. Baseline risk, publication bias	0.6977 (SE = 0.00459)	0.0063 (SE = 0.00079)	0.1126 (SE = 0.00316)	0.1541 (SE = 0.00361)	0.0955 (SE = 0.00294)	0.1608 (SE = 0.00367)

contain 30 trials [scenario 3] and approximately 30% of the time when meta-analyses contain 15 trials [scenario 7]).

When baseline risk × treatment interactions and publication bias are both present the funnel plot asymmetry test on the observed data identifies publication bias 99% of the time if meta-analyses consist of 30 trials (scenario 4) and 70% of the time if meta-analyses consist of 15 trials (scenario 8). Note, however, that these results need to be viewed within the context that, when only baseline risk was present, the test for publication bias was "already" significant 60% (30 trials) and 39% (15 trials) of the time (scenarios 2 and 6, respectively). This reduces to around 34% and 10% of the time when sample size or its inverse are used in the test (vs observed) for 30 and 15 study meta-analyses, respectively. After adjustment for a baseline risk interaction, the funnel plot asymmetry test on the residuals using SE as predictor correctly identifies publication bias <1% of the time (both 30 and 15 studies). However, when sample size or its inverse is used as predictor and residuals as outcome, significant asymmetry is identified approximately 36% to 40% (30 studies) and 15% to 16% (15 studies) of the time.

Figure 3 presents funnel plots for the first 4 simulated datasets for each of scenarios 1 to 4, firstly (in blue) plotted using the residuals, having adjusted for baseline risk, vs inverse sample size (analysis approach 6/column 6 of Table 2), and secondly (in red) using the observed effects vs SE (analysis approach 1/column 1 of Table 2). In each case the regression line for the associated test is also plotted. These plots broadly reflect the overall results of the simulation, in which performance is vastly superior when adjusting for baseline risk and using inverse sample size as the test regressor for the situation where a baseline risk × treatment interaction but no publication bias exists (scenario 2) as the regressions are visibly much closer to vertical lines, correctly implying little or no asymmetry. The two approaches are broadly comparable (ie, regression slopes are similar between pairs of plots) when no publication bias or interaction exist (scenario 1) or only publication bias exists (scenario 3). The lack of strong trend among the residual-based plots for scenario 4 (both baseline risk interaction and publication bias present) reflects the diminished power of the residual-based test when both publication bias and baseline risk interaction effects are present.

5 | DISCUSSION

The prevalence of publication bias within meta-analyses is estimated to be around 25% to 40%.[14] Within the anaesthesia literature, using a sample of systematic reviews from leading anaesthetic journals, the prevalence of publication bias may be as high as 50% to 80%.[15] This has important implications for the validity of systematic

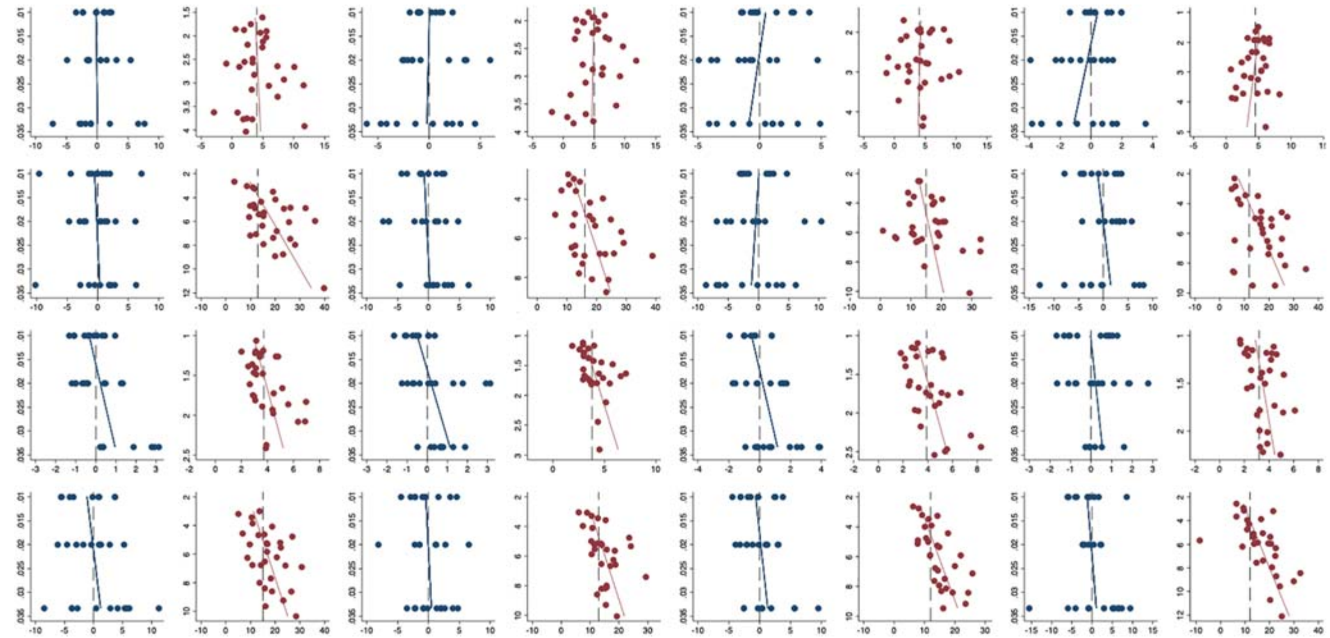

FIGURE 3 Example funnel plots of simulated meta-analyses with residuals (*x*-axis) vs inverse sample size (on a reverse scale, blue plots, *y*-axis) and traditional plots with mean difference (*x*-axis) vs SE (on a reverse scale, red plots, *y*-axis). Each row represents example datasets from top to bottom: scenario 1, scenario 2, scenario 3 and scenario 4 (see Table 2) [Colour figure can be viewed at wileyonlinelibrary.com]

review findings, as publication bias may be the cause when meta-analyses and subsequent large randomised controlled trials disagree.[16] Meta-analyses are frequently used to inform clinical decision-making and guidelines; therefore, using invalid data may lead to the use of ineffective or even harmful interventions in clinical practice. Indeed, possible publication bias is one factor that can cause downgrading of evidence as per Grading of Recommendations, Assessment, Development and Evaluations (GRADE).[1]

Although there is a wealth of research into the use of funnel plots and quantitative tests for the detection of possible publication bias, little work has been conducted on research using continuous outcomes such as morphine consumption.[17] Moreover, less work has been undertaken examining continuous outcomes with considerable heterogeneity and variation in effect estimates dependent on baseline risk. For the motivating context, we have previously demonstrated that the results from any one study are dependent on control group morphine consumption (with higher baseline risk having larger reductions in morphine consumption).[10] In addition, randomised controlled trials of postoperative analgesics are often small (50-100 participants) and therefore SE calculations will be more dependent on SDs than for larger studies (as $\text{SE} = \text{SD}/\sqrt{N}$). As there is a tendency for studies with higher control group morphine consumption to have larger SDs there is a dependency between the mean difference (larger with higher baseline risk) and the SEs (larger with higher baseline risk). This could create an asymmetric funnel plot even in the presence of no publication bias. Indeed, when we simulated meta-analyses where no publication bias was present but outcomes were dependent on baseline risk, funnel plot asymmetry was evident in approximately 60% of analyses when 30 trials were present.

This study has identified evidence of asymmetric study effects, using conventional methods, in meta-analyses of postoperative analgesics. However, when examining the relation between control group morphine consumption (baseline risk) and precision, for most analgesics there was a statistically significant relationship between baseline risk and effect size SEs, implying that standard funnel plots may be an inaccurate method to assess publication bias where values are dependent on baseline risk. On simulated data, when baseline risk is present without publication bias, the conventional funnel plot asymmetry test based on the observed data frequently incorrectly suggested the presence of publication bias. Adjusting for baseline risk via meta-regression and conducting the funnel plot asymmetry test on the residuals results in fewer meta-analyses being incorrectly identified as having publication bias. However, when publication bias is present without baseline risk the funnel plot asymmetry tests on both the observed data and the residuals, following adjustment for baseline risk, correctly identify publication bias approximately the same number of times. In our simulation, since baseline risk interactions with treatment and simulated publication bias both induce funnel plot asymmetry, when SE is used in the regression test/on the funnel plot *y*-axis, the approach struggles to disentangle these causes of the asymmetry when both are present. However, much improved performance was gained by the use of (inverse) sample size for the regression test/*y*-axis scale since this avoids the issue (as explained above) that studies with higher baseline risks/effect sizes also tend to have larger SEs, although power was still lower than the case where baseline risk interactions were not present. Improved performance using (a function of) sample size instead of SE in such tests is consistent with previous findings for meta-analysis of odds ratios since a similar dependency exists in that context[18] as well as diagnostic odds ratios[19] and we recommend its adoption for continuous, mean difference type outcomes. Since using inverse sample size had fractionally superior performance to sample size in the simulation study we recommend this analysis approach (analysis variant 6).

We conclude that traditional funnel plots are not a reliable method to detect asymmetric study effects for morphine consumption and that this finding may also extend to other, similar continuous outcomes whose results are dependent on baseline risk (such as pain scores[8] or depression scores[20]). Indeed, improved stability of effects is an argument routinely given for using relative over absolute effect measures for binary outcomes due to varying baseline rates.[21] Since such variability may often go unacknowledged for continuous outcomes, we recommend further empirical work looking at the stability of such outcomes in meta-analysis and recommend meta-analysists explore the relationship between outcome and baseline risk routinely, using meta-regression, when conducting meta-analysis of continuous outcomes.

This dependency can also present issues beyond publication bias assessments for meta-analyses of continuous outcomes. If the results from a meta-analysis vary with baseline risk, this will affect the weighting of individual studies when calculating pooled effect estimates. As studies with lower baseline risk (lower control group morphine consumption) will have smaller SEs, they will receive a higher percentage weight than studies with higher baseline risk (using the inverse-variance method). This will mean effect estimates will be lower than the true average effect, leading to a possible underestimation of efficacy in high baseline risk scenarios. This further supports the argument to report effect estimates from a fixed value of baseline risk or provide meta-regression parameter estimates to allow review consumers to calculate specific effect estimates for the baseline risk of their clinical population.[10]

Clearly, the issues highlighted above have implications for the interpretation of results derived from meta-analyses. Incorrect conclusions regarding the presence of publication bias could lead to unnecessary downgrading of evidence as per GRADE.[22] In addition, the conduct of trim and fill analysis, or one of several other methods proposed to adjust for publication bias, could reduce effect estimates and significantly alter a reviews conclusions, which may be inappropriate.[23] These factors need to be considered when performing meta-analyses using postoperative morphine consumption (a common outcome in postoperative pain trials) and similar continuous outcomes dependent on baseline risk (such as pain scores).

In terms of the conduct of future meta-analyses, if review authors are using mean differences as the effect estimate method, then the test based on meta-regression residuals may have advantageous properties over conventional tests (see Data S1 to perform test in Stata). Our simulations demonstrated it performs similarly to conventional tests when baseline risk is not present and may reduce type 1 errors in the presence of baseline risk. However, this test has lower power to detect publication bias in the presence of baseline risk if studies are 30 or less. Although conventional tests detected around 99% and 70% of cases of publication bias in the presence of baseline risk, the utility of this approach is questionable due to the higher number of false positives when no publication bias and baseline risk co-exist (60% and 39%).

Our work could be extended in a number of directions. Firstly, we have assumed that the relationship between baseline risk and outcome is linear throughout, and the impact of relaxing this could be investigated. A further continuous outcome routinely used in meta-analysis, but not considered here is the standardized mean difference (SMD). Like the (unstandardized) mean difference, this has received less attention when considering publication bias assessments; however, previous research has shown that use of SMDs and SEs can cause funnel plot distortion.[5,24] There is clearly further work needed to inform how best to assess SMD outcomes for publication bias, including how well the methods presented here translate.

A further option for meta-analysis of continuous outcomes is the use of a relative scale such as the ratio of means as the outcome measure. However, this method does not resolve the issue of statistical heterogeneity as shown in previous studies[25] and using the analgesic meta-analysis data considered here.[26] We can only speculate what the causes of this statistical heterogeneity are, although it essentially does not solve the problem that is solved by our baseline risk meta-regression models published previously.[10] In addition, relative measures may have less clinical significance than absolute measures. For example, with regards to morphine consumption, a 0.5 relative measure could correspond to a reduction of 50 mg (if 100 mg consumption) or 5 mg (10 mg consumption) which has particular relevance for reducing the dose-dependent adverse effects of opioids.[27] Despite this, ratio of means may offer an alternative method of publication bias assessment and could be the focus of future simulation studies similar to ours.

The first limitation of this study is the use of previously published reviews with variable search strategies in identifying the motivating analgesic data.[10] The fact that only a small number of included reviews searched for unpublished studies means our sample would be more likely susceptible to publication bias. Secondly, some of our analyses contained a low number of primary studies, which may render quantitative tests for publication bias underpowered. Of note, power may be improved (at the expense of increased type I errors) with our tests if P value thresholds were changed to recommended levels of $P < .1$ for example. Thirdly, regarding the methodology we used and developed, it has been well documented that there is structural dependency in a meta-regression of baseline risk on outcome[28] and specific methods are required to ensure regression to the mean does not bias results. Since we applied such methods to the motivating analgesic datasets in a previous paper and found they had minimal impact[10] we chose not to apply them here so as not to overcomplicate the analysis. However, regression to the mean may have a larger impact in other contexts and therefore the use of such methods is generally recommended. Fourthly, we simulated a limited set of conditions and it is therefore unknown how the novel test performs under the conditions of less-than-absolute selective publication bias simulated in our study. In addition, we did not consider extra unexplainable heterogeneity on top of that induced by the dependency of outcome on baseline risk (ie, explainable systematic heterogeneity). And we acknowledge this is a potential limitation of our simulation study, however from previous related work we strongly suspect such extra variability would reduce the power of the regression testing.[29] This could be the focus of future studies.

Our approach to assessing the likely presence of publication bias could be viewed as "2-stage", that is, the data is initially adjusted for the effects of baseline risk via a regression prior to a second regression to test for funnel asymmetry. It would be possible to achieve similar results using a single "1-stage" regression analysis simultaneously including terms for both baseline risk and (inverse) sample size. This may even be more efficient than the "2-stage" approach taken, however we did not pursue this for two reasons: (a) Standard random effect meta-regression models have additive heterogeneity variance parameters, while, funnel asymmetry tests have a history of incorporating multiplicative error terms instead (initially due to how the Egger's test was first conceived, and later because they were shown to have better statistical properties than models with additive errors[30]). If a single regression were used it is unclear whether the regression should have additive, or multiplicative heterogeneity parameters, or even both; to keep things manageable we stuck with convention. (b) We feel it is important to consider the visual impact of the adjusted funnel plot as well as the test *P*-value, and this is constructed using the residuals from the first regression of the "2-stage" approach. Finally, we note that we did not attempt to use selection modelling[31] to address the problem and since this has been used successfully in other publication bias contexts, it may provide a fruitful alternative approach.

In conclusion, using conventional methods, we found evidence of asymmetric study effects for most analgesics used to prevent postoperative pain. However, due to an association between baseline risk and SEs, this finding is a result of statistical artefact as demonstrated in our simulations of meta-analyses where no publication bias was present. In response to this we proposed a novel alternative approach to assessing whether publication bias is likely to be present, by first adjusting for baseline risk treatment interactions and regressing on inverse sample size (rather than SE). When evaluated using a simulation study, although power was low for meta-analyses with 15 studies, the approach performed considerably better than alternatives and thus may be advantageous for routine use with unstandardized mean difference outcome meta-analyses where treatment effects are dependent on baseline risk. Additionally, given accumulating evidence on the dependency of continuous outcomes on baseline rates, the possibility of such relationships should be explored in meta-analyses of continuous outcomes as a possible explanation for any between-study heterogeneity and to aid interpretation of results.

ACKNOWLEDGEMENTS

We would like to thank Dr Matthew Sherwin for help retrieving the articles used in this study. SCF is a member of the Complex Reviews Support Unit which is funded by the National Institute for Health Research (project number 14/178/29). The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the NIHR, NHS or the Department of Health.

CONFLICT OF INTEREST

The author reported no conflict of interest.

AUTHOR CONTRIBUTIONS

B.D. developed study conception, data analysis, writing manuscript and approving final version. S.F. worked on data analysis, writing manuscript and approving final version. J.N.L worked on data analysis, writing manuscript and approving final version. J.P.W. worked on data analysis, writing manuscript and approving final version. A.J.S worked on study conception, data analysis, writing manuscript and approving final version.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Brett Doleman https://orcid.org/0000-0003-4707-4755
Suzanne C Freeman https://orcid.org/0000-0001-8045-4405

REFERENCES

1. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. *J Clin Epidemiol*. 2011;64:1277-1282.
2. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev*. 2009;1: MR000006.
3. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *Br Med J*. 1997; 315:629-634.
4. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA*. 2006;295:676-680.
5. Zwetsloot PP, Van Der Naald M, Sena ES, et al. Standardized mean differences cause funnel plot distortion in publication bias assessments. *Elife*. 2017;6:1-20.
6. Sterne JA, Sutton AJ, Ioannidis JP *et al*. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *Br Med J* 2011; 343:1-8.
7. Hunter JP, Saratzis A, Sutton AJ, Boucher RH, Sayers RD, Bown MJ. In meta-analyses of proportion studies, funnel plots were found to be an inaccurate method of assessing publication bias. *J Clin Epidemiol*. 2014;67:897-903.
8. Doleman B, Heinink TP, Read DJ, Faleiro RJ, Lund JN, Williams JP. A systematic review and meta-regression analysis of prophylactic gabapentin for postoperative pain. *Anaesthesia*. 2015;70:1186-1204.
9. Doleman B, Read D, Lund JN, Williams JP. Preventive acetaminophen reduces postoperative opioid consumption, vomiting, and pain scores after surgery: systematic review and meta-analysis. *Reg Anesth Pain Med*. 2015;40:706-712.

10. Doleman B, Sutton AJ, Sherwin M, Lund JN, Williams JP. Baseline morphine consumption may explain between-study heterogeneity in meta-analyses of adjuvant analgesics and improve precision and accuracy of effect estimates. *Anesth Analg*. 2018;126:648-660.
11. StataCorp. *STATA Version 14.2*. Texas, USA: StataCorp; 2015.
12. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11): 1559-1573.
13. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Soft*. 2010;36(3):1-48.
14. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol*. 2000;53:1119-1129.
15. Hedin RJ, Umberham BA, Detweiler BN, Kollmorgen L, Vassar M. Publication bias and nonreporting found in majority of systematic reviews and meta-analyses in anesthesiology journals. *Anesth Analg*. 2016;123:1018-1025.
16. Sivakumar H, Peyton PJ. Poor agreement in significant findings between meta-analyses and subsequent large randomised controlled trials in perioperative medicine. *Br J Anesth*. 2016;117: 431-441.
17. Higgins JP, Green S. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1. 0 (updated March 2011)*. Chichester, UK: The Cochrane Collaboration; 2011.
18. Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. *Stat Med*. 2008;27(5):746-763.
19. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005; 58(9):882-893.
20. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med*. 2008;5:e45.
21. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med*. 2002;21(11):1575-1600.
22. Doleman B, Leonardi-Bee J, Heinink TP, Bhattacharjee D, Lund JN, Williams JP. Pre-emptive and preventive opioids for postoperative pain in adults undergoing all types of surgery. *Cochrane Database Syst Rev*. 2018;12:CD012624.
23. Rothstein HR, Sutton AJ, Borenstein M, eds. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Hoboken, New Jersey, USA: John Wiley & Sons; 2006.
24. Pustejovsky JE, Rodgers MA. Testing for funnel plot asymmetry of standardized mean differences. *Res Synth Methods*. 2019; 10(1):57-71.
25. Friedrich JO, Adhikari NK, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *J Clin Epidemiol*. 2011;64:556-564.
26. Doleman B, Lund JN, Williams JP. Ratio of means fails to resolve statistical heterogeneity in meta-analyses of postoperative analgesics. *Eur J Anaesth*. 2019;36:e57.
27. Zhao SZ, Chung F, Hanna DB, Raymundo AL, Cheung RY, Chen C. Dose-response relationship between opioid use and adverse effects after ambulatory surgery. *J Pain Symptom Manage*. 2004;28:35-46.
28. Ghidey W, Stijnen T, van Houwelingen HC. Modelling the effect of baseline risk in meta-analysis: a review from the perspective of errors-in-variables regression. *Stat Methods Med Res*. 2013;22(3):307-323.
29. Peters J, Sutton A, Jones D, Abrams K, Rushton L, Moreno S. Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *J R Stat Soc*. 2010;173(3):575-591.
30. Moreno SG, Sutton AJ, Ades AE, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol*. 2009;9(1):2.
31. Mavridis D, Welton NJ, Sutton AJ, Salanti G. A selection model for accounting for publication bias in a full network meta-analysis. *Stat Med*. 2014;33(30):5399-5412.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Doleman B, Freeman SC, Lund JN, Williams JP, Sutton AJ. Funnel plots may show asymmetry in the absence of publication bias with continuous outcomes dependent on baseline risk: presentation of a new publication bias test. *Res Syn Meth*. 2020;11: 522–534. https://doi.org/10.1002/jrsm.1414