Received: 24 January 2017 | Revised: 17 August 2017 | Accepted: 7 September 2017

DOI: 10.1002/jrsm.1266

RESEARCH ARTICLE

WILEY Research Synthesis Methods

Detecting small-study effects and funnel plot asymmetry in meta-analysis of survival data: A comparison of new and existing tests

Thomas P. A. Debray[1,2] | Karel G. M. Moons[1,2] | Richard D. Riley[3]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

[2]Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

[3]Research Institute for Primary Care and Health Sciences, Keele University, Newcastle, ST5 5BG, Staffordshire, UK

Correspondence
Thomas P. A. Debray, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.
Email: T.Debray@umcutrecht.nl

Funding information
The Netherlands Organization for Health Research and Development, Grant/Award Number: 91617050

Small-study effects are a common threat in systematic reviews and may indicate publication bias. Their existence is often verified by visual inspection of the funnel plot. Formal tests to assess the presence of funnel plot asymmetry typically estimate the association between the reported effect size and their standard error, the total sample size, or the inverse of the total sample size. In this paper, we demonstrate that the application of these tests may be less appropriate in meta-analysis of survival data, where censoring influences statistical significance of the hazard ratio. We subsequently propose 2 new tests that are based on the total number of observed events and adopt a multiplicative variance component. We compare the performance of the various funnel plot asymmetry tests in an extensive simulation study where we varied the true hazard ratio (0.5 to 1), the number of published trials ($N = 10$ to 100), the degree of censoring within trials (0% to 90%), and the mechanism leading to participant dropout (noninformative versus informative). Results demonstrate that previous well-known tests for detecting funnel plot asymmetry suffer from low power or excessive type-I error rates in meta-analysis of survival data, particularly when trials are affected by participant dropout. Because our novel test (adopting estimates of the asymptotic precision as study weights) yields reasonable power and maintains appropriate type-I error rates, we recommend its use to evaluate funnel plot asymmetry in meta-analysis of survival data. The use of funnel plot asymmetry tests should, however, be avoided when there are few trials available for any meta-analysis.

KEYWORDS
funnel plot, meta-analysis, publication bias, RCT, small-study effects, survival

1 | INTRODUCTION

The presence of small-study effects is a common threat to systematic reviews and meta-analyses, especially when it is due to publication bias, which occurs when small primary studies are more likely to be reported (published) if their findings were positive.[1,2] The presence of small-study effects is often verified by visual inspection of the funnel plot,[3-5] where for each included study of the meta-analysis, the estimate of the reported effect size is plotted against a measure of precision or sample size. The premise is that the scatter of plots should reflect a funnel shape, if small-study effects do not exist (provided that effect sizes are not substantially affected by the presence of between-study heterogeneity). However, when small studies are predominately in one direction (usually the direction of larger effect sizes), asymmetry will ensue. Because an inevitable degree of subjectivity exists in the interpretation of funnel plots, several tests have been proposed for detecting funnel plot asymmetry.[6,7] These tests may regress effect estimates against their standard error (the so-called Egger's test), their underlying sample size,[8] or the inverse of their underlying sample size.[9] Guidelines for conducting funnel plot asymmetry tests recommend to include at least 10 studies to maintain sufficient power for distinguishing chance from real asymmetry.[4,10,11]

Funnel plot asymmetry tests are currently being used for numerous types of estimates, such as odds ratios, risk ratios, and mean differences. As far as we are aware, the performance of these tests has never been evaluated for survival data,[4,12] and it is unclear whether 10 studies are indeed sufficient to detect small-study effects in such data.

In this paper, we investigate several approaches to evaluate small-study effects (funnel plot asymmetry) in a meta-analysis of hazard (rate) ratios. We propose a novel test that is tailored for survival data and illustrate its implementation in 3 exemplar reviews. Afterwards, we compare their performance in an extensive simulation study where meta-analyses are generated on a set of characteristics intended to reflect meta-analyses of randomized clinical trials in the medical literature. This simulation study adopts a new probabilistic mechanism to generate meta-analyses that are affected by selection but do not necessarily suffer from substantial funnel plot asymmetry. In this manner, we aim to assess to what extent funnel plot asymmetry test results can be used as an indication of publication bias.

2 | METHODS

Consider a meta-analysis of randomized trials, each containing a treatment and control group. Let X_{ij} and T_{ij} denote the allocation group (where $X_{ij} = 1$ is treated) and uncensored survival time respectively for subjects $i = 1, \dots, n_j$ in trial $j = 1, \dots, m$. Assuming a common treatment effect across studies, let β denotes the log hazard ratio of the *true* treatment effect and $\hat{\beta}_j$ the estimated treatment effect in trial j. Furthermore, let Δ_{ij} represents an indicator variable, which denotes for each subject in each study whether the event occurred during the study period ($\Delta_{ij} = 1$) or if the survival time was censored ($\Delta_{ij} = 0$). The total number of events for study j is then given as $d_j = \sum_{i=1}^{n_j} \Delta_{ij}$. Finally, we denote the censoring time points as L_{ij}, such that the observed follow-up time for each subject is given as $Z_{ij} = \min(L_{ij}, T_{ij})$ and the total follow-up time for study j is given as $z_j = \sum_{i=1}^{n_j} Z_{ij}$.

The observed data for subject i in trial j consist of $[X_{ij}, \Delta_{ij}, Z_{ij}]$. Subsequently, it is assumed that each trial yields a vector of study-level data $\boldsymbol{S}_j = [\hat{\beta}_j, \widehat{\mathrm{SE}}(\hat{\beta}_j), n_j, d_j, z_j]$. Let $\boldsymbol{V} \subseteq \boldsymbol{S}$ now denotes the study-level data for $m^{\mathrm{pb}} \leq m$ trials that are actually reported in the literature (ie, published). Below, we consider 5 methods for examining funnel plot asymmetry in published hazard ratios using information from $\boldsymbol{V}$.

The most common method to test the presence of small-study effects is given as the following (unweighted) regression model[11]:

$$\hat{\beta}_k = a + b\ \widehat{\mathrm{SE}}(\hat{\beta}_k) + \epsilon_k\ ,\ \epsilon_k \sim \mathcal{N}\left(0, \sigma^2\right) \qquad \text{(E-UW)}$$

where $k = 1, \dots, m^{\mathrm{pb}}$ and $\hat{\beta}_k$ is the estimated log hazard ratio in study k. The unknown parameters are the intercept term a, the slope b, and the error variance σ^2. Whereas a indicates the size and direction of the treatment effect, b provides a measure of asymmetry; the larger its deviation from zero is, the more pronounced the asymmetry is. Otherwise, if $b = 0$, there is no association between the estimated effect sizes $\hat{\beta}_k$ and their corresponding estimates for the standard error $\widehat{\mathrm{SE}}(\hat{\beta}_k)$ among the reported studies, indicating no asymmetry and thus no small-study effects.

It is possible to allow for potential heteroscedasticity by replacing σ^2 with a multiplicative overdispersion parameter[7,13] involving $\widehat{\mathrm{var}}(\hat{\beta}_k)$:

$$\hat{\beta}_k = a + b\widehat{\mathrm{SE}}(\hat{\beta}_k) + \epsilon_k\ ,\ \epsilon_k \sim \mathcal{N}(0, \phi\ \widehat{\mathrm{var}}(\hat{\beta}_k)) \qquad \text{(E-FIV)}$$

The corresponding model can be implemented by weighting the study estimates by the inverse variance of their estimated treatment effect (hence FIV, funnel inverse variance). Another method of incorporating residual heterogeneity is to include an additive between-study variance component[14] τ^2. The model is then

$$\hat{\beta}_k = a + b\widehat{\mathrm{SE}}(\hat{\beta}_k) + \epsilon_k\ ,\ \epsilon_k \sim \mathcal{N}(0, \widehat{\mathrm{var}}(\hat{\beta}_k) + \tau^2) \qquad \text{(TS)}$$

where TS stands for Thompson-Sharp.

There are, however, several problems with using $\widehat{\mathrm{SE}}(\hat{\beta}_k)$ as independent (predictor) variable.[15,16] First of all, the independent variable $\widehat{\mathrm{SE}}(\hat{\beta}_k)$ is estimated from the observed data and therefore prone to measurement error.[17] This error becomes particularly pronounced when standard errors are derived from small samples,[18-20] thereby causing bias in estimates for the regression slope b. Additional bias may appear when there is a correlation between the measurement error and the true value of the independent variable.[8,16] This effect has previously been discussed for funnel plot asymmetry tests involving log odds ratios,[15] but may also appear when dealing with log hazard ratios. In general, aforementioned issues imply that estimates for b cannot reliably be used for hypothesis testing when $\widehat{\mathrm{SE}}(\hat{\beta}_k)$ is used as independent variable in the above mentioned regression models with $\hat{\beta}_k$ as outcome. For this reason, Macaskill et al proposed using study sample size (n_k) as an independent variable[8]

$$\hat{\beta}_k = a + b\; n_k + \epsilon_k\ ,\ \epsilon_k \sim \mathcal{N}(0, \phi\; \widehat{\mathrm{var}}(\hat{\beta}_k)) \tag{M-FIV}$$

Again, the intercept term a indicates the size and direction of the treatment effect, and the slope b provides a measure of asymmetry. Note that $\widehat{\mathrm{var}}(\hat{\beta}_k)$ is still included to allow for possible heteroscedasticity, as this strategy was found to generate favorable type-I error rates.[8] To avoid bias in estimates for b resulting from $\widehat{\mathrm{var}}(\hat{\beta}_k)$, Macaskill proposed an alternative test where $\widehat{\mathrm{var}}(\hat{\beta}_k)$ is replaced with the variance of a *pooled* estimate of the outcome proportion[8]:

$$\hat{\beta}_k = a + b\; n_k + \epsilon_k\ ,\ \epsilon_k \sim \mathcal{N}\left(0, \phi\; \frac{1}{d_k(1-d_k/n_k)}\right) \tag{M-FPV}$$

where FPV stands for funnel pooled variance.

Finally, a modification of Macaskill's test was proposed by Peters et al to obtain more balanced type-I error rates in the tail probability areas.[6,9]

$$\hat{\beta}_k = a + b\; \frac{1}{n_k} + \epsilon_k\ ,\ \epsilon_k \sim \mathcal{N}\left(0, \phi\; \frac{1}{d_k(1-d_k/n_k)}\right) \tag{P-FPV}$$

Although aforementioned tests have been evaluated for meta-analyzing odds ratios, their application may be less appropriate for survival data where censoring influences statistical significance (and thus selective reporting) of the hazard ratio. In particular, study sample size has limited influence on the precision of estimated hazard ratios and is therefore likely to have limited power for detecting funnel plot asymmetry of reported hazard ratios. Furthermore, the study weights from the FPV methods (ie, M-FPV and P-FPV) are applicable to proportions and may therefore not be appropriate when some events remain unobserved because of participant dropout.

For this reason, we propose the following regression tests that are based on the total number of events:

$$\hat{\beta}_k = a + b\; \frac{1}{d_k} + \epsilon_k\ ,\ \epsilon_k \sim \mathcal{N}(0, \phi\; \widehat{\mathrm{var}}(\hat{\beta}_k)) \tag{D-FIV}$$

Note that when the event rate is constant over time, $\widehat{\mathrm{var}}\,(\hat{\beta}_k)$ can be approximated[21] by $d_{k1}^{-1}+d_{k2}^{-1}$. In this expression, d_{k1} and d_{k2} represent the number of events in the 2 compared groups (eg, treated or exposed group versus control or unexposed group) of study k. Hence, when $\widehat{\mathrm{var}}\,(\hat{\beta}_k)$ is unknown or derived from small samples, we may use

$$\hat{\beta}_k = a + b\; \frac{1}{d_k} + \epsilon_k\ ,\ \epsilon_k \sim \mathcal{N}\left(0, \phi\; \left(\frac{1}{d_{k1}} + \frac{1}{d_{k2}}\right)\right) \tag{D-FAV}$$

where FAV stands for *funnel approximate variance.*

For all methods, a 2-tailed t-test with test statistic $\hat{b}/\widehat{\mathrm{SE}}(\hat{b})$ and $(m^{\mathrm{pb}}-2)$ degrees of freedom can be performed to formally assess whether asymmetry occurs.[22] Hereby, it is common to use a 10% level of significance because the number of studies in a meta-analysis is usually low.

3 | EXAMPLES

We illustrate the implementation of aforementioned tests for detecting funnel plot asymmetry in 3 example datasets.

3.1 | Meta-analysis of the association between plasma fibrinogen concentration and the risk of coronary heath disease

The first example dataset composes of 31 studies in which the association between plasma fibrinogen concentration and the risk of coronary heath disease was estimated as a log hazard ratio separately for each study.[23,24] Across the 31 studies, the number of coronary heath disease events ranged from 17 to 1474 and the follow-up ranged from 4 to 33 years. Because of the low number of events, the proportion of censored events was quite high, with a median value of $\pi^{\mathrm{cens}} = 0.96$ (interquartile range , 0.90 to 0.97). The corresponding funnel plots are provided in Figure 1 and suggest that reported hazard ratios are larger for (small) studies with low precision. We subsequently applied the regression models presented in this article to obtain the test statistic for funnel plot asymmetry, ie, $\hat{b}/\mathrm{SE}(\hat{b})$. Afterwards, we applied a 2-tailed t-test with 29 degrees of freedom, yielding a *P* value of <0.01(E-UW), 0.05 (M-FPV), 0.06 (TS), 0.07 (E-FIV), 0.10 (D-FIV), 0.16 (M-FIV), and 0.18 (P-FPV). Hence, given a nominal level of 10%, the presence of small-study effects was (borderline) statistically significant for all tests except M-FIV and P-FPV.

3.2 | Meta-analysis of the effect of erythropoiesis-stimulating agents on overall survival

In 2009, Bohlius et al performed an individual participant data meta-analysis to examine the effect of erythropoiesis-stimulating agents on overall survival in cancer patients.[25] They summarized the hazard ratio of 49 randomized controlled trials that compared epoetin or darbepoetin plus red blood cell transfusions (as necessary) versus red blood cell transfusions (as necessary) alone. Also in this example, the proportion of censored events was quite high, with a median value of $\pi^{\mathrm{cens}} = 0.92$ (interquartile range, 0.85 to 0.96). There was little evidence of between-study heterogeneity ($I^2 = 0\%$), and the fixed effect summary of the hazard ratio was 1.17 with a 95%

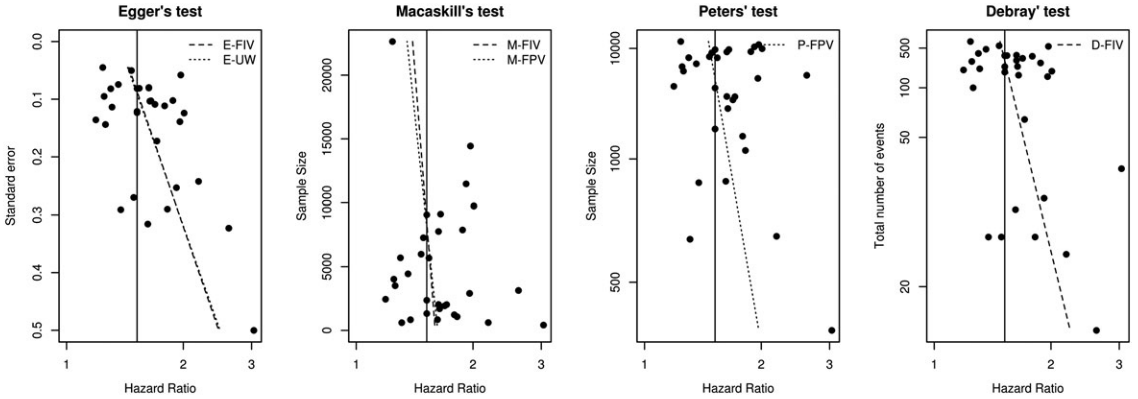

FIGURE 1 Funnel plots for a meta-analysis of the association between plasma fibrinogen concentration and the risk of coronary heath disease ($\pi^{\text{cens}} = 0.96$). The vertical line indicates the fixed effect estimate

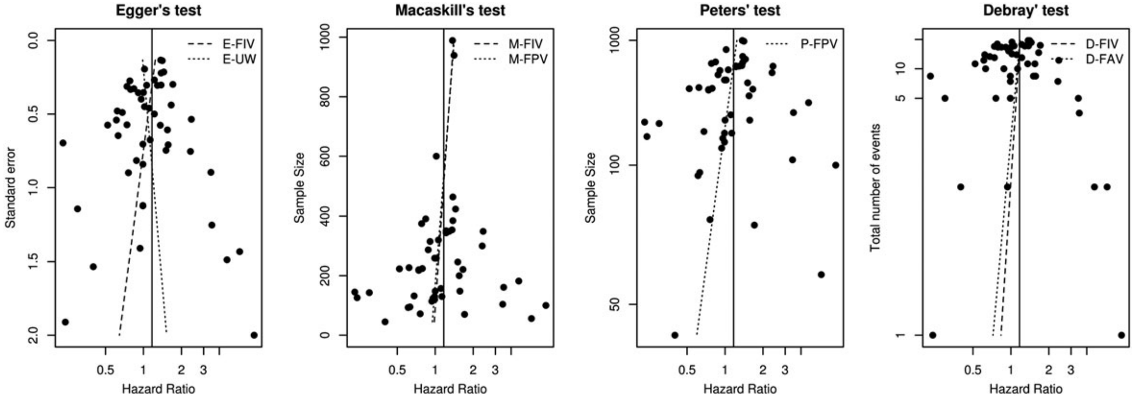

FIGURE 2 Funnel plots for a meta-analysis of the effect of erythropoiesis-stimulating agents on overall survival ($\pi^{\text{cens}} = 0.92$). The vertical line indicates the fixed effect estimate

confidence interval from 1.06 to 1.30. Visual inspection of the funnel plots in Figure 2 did not indicate any apparent asymmetry. Aforementioned funnel plot asymmetry tests yielded a P value of 0.01 (M-FIV and M-FPV), 0.06 (P-FPV), 0.14 (E-FIV), 0.18 (TS), 0.30 (E-UW), 0.50 (D-FAV), and 0.67 (D-FIV). In summary, with exception of the tests proposed by Macaskill and Peters, no evidence was found for the presence of small-study effects.

3.3 | Meta-analysis of adjusted hazard ratios of total stroke for depressed subjects versus nondepressed subjects

In 2011, Pan et al conducted a systematic review and meta-analysis of prospective studies assessing the association between depression and risk of developing stroke in adults.[26] The search yielded 28 prospective cohort studies that reported 8478 stroke cases (morbidity and mortality) during a follow-up period ranging from 2 to 29 years. Most of the results were adjusted for age (25 studies), smoking status (20 studies), body mass index (14 studies), alcohol intake (9 studies), physical activity (7 studies), and comorbidities (23 studies; such as diabetes, hypertension, and coronary heart disease). The median proportion of censored events within studies was $\pi^{\text{cens}} = 0.94$ (interquartile range, 0.91 to 0.98). A moderate extent of heterogeneity was detected in the meta-analysis ($I^2 = 66\%$), and the hazard ratio from the random-effects model was 1.45 (95% confidence interval, 1.29 -1.63). Visual inspection of the funnel plot revealed some degree of asymmetry (Figure 3), and this was confirmed by D-FIV ($P = 0.01$), E-FIV ($P = 0.04$), and TS ($P = 0.06$). Conversely, the presence of small-study effects was not supported by P-FPV ($P = 0.11$), M-FIV ($P = 0.16$), M-FPV ($P = 0.26$), and E-UW ($P = 0.54$).

4 | SIMULATION STUDY

We conducted an extensive simulation study to assess the type-I error rates and power of all aforementioned statistical tests.[27] Hereto, we generated several scenarios where we varied the number of generated trials (m), the number of *published* trials ($m^{\text{pb}} = 10, 20, 50, \text{or } 100$), the size of the true effect ($\exp(\beta) = 1, 0.75, \text{or } 0.5$), the proportion of

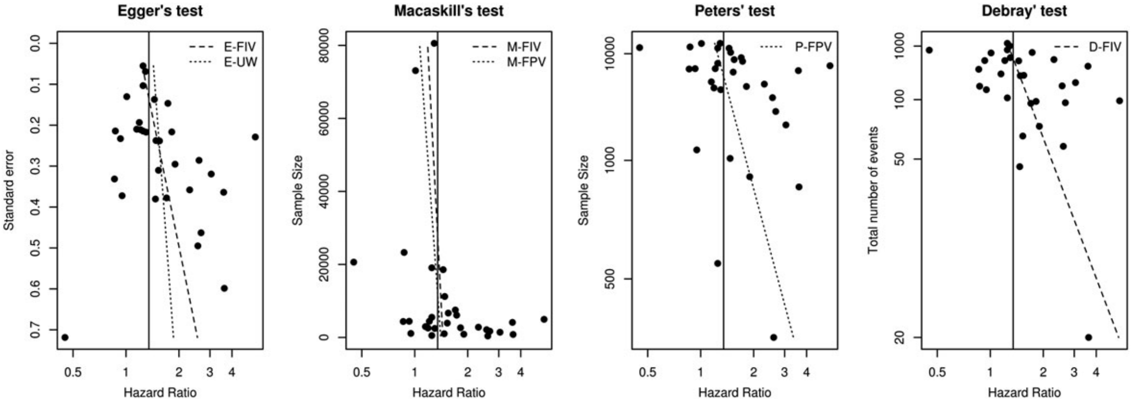

FIGURE 3 Funnel plots for a meta-analysis of adjusted hazard ratios of total stroke for depressed subjects versus nondepressed subjects ($\pi^{\text{cens}} = 0.94$). The vertical line indicates the fixed effect estimate

censored events ($\pi^{\text{cens}} = 0\%, 30\%$, or 90%), and the mechanism of censoring (noninformative versus informative). All scenarios were repeated 10000 times.

4.1 | Data generation

For each trial, the sample size was generated from a log-normal distribution with mean 6 and variance 0.6. This reflects the greater number of small trials compared to large trials as commonly observed in real meta-analyses and results in a mean size of $\exp(6 + 0.6/2) = 545$ subjects per trial and a standard deviation of $\sqrt{\exp(2 \times 6 + 0.6^2) \times (\exp\left(0.6^2\right) - 1)} = 318$. We subsequently generated a survival time for each subject in each trial according to a Weibull distribution[28]:

$$T_{ij} = \left(-\frac{\log(u)}{\lambda \ \exp\left(\beta X_{ij}\right)} \right)^{\frac{1}{\upsilon}} \tag{1}$$

with $u \sim U(0, 1)$, shape parameter $\upsilon > 0$, and a scale parameter $\lambda > 0$. We assumed a fixed treatment effect across trials. We set the probability of receiving a treatment in each trial to 50%, ie, $X \sim$ Bernoulli(0.50), and defined β as a protective treatment effect such that $\exp(\beta) \le 1$. Furthermore, we chose $\lambda = 0.03$ and $\nu = 0.65$ in accordance with trial data from Hodgkin's disease.[29]

4.2 | Dropout of participants

For each trial, we introduced noninformative censoring to mimic random dropout of participants. Hereto, we generated censoring time points L_{ij} from a uniform distribution $U(0, Q)$, where Q was determined by iteration to yield the prespecified censoring proportion π^{cens}.[30] The specific value for Q is then dependent on the choice for λ, υ, β, and π^{cens} (Supporting Information).

As an alternative scenario, we also considered informative right censoring to mimic nonrandom dropout of trial participants. Hereto, we set the observed survival time Z_{ij} equal to

$$Z_{ij} = \begin{cases} T_{ij} & : c_{ij} = 0 \\ L_{ij} \sim U\left(0, T_{ij}\right) & : c_{ij} = 1 \end{cases} \tag{2}$$

with $c_{ij} \sim$ Bernoulli(π^{cens}).

4.3 | Analysis of individual trials

After the introduction of participant dropout, each trial was analyzed using Cox regression to estimate the log hazard ratio β and its corresponding standard error.

4.4 | Introduction of small-study effects

For all scenarios, we varied the number of published trials by considering the following mechanisms of small-study effects:

1. Absence of small-study effects. All generated trials are included in the meta-analysis, such that $m^{\text{pb}} = m$.
2. Presence of small-study effects. A predefined fraction of the generated trials remain unpublished, according to $m^{\text{pb}} = m/1.20$. To determine which trials remain unpublished, we calculated the one-sided P value (given a null hypothesis of $\beta \ge 0$) for each trial and subsequently sorted all trials by their P value in ascending order. The rank for trial j is then given by r_j, which can be used as follows to define the probability w_{r_j} of excluding trial j from the meta-analysis (Supporting Information):

$$w_{r_j} = a^{r_j - 1} \frac{1 - a}{1 - a^m} \tag{3}$$

Because $\sum w_{r_j} = 1$ by definition, we can use the cumulative distribution to iteratively exclude a trial (after which we recalculate w_{r_j} for the remaining trials) until m^{pb} trials remain for meta-analysis. We here choose $a = 1.2$ to ensure that the exclusion probability for trial r_j is 1.20 times higher than the exclusion probability for trial r_{j-1}, which implies that the probability of exclusion modestly increases for trials with larger *P* values. Figure S1 indicates that for $m = 11$ generated trials, the study with the smallest *P* value has an exclusion probability of 3% whereas the study with the largest *P* value has an exclusion probability of 19%. When the number of generated trials increases to $m = 101$, exclusion probabilities range from 2e-7% (for $r_j = 1$) to 17% (for $r_j = 101$).

4.5 | Evaluation of funnel plot asymmetry

For all scenarios, we evaluated the presence of funnel plot asymmetry using aforementioned regression models. When applying M-FPV, P-FPV, D-FIV, or D-FPV, we added 0.5 to d_{k1} and d_{k2} and 2 to n_k for trials with zero cell counts. We subsequently used a 2-tailed t-test with $m^{pb} - 2$ degrees of freedom and a nominal level of 10%.

We estimated the type-I error rate of each test by calculating the total number of *positive* test results in the meta-analyses with absence (ie, no missing studies) of small-study effects. Conversely, for meta-analyses with presence of small-study effects, the total number of *positive* test results was used to estimate the power of each test.

All methods were implemented in R. The corresponding source code is available from the Supporting Information.

5 | RESULTS

For the sake of simplicity, we here focus on funnel plot asymmetry tests that use the inverse variance of the estimated treatment effect as weight in the regression analysis (ie, E-FIV, M-FIV, and D-FIV). We also discuss the performance of TS, as this is the only method adopting an additive heterogeneity component. The corresponding results are presented in Figure 4 (type-I error rates) and Figure 5 (power) for scenarios with noninformative

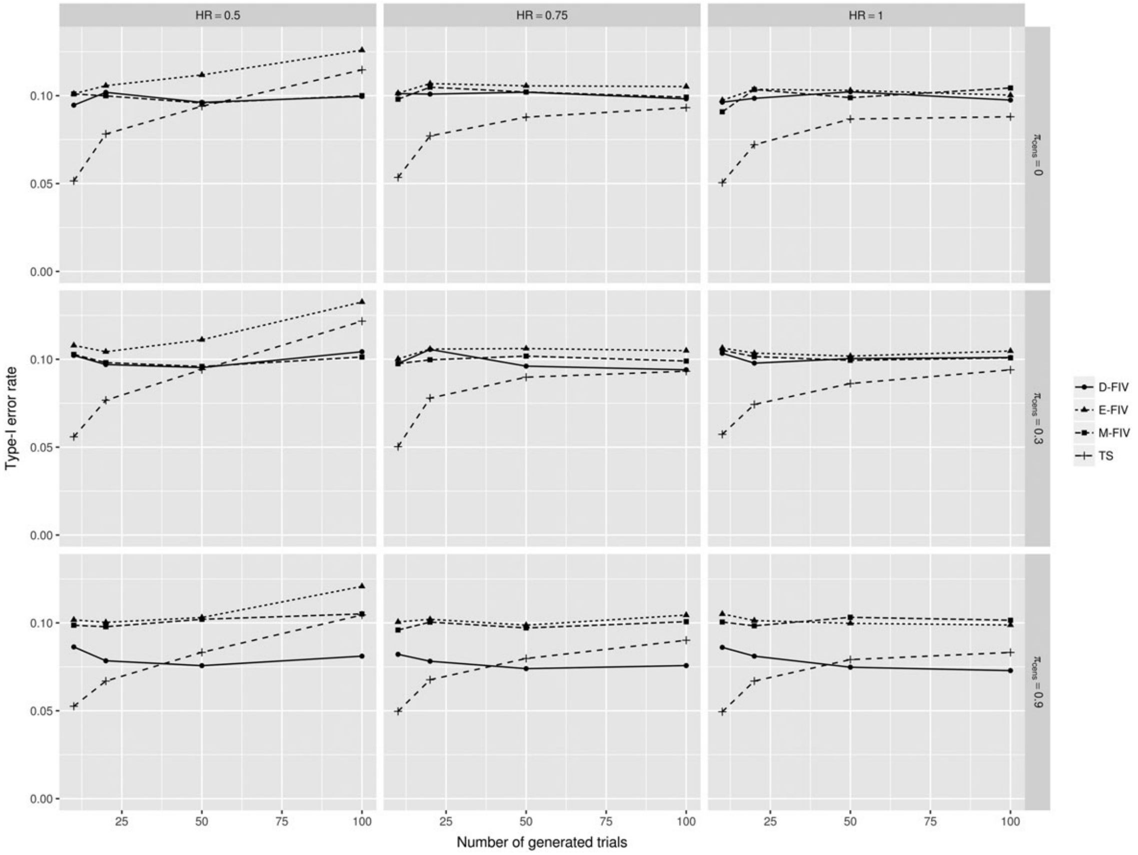

FIGURE 4 Type-I error (false positive) rates in the absence of small-study effects. Results are presented for 3 variations of the true hazard ratio (0.5, 0.75, and 1) and for 3 variations of noninformative participant dropout within trials (values for π^{cens} by row). Results for each scenario are based on 10 000 simulations

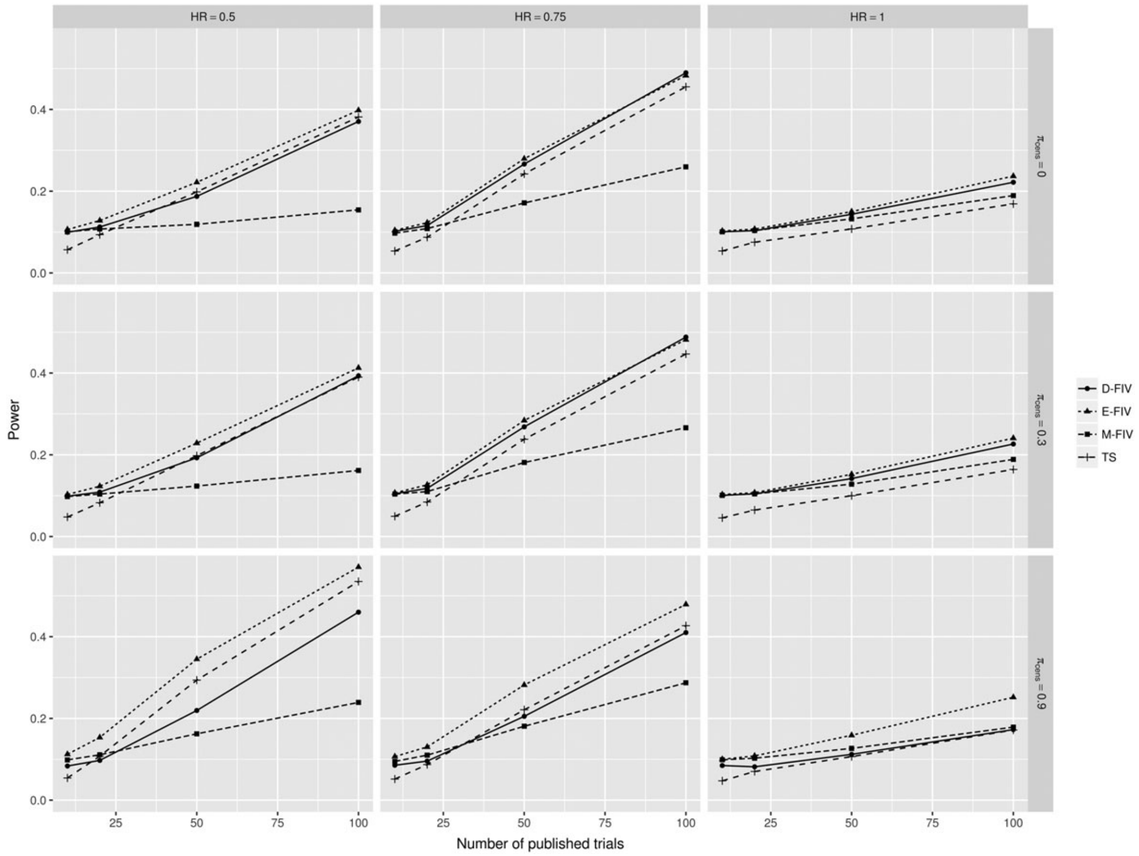

FIGURE 5 Power (true positive) rates in the presence of small-study effects. Results are presented for 3 variations of the true hazard ratio (0.5, 0.75, and 1) and for 3 variations of noninformative participant dropout within trials (values for π^{cens} by row). Results for each scenario are based on 10 000 simulations

dropout of participants. Results for the other tests (E-UW, P-FPV, M-FPV, and D-FAV) and for scenarios involving informative dropout of participants are presented in the Supporting Information.

Conversely, we found that type-I errors were below the nominal level for TS in meta-analyses with few studies and for D-FIV in meta-analyses with substantial participant dropout.

5.1 | Absence of small-study effects

Results in Figure 4 demonstrate that all FIV tests (E-FIV, M-FIV, and D-FIV) yielded appropriate type-I error rates (ie, around 10%) when there was no underlying effect of treatment. In scenarios where treatment efficacy was present, tests that use (estimates of) the standard error as independent predictor variable (E-UW, E-FIV, and TS) tended to yield excessive type-I error rates when meta-analyses included many studies or when studies were affected by informative dropout (Figure S5). For instance, when the true hazard ratio was 0.50 (given $m = 100$, $\pi^{\text{cens}} = 0.90$), the type-I error rate of E-FIV was 12% (noninformative dropout) and, respectively, 24% (informative dropout). Results in Figure S3 further indicate that problematic type-I error rates also occurred for E-UW (with estimates as high as 76%) and, in cases of excessive dropout, for P-FPV and D-FAV.

5.2 | Presence of small-study effects

The power for detecting small-study effects was relatively low for all tests, particularly when few trials were available for meta-analysis (Figure 5). For instance, the power of D-FIV was only 10% when $\exp(\beta) = 1.00$, $\pi^{\text{cens}} = 0$, and $m^{\text{pb}} = 10$. As anticipated, we found that the power of all tests substantially improved as more trials were included in the meta-analysis. The highest power was achieved by E-UW and E-FIV (which also yielded the highest type-I error rates). Results in Figures S4, S6, and S8 further indicate that M-FIV and M-FPV yielded low power across all situations and that TS yielded low power when there was no underlying treatment effect ($\exp(\beta) = 1.00$). Conversely, P-FPV and D-FAV performed relatively well in cases of excessive dropout. Finally, the power of all tests was not much affected by the presence of informative dropout (Supporting Information).

6 | DISCUSSION

Small-study effects are a major concern in systematic reviews as they may signal publication bias, such that small studies with less favorable results are potentially missing from the meta-analysis. For this reason, it is generally recommended to evaluate and formally test funnel plot asymmetry, provided that sufficient studies are available.[4,10] Because most tests for evaluating small-study effects were designed for meta-analysis of odds ratios, we performed a simulation study to verify whether their use is also justified in survival data where hazard ratios are to be synthesized.[21] In particular, because survival data are often affected by censoring, common predictors of funnel plot asymmetry (such as total sample size) may no longer be reliable.

In line with previous findings,[7-9,31] our results demonstrate that the performance of funnel plot asymmetry tests is rather poor. Most tests yield inadequate type-I error rates or suffer from low power, also when applied to meta-analysis of survival data. Although Egger's tests (E-UW and E-FIV) achieve the highest power, their type-I error rates are too high, particularly when many studies are available for meta-analysis or when they are affected by informative dropout. Conversely, other existing tests with appropriate type-I error rates tend to have poor power. This clearly casts doubt about the clinical utility of funnel plot asymmetry tests for assessing the presence of small-study effects in meta-analyses of survival data. We therefore developed a novel test that yields higher power than existing tests with appropriate type-I error rates. Our novel test D-FIV is loosely based on the test proposed by Peters et al[22] but adopts different study weights and uses the inverse of the number of events, rather than the inverse of the total sample size. Although D-FIV and Peters' test performed very similar in the absence of participant dropout, D-FIV yields more favorable type-I error rates in the presence of censoring. For this reason, its use appears more appropriate when dealing with time-to-event data.

Although D-FIV and D-FAV are not designed for meta-analyses with binary outcomes, they may offer an appealing choice when observed event rates are close to 1. In particular, results from the simulation study demonstrate that P-FPV and M-FPV become problematic when all participants experience an event (as study weights are no longer identifiable) and that other tests suffer from low power (M-FPV and M-FIV) or inappropriate type-I error rates (E-UW and E-FIV). In most situations, however, D-FIV, D-FAV, and P-FPV are likely to perform similarly when participant dropout is not an issue as the total number of events is then strongly related to the total sample size (provided that baseline risk and treatment effects do not vary much across studies).

Because it has been argued that the statistical rationale for a multiplicative variance inflation factor is rather weak,[14] further improvements of D-FIV and D-FAV are possible by considering an additive between-study heterogeneity component (as implemented by TS). Recent simulation studies suggest that TS-related models perform relatively well in the presence of between-study heterogeneity,[31,32] but have limited power when the extent of between-study heterogeneity is low. Our results indicate that in the absence of heterogeneity, TS essentially trades power to reduce type-I error rates. This effect is rather problematic, as the power of most funnel plot asymmetry tests is already very low. Furthermore, the implementation of TS-related models requires careful thought with respect to distributional assumptions and estimation methods.[15]

Previously, Sterne et al recommended 10 or more studies for testing for funnel plot asymmetry.[4] However, results from our simulation study indicate that even when many ($\geq$ 50) studies are available for meta-analysis, the power for detecting the presence of small-study effects usually remains below 50%. The performance of the presented funnel plot asymmetry tests may further deteriorate when studies are relatively small or have limited follow-up (and continuity corrections are needed)[33] or when larger studies are conducted more often in settings with small event probabilities. For this reason, it will often be necessary to explore alternative strategies to address whether reporting biases are of any concern. Suggestions for this have recently been proposed.[4,10,12]

The methods discussed in this paper merely test for the presence of funnel plot asymmetry. Several authors have discussed the implementation of alternative methods that attempt to *correct* meta-analyzed estimates of treatment effect for the presence of small-study effects.[34-37] These methods make different assumptions about the mechanisms of selective reporting and can be applied even when there is no evidence of funnel plot asymmetry. However, it has also been shown that in the strong presence of small-study effects, regression-based approaches may still be preferred.[38] In any case, the methods presented in this paper should not be used for assessing the quality of a meta-analysis, but rather to explore the presence of small-study effects and to facilitate the conduct of sensitivity analyses to the potential impact of publication bias. Further, it is important to recognize that small-study effects may rather be caused by heterogeneity than by publication bias.[39] Although we did not generate heterogeneous hazard ratios in our simulation study, recent studies have demonstrated that the performance of most funnel plot asymmetry tests deteriorates when reported effect sizes are substantially heterogeneous.[15,32]

In conclusion, when examining the presence of small-study effects in meta-analyses of hazard (rate) ratios, we recommend the use of our novel test D-FIV. Our test is loosely related to Peters' regression test but achieves better type-I error rates when studies are affected by participant dropout. However, because funnel plot asymmetry does not necessarily arise because of small-study effects, and because proper testing of funnel plot asymmetry requires access to many studies, their implementation is no panacea against selective reporting.[4] Further, with few studies available for meta-analysis, all tests have very low power for asymmetry detection of hazard ratios and thus are best avoided.

ACKNOWLEDGEMENTS

We thank Prof Paul Lambert for his valuable comments on how to simulate individual participant survival data. We also thank the editors and anonymous reviewers for their constructive feedback on this work. We gratefully acknowledge the financial support by the Netherlands Organization for Health Research and Development (91617050).

ORCID

Thomas P. A. Debray http://orcid.org/0000-0002-1790-2719

REFERENCES

1. Rothstein H, Sutton AJ, Borenstein M. *(Eds.) Chapter 10. Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester, England; Hoboken, NJ: Wiley; 2005.
2. Ioannidis JP, Contopoulos-Ioannidis DG, Lau J. Recursive Cumulative Meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. *J Clin Epidemiol*. 1999;52:281-291.
3. Sterne JAC, Harbord RM. Funnel plots in meta-analysis. *The Stat J*. 2004;4:127-141.
4. Sterne JAC, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*. 2011;343:d4002.
5. Deeks JJ, Higgins JPT, Altman DG. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JP, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 5.1.0, Vol. 9: Cochrane; 2011. www.handbook.cochrane.org.
6. Jin Z-C, Zhou X-H, He J. Statistical methods for dealing with publication bias in meta-analysis. *Stat Med*. 2014;34:343-360.
7. 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;9:2.
8. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med*. 2001;20:641-654.
9. 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.
10. Sterne JAC, Egger M, Moher D. Chapter 10: Addressing reporting biases. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 5.1.0, Vol. 10: Cochrane; 2011. www.cochrane-handbook.org.
11. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315:629-634.
12. Song F, Parekh S, Hooper L, et al. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*. 2010;14. https://doi.org/10.3310/hta14080
13. 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*. November 2000;53:1119-1129.
14. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med*. 1999; 18:2693-2708. https://doi.org/10.1002/(SICI)1097-0258(19991030)18:20<2693::AID-SIM235>3.0.CO;2-V
15. Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med*. 2006;25:3443-3457. https://doi.org/10.1002/sim.2380
16. Irwig L, Macaskill P, Berry G, Glasziou P. Bias in meta-analysis detected by a simple, graphical test. Graphical test is itself biased. *BMJ*. 1998;316:470-471.
17. Cao X. Relative performance of expected and observed {Fisher} information in covariance estimation for maximum likelihood estimates. *Ph.D. Thesis*: Johns Hopkins University. Baltimore, Maryland; 2013.
18. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *J Clin Epidemiol*. 1995;48:1495-1501.
19. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995;48:1503-1510.
20. Gart JJ, Zweifel JR. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*. 1967;54:181-187.
21. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med*. 1998;17:2815-2834. https://doi.org/10.1002/(SICI)1097-0258(19981230)17:24<2815::AID-SIM110>3.0.CO;2-8
22. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of tests and adjustments for publication bias in the presence of heterogeneity. Technical Report 05-01, Leicester, England: Dept of Health Sciences, University of Leicester; 2005.
23. Thompson S, Kaptoge S, White I, et al. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *Int J Epidemiol*. 2010;39:1345-1359. https://doi.org/10.1093/ije/dyq063
24. Fibrinogen Studies Collaboration. Collaborative meta-analysis of prospective studies of plasma fibrinogen and cardiovascular disease. *Eur J Cardiovasc Prev Rehabil*. 2004;11:9-17.
25. Bohlius J, Schmidlin K, Brillant C, et al. Erythropoietin or Darbepoetin for patients with cancer–meta-analysis based on individual patient data. *Cochrane Database Syst Rev*. 2009:CD007303. https://doi.org/10.1002/14651858.CD007303.pub2
26. Pan A, Sun Q, Okereke OI, Rexrode KM, Hu FB. Depression and risk of stroke morbidity and mortality: A meta-analysis and systematic review. *JAMA*. 2011;306:1241-1249.

27. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25:4279-4292. https://doi.org/10.1002/sim.2673
28. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24:1713-1723. https://doi.org/10.1002/sim.2059
29. Feuer EJ, Kessler LG, Baker SG, Triolo HE, Green DT. The impact of breakthrough clinical trials on survival in population based tumor registries. *J Clin Epidemiol*. 1991;44:141-153.
30. Qian J, Li B, Chen P-Y. Generating Survival Data in the Simulation Studies of Cox Model. IEEE; June 2010:93-96. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5514026, https://doi.org/10.1109/ICIC.2010.294
31. Jin Z-C, Wu C, Zhou X-H, He J. A modified regression method to test publication bias in meta-analyses with binary outcomes. *BMC Med Res Methodol*. 2014;14:132.
32. Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. *Stat Med*. 2008;27:746-763. https://doi.org/10.1002/sim.2971
33. Schwarzer G, Antes G, Schumacher M. Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Stat Med*. 2002;21:2465-2477. https://doi.org/10.1002/sim.1224
34. McShane BB, Bckenholt U, Hansen KT. Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspect Psychol Sci*. 2016;11:730-749. https://doi.org/10.1177/1745691616662243
35. Mavridis D, Sutton A, Cipriani A, Salanti G. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Stat Med*. 2013;32:51-66. https://doi.org/10.1002/sim.5494
36. Sutton AJ, Song F, Gilbody SM, Abrams KR. Modelling publication bias in meta-analysis: A review. *Stat Methods Med Res*. 2000;9:421-445.
37. Copas JB. What works?: Selectivity models and meta-analysis. *J R Stat Soc Ser A Stat Soc*. 1999;162:95-109. https://doi.org/10.1111/1467-985X.00123
38. Rücker G, Carpenter JR, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. *Biom J*. 2011;53:351-368. https://doi.org/10.1002/bimj.201000151
39. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Stat Med*. 2003;22:2113-2126. https://doi.org/10.1002/sim.1461

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Debray TPA, Moons KGM, Riley RD. Detecting small-study effects and funnel plot asymmetry in meta-analysis of survival data: A comparison of new and existing tests. *Res Syn Meth*. 2018;9:41–50. https://doi.org/10.1002/jrsm.1266