Received: 4 December 2019 | Revised: 15 March 2020 | Accepted: 16 March 2020

DOI: 10.1002/bimj.201900379

Biometrical Journal

RESEARCH PAPER

Comparison of random-effects meta-analysis models for the relative risk in the case of rare events: A simulation study

Marie Beisemann[1] | Philipp Doebler[1] | Heinz Holling[2]

[1] Faculty of Statistics, TU Dortmund University, Dortmund, Germany

[2] Faculty of Psychology and Sports Sciences, University of Münster, Münster, Germany

Correspondence
Marie Beisemann, Faculty of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany.
Email: beisemann@statistik.tu-dortmund.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: HO 1286/16-1

Correction added on 19 January 2021, after first online publication: Projekt Deal funding statement has been added.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

Abstract

Pooling the relative risk (RR) across studies investigating rare events, for example, adverse events, via meta-analytical methods still presents a challenge to researchers. The main reason for this is the high probability of observing no events in treatment or control group or both, resulting in an undefined log RR (the basis of standard meta-analysis). Other technical challenges ensue, for example, the violation of normality assumptions, or bias due to exclusion of studies and application of continuity corrections, leading to poor performance of standard approaches. In the present simulation study, we compared three recently proposed alternative models (random-effects [RE] Poisson regression, RE zero-inflated Poisson [ZIP] regression, binomial regression) to the standard methods in conjunction with different continuity corrections and to different versions of beta-binomial regression. Based on our investigation of the models' performance in 162 different simulation settings informed by meta-analyses from the Cochrane database and distinguished by different underlying true effects, degrees of between-study heterogeneity, numbers of primary studies, group size ratios, and baseline risks, we recommend the use of the RE Poisson regression model. The beta-binomial model recommended by Kuss (2015) also performed well. Decent performance was also exhibited by the ZIP models, but they also had considerable convergence issues. We stress that these recommendations are only valid for meta-analyses with larger numbers of primary studies. All models are applied to data from two Cochrane reviews to illustrate differences between and issues of the models. Limitations as well as practical implications and recommendations are discussed; a flowchart summarizing recommendations is provided.

KEYWORDS
beta-binomial regression, (zero-inflated) Poisson regression, random-effects meta-analysis, rare events, relative risk

1 | INTRODUCTION AND BACKGROUND

Meta-analysis is a common and popular statistical tool for research synthesis that hardly needs an introduction at this point (Jackson, Law, Stijnen, Viechtbauer, & White, 2018). Of interest in the present paper is the research synthesis of count data in settings, where researchers study the occurrence of an event that is rare. We refer to an event as rare when it has a very low event occurrence probability that results in only very few to no observations of the respective event in a study, even if observation times and experimental groups are not short/small (Böhning, Mylona, & Kimber, 2015). This topic has recently received an increasing amount of attention in the statistical literature (e.g., Böhning et al., 2015; Efthimiou, 2018; Jackson et al., 2018; Kuss, 2015). A common example of rare events is adverse side effects to treatment, for example, cardiovascular mortality in response to medications, such as prophylactics (Squizzato, Bellesini, Takeda, Middeldorp, & Donadini, 2017) or anti-inflammatory drugs (Hemkens et al., 2016). These two Cochrane reviews shall serve as examples in the present work. Excerpts of the data for the outcomes cardiovascular mortality and fatal stroke from these two reviews are shown in Table 1 to give the reader an impression of the data setting discussed in the present work (four studies for each outcome shown). Meta-analysis is especially important in rare-events contexts, as it might constitute the only means by which researchers may obtain reliable evidence for the clinical phenomena they are studying (Higgins & Green, 2011). Meta-analyses may also serve as a basis for decisions on medication approval, making reliable statistical methods all the more necessary. For an overview and methodological evaluation of such applications, see Warren, Abrams, Golder, and Sutton (2012). Typically, primary studies examine how often an event of interest occurred in either one of two experimental groups (i.e., treatment and control group), yielding 2 × 2 frequency tables like Table 2. Then, univariate summary statistics, such as the odds ratio (OR) or the relative risk (RR), are calculated. These in turn are then pooled in the meta-analysis. Among meta-analysis models, we distinguish between fixed- and random-effects (RE) approaches. While the former assumes one single "true" effect underlying all studies, with differences in effects between studies being only due sampling error, the

TABLE 1 Excerpts from the data from three meta-analyses conducted in the context of two Cochrane reviews (Hemkens et al., 2016; Squizzato et al., 2017) that were used as examples to illustrate the models compared in this simulation study (from each meta-analysis, four included primary studies are shown)

Squizzato et al. (2017): Cardiovascular mortality				
Study ID*	Y_{i1}	Y_{i2}	n_{i1}	n_{i2}
STD-CURE-2001	318	345	6,259	6,303
STD-Vavuranakis-2006	1	1	43	43
STD-CASCADE-2010	0	1	56	57
STD-CRYSSA-2012	1	2	150	150
Hemkens et al. (2016): Cardiovascular mortality				
Study ID*	Y_{i1}	Y_{i2}	n_{i1}	n_{i2}
STD-Nidorf-2013	1	10	282	250
STD-Kaplan-1986	1	0	28	29
STD-Kershenobich-1988	0	2	54	46
STD-Parise-1995	0	0	21	20
Hemkens et al. (2016): Fatal stroke				
Study ID*	Y_{i1}	Y_{i2}	n_{i1}	n_{i2}
STD-Deftereos-2013	1	0	112	110
STD-Nidorf-2013	0	0	282	250
STD-Parise-1995	0	0	21	20
STD-Yurdakul-2001	0	0	60	60

Note. *Study ID as used in the respective review.

TABLE 2 Illustration of a 2×2 Frequency Table for a primary study *i*

Group	Event	No event	Total
Treatment	Y_{i1}	$n_{i1} - Y_{i1}$	n_{i1}
Control	Y_{i2}	$n_{i2} - Y_{i2}$	n_{i2}
Total	$n_{event,i}$	$n_{no\ event,i}$	$n_{total,i}$

latter assumes that underlying true effects differ systematically between primary studies additionally to sampling error. In real-world applications, an RE approach is commonly considered more adequate (Bai, Chen, & Wang, 2016; Jackson, Law, Stijnen, Viechtbauer, & White, 2018) and therefore constitutes the focus of this paper.

Rare events present a challenge in this context for multiple reasons, specifically when they result in the lack of observations of the event of interest in either one of treatment or control group (*single-zero study*) or in both (*double-zero study*). First, zero studies leave estimators of measures like the RR (formally defined as RR $= P(\text{event} \mid \text{treatment group}) \,/\, P(\text{event} \mid \text{control group})$, and estimated by $\widehat{\mathrm{RR}_i} = (Y_{i1}/n_{i1}) \,/\, (Y_{i2}/n_{i2})$, using the notation introduced in Table 2) as either 0 or undefined; their logarithm is undefined in any case that causes a problem for the standard meta-analytical approach, which pools the log RR rather than the RR. The same issue applies to the OR (see Böhning et al., 2015, for more details). A related problem that also concerns the standard model of meta-analysis, commonly referred to as the inverse-variance model, is the following: The inverse-variance method computes a weighted average of the log $\widehat{\mathrm{RR}_i}$ estimated for the primary studies included in the meta-analysis. With an RE approach, the weight for study i is given by $1/(\tau^2 + s_i^2)$, with τ^2 denoting the between-study variance in the underlying effects, and s_i^2 denoting the sampling variance, estimated as $\widehat{s_i^2} = 1/Y_{i1} - 1/n_{i1} + 1/Y_{i2} - 1/n_{i2}$. The estimator for s_i^2 is based on the assumption that the sample distribution of log $\widehat{\mathrm{RR}_i}$ is approximately normal (Jackson & White, 2018). However, this assumption is often violated in the context of rare events (Friede, Röver, Wandel, & Neuenschwander, 2017; Jackson & White, 2018). A general criticism of the inverse-variance method is that this method treats the estimates of τ^2 and s_i^2 as if they were the true values in the computation of the meta-analytical weights (Bakbergenuly & Kulinskaya, 2018; Jackson & White, 2018; Malzahn, Böhning, & Holling, 2000; Stijnen, Hamza, & Özdemir, 2010). The common approaches to enable the computation of a meta-analysis with the inverse-variance method in the presence of single- or double-zero studies are either the exclusion of the problematic primary studies or the application of a continuity correction, either using a fixed (standard correction, often 0.5; Higgins & Green, 2011; Cox, 1970) or a variable value (alternative correction; Sweeting, Sutton, & Lambert, 2004, see also de Rooi, 2008, for another alternative correction). Both approaches (exclusion and correction) have been heavily criticized as they might introduce bias (Friedrich, Adhikari, & Beyene, 2007; Kuss, 2015; Sankey, Weissfeld, Fine, & Kapoor, 1996), and the latter is also warned against by the Cochrane collaboration (Higgins & Green, 2011). In general, authors are usually in unison when discouraging the use of the inverse-variance method in the context of rare events (Bradburn, Deeks, Berlin, & Russell Localio, 2007; Sweeting, Sutton, & Lambert, 2004; Tang, 2000), or at least recommend robustness assessment due to bias in the method (Keus, Wetterslev, Gluud, Gooszen, & Van Laarhoven, 2009).

Thus, it is unsurprising that research into alternative RE approaches of pooling the RR in a setting with rare events has recently grown. So far, a focus on the OR as an effect measure (e.g., Bhaumik et al., 2012; Bradburn, Deeks, Berlin, & Russell Localio, 2007; Cheng, Pullenayegum, Marshall, Iorio, & Thabane, 2016; Jackson, Law, Stijnen, Viechtbauer, & White, 2018; Li, Bai, & Wang, 2018; Sankey, Weissfeld, Fine, & Kapoor, 1996; Shuster, 2010; Sweeting, Sutton, & Lambert, 2004) and fixed-effects approaches (Bradburn et al., 2007; Sweeting et al., 2004 for the OR; Guevara, Berlin, & Wolf, 2004; Spittal, Pirkis, & Gurrin, 2015 for the incidence rate ratio, extendable to the RR; Böhning et al., 2015) has prevailed. For an overview of existing methods and recommendations—also extending to other effect measures not considered here, such as the risk difference (RD) or the arcsine difference (Rücker, Schwarzer, Carpenter, & Olkin, 2009)— please see Efthimiou (2018), Jackson et al. (2018), Lane (2013), or Kuss (2015). However, considering that approximately one third of adverse events meta-analyses alone use the RR as an effect measure (Warren et al., 2012), the development of appropriate methods for the RR seems much warranted. Providing researchers with appropriate and updated guidelines is especially important as the Cochrane collaboration focuses their recommendations on the OR (Higgins & Green, 2011). Their recommendations are also mainly fixed-effects models. However, RE models for the RR are available. A recent paper by Böhning et al. (2015) discussed several models to this end and demonstrated them using a real-life example. The aim of the present paper is to extend their work by comparing the models discussed in their paper to other proposed alternatives and the standard meta-analysis approach (i.e., the inverse-variance method). Before discussing the models we included in our comparison, we are going to give an overview of the existing literature: Even though there is a RE version of the Mantel–Haenszel (MH) method (Mantel & Haenszel, 1959), it is merely the fixed-effects MH estimator embedded into the inverse-variance method, thus subject to the same limitations (Efthimiou, 2018) and so not considered here. RE Poisson regression models have been suggested for pooling incidence rate ratios (Guevara et al., 2004; Spittal et al., 2015; Stijnen et al., 2010). The model is also applicable to the RR and discussed as one of the models in Böhning et al. (2015). A comparable, but supposedly numerically more stable approach has been recommended by Cai, Parast, and Ryan (2010) (a gamma RE Poisson regression model). Böhning et al. (2015), the first to our knowledge, also suggested to use an extension of the RE Poisson regression model: a (RE) zero-inflated Poisson (ZIP) regression model (Lambert, 1992). Promisingly, an analogous model for the OR in a fixed-effects setting, that is, zero-inflated binomial regression, has shown good performance (Dong, Zhao, & Tiwari, 2019). Based on an extensive simulation study, Kuss (2015) recommended beta-binomial regression using a log link for pooling the RR (and also the OR and the RD, using different link functions). His recommendation is based on a comparison including the models proposed by Stijnen et al. (2010) and Cai et al. (2010) that—taking into consideration all measures of performance administered in this study—were out-performed by the beta-binomial model. Other proposed models also discussed by Böhning et al. (2015) include binomial RE models that predict the number of events in the treatment group conditioned on the total number of events observed across both groups (Stijnen et al., 2010). Again, Cai et al. (2010) have suggested an alternative version of this model that has a closed-form likelihood. Yet, it should be pointed out that these models warrant the exclusion of double-zero studies.

As previously mentioned, the aim of the present simulation study with a strong focus on the RR was to extend the work by Böhning et al. (2015). We compared the RE models they discussed to other existing models in different rare-events settings of meta-analysis of randomized trials. An important contribution of this simulation study is that it includes the evaluation of the newly suggested application of the zero-inflated Poisson regression model (Böhning et al., 2015). The models were compared in terms of convergence rates as well as mean and median bias, root mean squared error (RMSE), mean absolute error and maximum absolute error of their pooled (log) RR estimates, and coverage of the respective 95% confidence intervals (CIs). Models that had already been included in a previous study by Kuss (2015) and had been clearly outperformed by the model Kuss (2015) ended up recommending were not included again to avoid repetition (except for the Poisson regression model as it was also discussed in Böhning et al., 2015). We derived standard errors for some of the pooled effect estimators, as well as in one case even the estimator itself (based on the ZIP regression model and considerations by Dong et al., 2019). This may be considered another contribution of the present study. To limit the extensive scope of our undertaking, we restricted our investigation to frequentist RE approaches and comparable models (but please see Shuster, Guo, & Skyler, 2012; Shuster & Walker, 2016, for a discussion of the distinction between different variants of RE and the resulting implications for employed methods). Bayesian methods were excluded from the current investigation mainly for two reasons: (a) arguably, they are still less popular in meta-analytical work, at least the vast majority of Cochrane reviews do not employ them, and (b) more importantly, typical Bayesian methods employ Markov chain Monte Carlo (MCMC) methods for parameter estimation, which are often more computer intensive and would have severely limited the number of replications of the simulations. Even though some of the models, for example, the Poisson regression model (see Section 2 for details), are able to account for varying person-time across studies, we do not consider scenarios with varying person-time across studies here. We further concentrated on a setting free of any further methodological issues likely to occur in the context of rare-events studies in our simulation study, such as publication bias or outcome reporting bias.

The paper is organized as follows. In Section 2, we describe the models compared in this simulation study. In Section 3, we describe the conducted Monte Carlo simulation study and in Section 4, we present the respective results. An example that illustrates model performance using data from two different Cochrane reviews is presented in Section 5. We conclude with a discussion of our findings and their implications as well as give practical recommendations in Section 6.

2 | MODELS

2.1 | RE Poisson model

Poisson regression models with RE are a member of the family of generalized linear mixed models (GLMM). They are used to predict count data. For the model proposed by Böhning et al. (2015) (see also Spittal et al., 2015; Stijnen et al., 2010), we assume that the number of observed events of interest Y_{ij} in study i and group j follows a Poisson(λ_{ij}) distribution. The expected value can be written as $\lambda_{ij} = E(Y_{ij}) = \mu_{ij} n_{ij}$, where μ_{ij} is the incidence rate in group j of the ith study and n_{ij} is the number of total observational units (i.e., subjects) in group j of study i. The incidence rate μ_{ij} in group j of study i is thus given by $\mu_{ij} = E(Y_{ij})/n_{ij}$. Applying the log function to $\lambda_{ij} = E(Y_{ij}) = \mu_{ij} n_{ij}$, yields $\log \lambda_{ij} = \log E(Y_{ij}) = \log \mu_{ij} n_{ij} = \log n_{ij} + \log \mu_{ij}$. As common in regression models, we may predict $\log \mu_{ij}$ by a linear combination, yielding

$$\log \lambda_{ij} = \log E(Y_{ij}) = \log \mu_{ij} n_{ij} = \log n_{ij} + \log \mu_{ij} = \log n_{ij} + \alpha_i + \beta_i X_{i1}, \tag{1}$$

which describes the Poisson regression model (using a log link) to predict the number of events of interest in group j of study i. X_{i1} is a dummy-coded variable with $X_{i1} = 1$ for the treatment group, and else 0 (control group). Note that the multiplicative relationship between μ_{ij} and n_{ij} translates into an additive relationship on the log scale. Consequently, an offset of $\log n_{ij}$ is included in our regression equation on the log scale. This offset allows for accounting for different group sizes in different studies. In fact, it actually allows to even model different group sizes in conjunction with different observation times for the groups and/or studies. To simplify matters, we do not further discuss this option here but please see Böhning et al. (2015) for more details. Further, note the subscripts on the intercept parameter α_i and the slope parameter β_i, indicating that these parameters differ across studies. We assume that $\alpha_i \sim \text{Normal}(\alpha, \sigma_\alpha^2)$ and $\beta_i \sim \text{Normal}(\beta, \sigma_\beta^2)$, modeling both regression coefficients as RE. The likelihood of the model is given by

$$L_m(\boldsymbol{\theta} \mid Y_{ij}, X_{i1}, n_{ij}) = \prod_{i=1}^{m} \int \text{Po}(y_{i2} \mid \lambda_{i2}) \times \left[\int \text{Po}(y_{i1} \mid \lambda_{i1}) \phi(\beta_i \mid \beta, \sigma_\beta^2) d\beta_i \right] \phi(\alpha_i \mid \alpha, \sigma_\alpha^2)\, d\alpha_i \tag{2}$$

with m denoting the number of studies included in the meta-analysis, $\boldsymbol{\theta} = [\alpha, \beta, \sigma_\alpha^2, \sigma_\beta^2]'$ denoting the vector of unknown parameters to be estimated in the model, and $\text{Po}(y_{ij} \mid \lambda_{ij})$ representing the (discrete) density of the Poisson distribution, that is,

$$\text{Po}(y_{ij} \mid \lambda_{ij}) = \frac{\exp(-\lambda_{ij}) \lambda_{ij}^{y_{ij}}}{y_{ij}!}.$$

The intercept α_i of the model, representing the log risk in the control group (i.e., the log baseline risk), is modeled as a random effect to account for different baseline risks across studies. The slope β_i represents the log RR, with the fixed effect β representing the pooled log RR (i.e., our parameter of interest). Here, we have the option of modeling the log RR as constant across studies, yielding a fixed-effects model of meta-analysis, or, as we have done here, we can assume that the log RR varies across studies. Please note that the distinction between the fixed- and random-effects meta-analytical model is made in reference to the slope. In both cases, the intercept is modeled as a random effect.

2.2 | Zero-inflated Poisson models

As mentioned above, in a rare events setting, one of the crucial problems with conducting meta-analysis is the occurrence of zero events. Although the Poisson model is principally able to handle single- as well as double-zero studies, it may not fit the data as well when we observe more zeros than we would expect under the Poisson distribution: an excess of zero counts. This problem can be alleviated by explicitly modeling the excess of zeroes using a variant of the Poisson model called ZIP regression (Böhning et al., 2015; Lambert, 1992). ZIP regression models handle zero inflation by modeling a two-component process for data generation: the observed count is an excess zero with probability π_{ij} and a realization of a Poisson-distributed random variable with probability $1 - \pi_{ij}$. As in Equation (2), $\text{Po}(y_{ij} \mid \lambda_{ij})$ denotes the (discrete) density of the Poisson distribution. The ZIP model then predicts the excess-zero probability π_{ij} using logistic regression as well as λ_{ij}, the expectation of the Poisson distribution, using the same Poisson regression model as described in Equation (1). More formally, we write

$$\begin{aligned} Y_{ij} &= 0, \quad \text{with probability } \pi_{ij} + (1 - \pi_{ij})\,\text{Po}(0 \mid \lambda_{ij}) \\ &= y, \quad \text{with probability} (1 - \pi_{ij})\,\text{Po}(y \mid \lambda_{ij}) \end{aligned}$$

with $y = 1, 2, \ldots$. Further, λ_{ij} and π_{ij} satisfy

$$\log \lambda_{ij} = \log \mu_{ij} n_{ij} = \log n_{ij} + \log \mu_{ij} = \log n_{ij} + \alpha_i + \beta_i X_{i1}$$

with $\alpha_i \sim \text{Normal}(\alpha, \sigma_\alpha^2)$ and $\beta_i \sim \text{Normal}(\beta, \sigma_\beta^2)$, as well as

$$\text{logit}\, \pi_{ij} = \log(\pi_{ij}/(1 - \pi_{ij})) = \eta_{ij} \tag{3}$$

with η_{ij}denoting the linear prediction term that models logit π_{ij}. In the estimation of this model, we would maximize the product of the likelihood of the observation Y_{ij}, integrating over the RE, analogously to what we have shown in Equation (2) for the Poisson model. We generically use η_{ij} in Equation (3) as we included several versions of this ZIP model in our simulations differentiated by different η_{ij}. Böhning et al. (2015) presented and illustrated a fixed-effects ZIP model (see eqs. 10 and 11 in Böhning et al., 2015) in which both λ_{ij} and π_{ij} were predicted by an intercept and the indicator of experimental group membership (treatment or control). A full RE model is given by substituting $\alpha_i' + \beta_i' X_{i1}$ (with $\alpha_i' \sim \text{Normal}(\alpha', \sigma^2_{\alpha'})$ and $\beta_i' \sim \text{Normal}(\beta', \sigma^2_{\beta'})$) for η_{ij} in Equation (3). Preliminary trial simulations showed that this model was practically not viable as it hardly ever converged in any of our simulation conditions. Alternatively, we could model the treatment effect in the zero-inflation arm as a fixed effect, in conjunction with either (a) a fixed, or (b) a random intercept, yielding either (1) $\eta_{ij} = \alpha_i' + \beta' X_{i1}$ with $\alpha_i' \sim \text{Normal}(\alpha', \sigma^2_{\alpha'})$, or (2) $\eta_{ij} = \alpha' + \beta' X_{i1}$. It is also possible that our model would not benefit considerably from modeling a treatment effect in the zero-inflation arm. If that were the case, merely modeling a random intercept ($\eta_{ij} = \alpha_i'$, with $\alpha_i' \sim \text{Normal}(\alpha', \sigma^2_{\alpha'})$) or possibly just a fixed intercept ($\eta_{ij} = \alpha'$) might suffice.

An issue that arises with employing the ZIP model is that if we include a treatment effect in the zero-inflation arm of our model, then our parameter of interest, the RR (or rather the log RR), is no longer a parameter in the model. The (fixed effect of the) slope for the experimental group, β, (in the Poisson-arm of the model) represents now the pooled log RR conditional on being in the Poisson-arm of the model (Dong et al., 2019). For their fixed-effects zero-inflation binomial model, Dong et al. (2019) have shown how an "unconditional" pooled OR can be obtained from the estimated model parameters. Adapting their work to the ZIP model, we can arrive at an "unconditional" pooled RR. It is

$$\begin{aligned}\text{RR}_{pooled} &= \frac{E(Y_1)/n_1}{E(Y_2)/n_2} = \frac{((1 - \text{expit}(\alpha' + \beta'))\exp(\alpha + \beta + \log n_1))/n_1}{((1 - \text{expit}(\alpha'))\exp(\alpha + \log n_2))/n_2} \\ &= \frac{1 + \exp(\alpha')}{1 + \exp(\alpha' + \beta')}\exp(\beta)\end{aligned}$$

with $\text{expit}(x) = \text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$, α' and β' denoting the fixed effects for the intercept and slope, respectively, in the zero-inflation arm, and α and β denoting the fixed effects for the intercept and slope, respectively, in the Poisson arm. Please note that here, we use α', β', α, and β without any subscript to indicate that we are referring to the respective fixed effects. Under the assumption of a normal sampling distribution, we derived a standard error for the "unconditional" pooled RR by employing the delta method relying upon a first-order Taylor-series approximation:

$$\text{SE}(\widehat{\text{RR}}_{pooled}) = \sqrt{J(\alpha', \beta', \beta) V J(\alpha', \beta', \beta)^T}$$

with $J(\alpha', \beta', \beta)$ denoting the Jacobian matrix for the "unconditional" RR estimator, given by

$$J(\alpha', \beta', \beta) = \left[\frac{(1 - \exp(\beta'))\exp(\beta + \alpha')}{(1 + \exp(\alpha' + \beta'))^2}, \quad -\frac{(1 + \exp(\alpha'))\exp(\alpha' + \beta' + \beta)}{(1 + \exp(\alpha' + \beta'))^2}, \quad \frac{\exp(\beta) + \exp(\alpha' + \beta)}{1 + \exp(\alpha' + \beta')}\right]$$

and V denoting the variance–covariance matrix of the parameters α', β', and β. Please note that computing an "unconditional" pooled RR is only necessary for those ZIP models that include a treatment effect in the zero-inflation arm.

2.3 | Conditional (beta-)binomial models

The third model discussed in Böhning et al. (2015) is a model proposed by Stijnen et al. (2010). Here, we consider Y_{i1} conditional on $Y_i = Y_{i1} + Y_{i2}$. Stijnen et al. (2010) state that given this assumption, Y_{i1} follows a binomial distribution: $Y_{i1} \sim \text{Binom}(Y_i, \pi_i)$ where we predict the probability π_i that an observed event was observed in the treatment group using a binomial regression model with the canonical logit-link (i.e., logistic regression) with a random intercept and an offset variable that is the log group size ratio n_{i1}/n_{i2}. Please note again that instead of only taking the group sizes into account, we could also consider different observation times in conjunction with different group sizes across studies. To simplify, we again do not touch on this option further and refer to Böhning et al. (2015). We write this model as

$$\text{logit}\,\pi_i = \alpha_i + \log(n_{i1}/n_{i2})$$

with $\alpha_i \sim \text{Normal}(\alpha, \sigma_\alpha^2)$. The fixed effect of the intercept in this model represents the pooled log RR across studies, with σ_α^2 indicating the between-study variation in the log RR, that is, τ^2. Please note that here, the log RR is represented by the intercept parameter while it was represented by the slope parameter in the models described above. The likelihood for this model is given by

$$L_m(\alpha, \sigma_\alpha^2 \mid Y_{i1}, Y_i, n_{ij}) = \prod_{i=1}^{m} \int \binom{Y_i}{Y_{i1}} \text{expit}(\alpha_i + \log(n_{i1}/n_{i2}))^{Y_{i1}} \times$$
$$(1 - \text{expit}(\alpha_i + \log(n_{i1}/n_{i2}))^{Y_i - Y_{i1}} \phi(\alpha_i \mid \alpha, \sigma_\alpha^2) d\alpha_i,$$

where m denotes the number of studies included in the meta-analysis. Böhning et al. (2015) point out a disadvantage of this model that has also been criticized by Kuss (2015): it does not use any information from double-zero studies as they would result in $Y_i = 0$. Cai et al. (2010) view the fact that the model relies upon numerical approximations in order to be estimated as another disadvantage. They attempt to remedy this problem by conceiving a beta-binomial model that models π_i as varying between studies and also offers a closed-form likelihood solution that does not have to rely upon numerical approximations (see subsection 2.2.1 in Cai et al., 2010, p. 2080 for more details). Assuming again that $Y_{i1} \sim \text{Binom}(Y_i, \pi_i)$ and now also that $\pi_i \sim \text{Beta}(\psi\gamma, \psi(n_{i2}/n_{i1}))$, the model along with its parameters ψ and γ can be estimated by maximizing

$$L_m(\psi, \gamma \mid Y_{ij}, n_{ij}) = \prod_{i=1}^{m} \int \pi^{Y_{i1}}(1-\pi)^{Y_{i2}} \frac{\pi^{\psi\gamma - 1}(1-\pi)^{\psi(n_{i2}/n_{i1})-1}}{B(\psi\gamma, \psi(n_{i2}/n_{i1}))} d\pi$$
$$= \prod_{i=1}^{m} \frac{B(\psi\gamma + Y_{i1}, \psi(n_{i2}/n_{i1}) + Y_{i2})}{B(\psi\gamma, \psi(n_{i2}/n_{i1}))}.$$

$B(a, b)$ denotes the beta function. Please note that none of the model parameters directly represent the parameter of interest, the pooled (log) RR. The parameter ψ provides a way to adjust the shape of the beta-binomial distribution according to the between-study variation in the treatment effects. The less between-study variation, the larger is ψ (Cai et al., 2010). The authors themselves only examined the estimates of the model parameters γ and ψ in terms of, for example, bias, not their proposed estimator for the pooled RR. Yet, for practical applications, these are likely much more important. Thus, we evaluated the estimates of the pooled RR in our simulation study rather than the model parameters γ and ψ. A pooled RR can be obtained on the basis of the model parameters ψ and γ as

$$\widehat{\text{RR}}_{pooled} = m^{-1} \sum_{i=1}^{m} \int \exp(\theta) f_{Bi}(\theta; \psi, \gamma) d\theta$$

with

$$f_{Bi}(\theta \mid \psi, \gamma) = \frac{\left(\frac{\exp(\theta)}{(n_{i2}/n_{i1})+\exp(\theta)}\right)^{\psi\gamma} \left(\frac{(n_{i2}/n_{i1})}{(n_{i2}/n_{i1})+\exp(\theta)}\right)^{\psi(n_{i2}/n_{i1})}}{B(\psi\gamma, \psi(n_{i2}/n_{i1}))}, \tag{4}$$

where $f_{Bi}(\theta; \psi, \gamma)$ denotes the density of the log RR_i and θ denotes the log RR (Cai et al., 2010). Cai et al. (2010) do not give an estimator for the standard error of $\widehat{\text{RR}}_{pooled}$. Thus, we again employed the delta method under the assumption of a normal sampling distribution to derive a standard error for the pooled RR estimator, relying on a first-order Taylor-series approximation:

$$\text{SE}(\widehat{\text{RR}}_{pooled}) = \sqrt{J(\psi, \gamma) V J(\psi, \gamma)^T}$$

with $J(\psi, \gamma)$ denoting the Jacobian matrix for the pooled RR estimator and V denoting the variance–covariance matrix of the parameters ψ and γ. $J(\psi, \gamma)$ is of the shape

$$J(\psi, \gamma) = \left[\frac{\delta g(\psi, \gamma)}{\delta \psi}, \quad \frac{\delta g(\psi, \gamma)}{\delta \gamma}\right],$$

where we use $g(\psi, \gamma)$ to denote $\widehat{\mathrm{RR}}_{pooled}$. The partial derivatives within the Jacobian matrix are presented in the Appendix. Cai et al. (2010) point out that the density function in Equation (4) can also be used to estimate the between-study variance of the RR_i. Yet, studying the respective estimates would have been difficult with our simulation design (compare Subsection 3.1), and as this model was not the focus of this paper, we neither estimated nor assessed estimates for the between-study variance of the RR_i as provided by the model by Cai et al. (2010).

Finally, we also included a different parameterization of the beta-binomial model for which we wished to assess the performance, one that directly included the parameter of interest, the pooled log RR. The respective model is a beta-binomial model using the canonical logit-link. Here, we once again assume that $Y_{i1} \sim \mathrm{Binom}(Y_i, \pi_i)$ and that $\pi_i \sim \mathrm{Beta}(\mu_i \upsilon, (1 - \mu_i)\upsilon)$, where $\upsilon > 0$ resembles the precision in the sense that it is roughly proportional to the precision, that is, the inverse of the variance. This can be seen from $\upsilon = (\mu(1 - \mu))/\sigma^2 - 1$, with σ^2 denoting the variance of the beta distribution and μ denoting the mean of the beta distribution. We can see that even though υ is not exactly the precision, which would be the inverse of the variance σ^2, it is roughly proportional to the precision. We predict the mean μ_i of the beta distribution via logit $\mu_i = \alpha + \log(n_{i1}/n_{i2})$. Please once again be reminded of our remarks above about being able to also model different observation times. Here, as in the conditional model proposed by Stijnen et al. (2010) and discussed in Böhning et al. (2015), the intercept α represents the pooled log RR, the parameter of interest. The corresponding likelihood is given by

$$L_m(\alpha, \beta, \upsilon \mid Y_{i1}, Y_i, n_{ij}) = \prod_{i=1}^{m} \binom{Y_i}{Y_{i1}} \frac{B(\mu_i \upsilon + Y_{i1}, Y_i + (1 - \mu_i)\upsilon - Y_{i1})}{B(\mu_i \upsilon, (1 - \mu_i)\upsilon)},$$

where $B(a, b)$ denotes the beta function, m the number of primary studies and $\mu_i = \mathrm{expit}(\alpha + \log(n_{i1}/n_{i2}))$. This model does not provide an estimate for the between-study heterogeneity τ^2. Conceptionally, we are unable to include double-zero studies with all three (beta-)binomial models in this section. Based on arguments brought forward by Kuss (2015), we would thus expect these models to perform less well than other models, such as the Poisson or ZIP models, which are naturally able to include all studies.

2.4 | Beta-binomial model by Kuss

Based on an extensive simulation study comparing a wide array of exact-likelihood methods that do not warrant the exclusion of double-zero studies or the administration of continuity corrections, Kuss (2015) recommended the use of a beta-binomial regression model for meta-analysis of the RR. This beta-binomial model differs from the beta-binomial model described above: the beta-binomial model is applied in conjunction with a log link, assuming that $Y_{ij} \sim \mathrm{Binom}(\pi_{ij}, n_{ij})$ and that the π_{ij} follow a beta distribution with a mean μ_{ij}. Then, $\log \mu_{ij}$ is predicted through an intercept α and an indicator of the experimental group membership (treatment or control) X_{i1}, yielding $\log \mu_{ij} = \alpha + \beta X_{i1}$. Note that the model does not conditionalize on the total number of observed events, therewith not introducing the restriction that double-zero studies cannot be included. An interpretation of the slope β in terms of the pooled log RR is enforced by using the log link (for this model, the canonical logit link would result in the slope β representing the pooled log OR). In the parameterization of the beta-binomial model used in Kuss (2015), the beta distribution is parameterized in terms of the mean μ_{ij} and the correlation between the observations from the same study in control and treatment group (please see subsection 3.4 in Kuss, 2015, p. 1101 for further details). Due to being unable to obtain a working implementation of the parameterization used in Kuss (2015) in R, we chose to use a parameterization of the beta-binomial model in which the beta distribution is parameterized in terms of its mean μ and a second parameter υ (similar to the precision), with $\upsilon > 0$ (same parameterization as used for the third conditional beta-binomial model, see above for further details). The crucial difference of this beta-binomial model to the one described in the previous section is the linear combination as well as the link used to predict the mean μ_{ij} of the distribution, resulting in $\log \mu_{ij} = \alpha + \beta X_j$. The corresponding likelihood can be written as

$$L_m(\alpha, \beta, \upsilon \mid Y_{ij}, X_{i1}, n_{ij}) = \prod_{i=1}^{m} \binom{n_{ij}}{Y_{ij}} \frac{B\left(\mu_{ij}\upsilon + Y_{ij}, n_{ij} + (1 - \mu_{ij})\upsilon - Y_{ij}\right)}{B\left(\mu_{ij}\upsilon, (1 - \mu_{ij})\upsilon\right)}, \qquad (5)$$

where $B(a, b)$ denotes the beta function and $\mu_{ij} = \exp(\alpha + \beta X_{i1})$. Note that for consistency's sake, we use α and β to denote the intercept and slope of the regression model here, not the shape parameters of the beta distribution, as is very common. Again, using the log link in conjunction with this model allows for interpreting the treatment effect β as the log RR. But, note that the application of the log link results in mapping the value of the linear combination with a value range of $(-\infty, \infty)$ onto the value range of the inverse log function, that is, $[0, \infty)$, albeit the value range of π being $[0, 1]$. Nonetheless, the maximum likelihood estimator may still converge onto sensible parameter values, but it should be noted that this could lead to numerical difficulties. This model does not provide an estimate for τ^2, the between-study heterogeneity in the RR. Based on the conclusions drawn by Kuss (2015), we would expect this model to perform well, serving as an important benchmark for yet uninvestigated models, such as the ZIP model.

2.5 | Standard RE meta-analysis

Finally, we also included the standard RE meta-analysis model (i.e., inverse-variance method). Even though based on the theoretical considerations elaborated above (Bakbergenuly & Kulinskaya, 2018; Friede et al., 2017; Jackson & White, 2018; Shuster & Walker, 2016; Stijnen et al., 2010) as well as simulative evidence (Pateras, Nikolakopoulos, Mavridis, & Roes, 2018), we do not expect the conventional meta-analysis model to perform particularly well, we still wanted to include it as a benchmark that is commonly used in practice (Warren et al., 2012). The standard RE meta-analysis model computes a pooled $\log \mathrm{RR}$ from the RR_i obtained in i studies included in one meta-analysis. It is also referred to as the two-stage approach: first, an RR_i, its logarithm, and the sampling variance s_i^2 of $\log \mathrm{RR}_i$ are estimated for each study i. Then, the $\log \mathrm{RR}_i$ are predicted using an RE model, with $\log \mathrm{RR}_i \sim \mathrm{Normal}(\theta_i, s_i^2)$ and $\theta_i \sim \mathrm{Normal}(\theta, \tau^2)$, so that equivalently $\log \mathrm{RR}_i \sim \mathrm{Normal}(\theta, \tau^2 + s_i^2)$. Here, θ_i denotes the true underlying $\log \mathrm{RR}_i$ for the ith study. The θ_i are assumed to be drawn from an underlying (RE) distribution with expected value θ and (between-study) variance τ^2. Variants of the standard RE meta-analysis model differ with regard to the estimator they use for τ^2. The most commonly used τ^2 estimator in the (medical) meta-analyses of rare events—as outlined in the systematic literature review by Warren et al. (2012)—is the DerSimonian–Laird (DL) estimator (DerSimonian & Laird, 1986). We included this estimator because of its prevalence. With the DL estimator, the between-study variance in the $\log \mathrm{RR}_i$ is estimated as $\hat{\tau}^2_{DL} = \max(0, (Q_{FE} - (m-1))/c_{FE})$, with $Q_{FE} = \sum_{i=1}^m w_i(Y_i - \bar{Y}_{FE})^2$ and $c_{FE} = \sum_{i=1}^m w_i(\sum_{i=1}^m w_i^2)/(\sum_{i=1}^m w_i)$, where $w_i = 1/s_i^2$, Y_i represents the log RR of study i and $\bar{Y}_{FE} = (\sum_{i=1}^m w_i Y_i)/(\sum_{i=1}^m w_i)$. For the same reason, we also included the restricted maximum likelihood (REML) estimator for τ^2 that is the default estimator in the `R` meta-analysis software package `metafor` (Viechtbauer, 2010) and thus also commonly used. The iterative estimator is given by $\hat{\tau}^2_{REML} = \max(0, (\sum_{i=1}^m \tilde{w}_i{}^2(Y_i - \bar{Y}_{ML}) - s_i^2)/(\sum_{i=1}^m \tilde{w}_i{}^2) + 1/(\sum_{i=1}^m \tilde{w}_i)$, with $\tilde{w}_i = 1/(s_i^2 + \tau^2)$ and $\bar{Y}_{ML}$ denoted the maximum likelihood estimate of the pooled log RR. Even though these estimators have been compared to other methods in rare event settings before (e.g., Pateras et al., 2018; Spittal et al., 2015) and we otherwise tried to avoid redundant comparisons in this simulation study, we still included them in our comparison to be able to compare the performance of more recently proposed models to models that are most commonly used in practice. We also included a third τ^2 estimator, the Sidik–Jonkman (SJ) estimator (Sidik & Jonkman, 2005), given by $\hat{\tau^2}_{SJ} = \max((\sum_{i=1}^m((Y_i - \bar{Y}_{FE})^2)/(r_i + 1))/(m-1), 0.01)$, with $r_i = s_i^2/(\sum_{i=1}^m(Y_i - \bar{Y}_{FE})/m)$. This estimator was chosen based on recommendations by Pateras et al. (2018), albeit their simulations concerned the OR. To avoid more redundant comparisons, we chose only one of the three equally recommended estimators of this paper.

2.6 | Continuity correction

As discussed above, the standard RE meta-analysis model has to rely on continuity corrections or exclusion of studies in the event of single- or double-zero studies. As all other models compared in our simulation study were able to at least include single-zero studies and in line with other simulation studies (Pateras et al., 2018; Spittal et al., 2015), we applied a continuity correction to all single-zero studies so that they could be included in the standard RE meta-analysis model. For double-zero studies, the circumstances were different: not only the standard RE meta-analysis model but also the conditional binomial models were unable to include them. Thus, we chose to exclude them from those models that were unable to handle them and only include them in models that could do so naturally, that is, the Poisson, ZIP models, and the beta-binomial model by Kuss (2015). For each standard RE meta-analysis model, we used two different continuity corrections: (a) We applied the standard continuity correction of adding a constant 0.5 to each cell of the 2×2 frequency tables of those studies in which zero events occurred. (b) We used the treatment arm continuity correction proposed by Sweeting et al. (2004) that is extendable to the RR and an RE scenario. Instead of adding the same constant value to all cells of the 2 × 2 frequency table, we add different values to the treatment and control group that depend on the group sizes. The starting point is again a constant, for example, 0.5, which is then divided by the respective group sizes. This continuity correction supposedly accounts better for group size imbalances than the standard continuity correction, at least in fixed-effects settings (Sweeting et al., 2004). Based on the findings by Sweeting et al. (2004) and hoping to extend them to an RE setting, we would expect the conventional RE meta-analysis models to perform better in conjunction with the treatment arm (as opposed to the standard) continuity correction.

3 | MONTE CARLO STUDY

Our simulations were carried out using `R` (R Core Team, 2019), with the help of the `R` packages `doParallel` (Watson, 2018), `dplyr` (Wickham, François, Henry, & Müller, 2019), `tidyr` (Wickham & Henry, 2019), `truncnorm` (Mersmann, Trautmann, Steuer, & Bornkamp, 2018), and `nleqslv` (Hasselman, 2018), in addition to the `R` packages used to implement the investigated models below (for details, please see Subsection 3.2). The simulations were run on the computing cluster PALMA II (`https://www.uni-muenster.de/ZIV/Technik/Server/HPC.html`) at the University of Münster. You may find `rds` files of the simulation results for each condition on the OSF repository for this paper (`https://osf.io/h4vp6/`) and all `R` scripts used for this simulation, the data preparation and the results visualization on the GitHub repository for this paper (`https://github.com/mariebeisemann/metaanalysis_for_rare_events_simulation`).

3.1 | Simulation set-up

We designed the simulation scenarios to align with settings used in previous work (Cheng et al., 2016; Pateras et al., 2018; Spittal et al., 2015; Sweeting et al., 2004) and to reflect relevant situations in which applied work would use the models examined here (e.g., Cheng et al., 2016). Our aim was to investigate the performance of the models presented above in a setting with rare events. Although such a setting often coincides with other methodological challenges, such as extremely small numbers of primary studies (as low as two studies, see below), publication bias, or outcome reporting bias, our investigation focused only on rare events, operating under the assumption that none of these other methodological challenges were present.

To achieve realistic simulation scenarios (this term is used interchangeably with the term "simulation scenarios" and "simulation settings" throughout the remaining paper), we re-analyzed a set of Cochrane reviews investigating rare events. To this end, we searched the Cochrane Library (https://www.cochranelibrary.com) for the key word "adverse events." Of the 2,953 search results, we downloaded the 2,218 data sets that were available to us and not meta-reviews. We selected the (sub-)data sets with dichotomous outcomes and searched the outcomes for key words typically associated with adverse events ("adverse," "morbidity," "mortality," "death," "trauma," "infection") and selected them accordingly. A total of 1,447 Cochrane reviews met these criteria, leading (with multiple suitable outcomes in some studies) to a re-computation of 15,537 meta-analyses. For each outcome measure, we computed a standard RE meta-analysis (with a continuity correction of 0.5) using the RR as an effect measure. We also used the standard approach to pool the baseline risks (in conjunction with a logit transformation). Additionally, we applied the RE Poisson model to the data that yielded estimates for both the pooled baseline risk and the pooled RR. Excluding results for outcome measures for which only one primary study was available, we computed the mean, the median, and the variance of the RR as well as the baseline risk, as estimated by both approaches. We found a mean baseline risk of 0.13 ($Mdn = 0.07$) with the standard and of .11 ($Mdn = 0.06$) with the Poisson approach. The estimate of the mean RR obtained using the Poisson model appeared distorted (possibly by some models with undetected convergence issues—we did not check all 15,537 models manually), but the mean RR obtained using the standard approach was 1.06 ($Mdn = 1.00$). The log RR (as computed with the standard approach) varied across outcome measures with a variance (on the log scale) of 0.36. The average number of studies included in a meta-analysis (after exclusion of outcome measures with only one study available) was 5 (min = 2, max = 105). We then used these results as inspiration for the parameter values we chose for the parameters we varied to derive our simulation scenarios.

We ran 1,000 simulation trials for each simulation condition. The simulation conditions resulted from a full crossing of six design factors that we are going to describe in the following and an overview of which is shown in Table 3. In accordance with other simulation studies on the topic (Bai et al., 2016; Bhaumik et al., 2012; Bradburn et al., 2007; de Rooi, 2008; Jackson et al., 2018; Sweeting et al., 2004), we generated the data for each one of the $n_{studies}$ primary studies of each respective meta-analysis using a binomial distribution Binom(n, p). Depending on the group (treatment or control),

TABLE 3 Parameters and their values in the simulation study

Parameter	Values
True RR	$\{0.5, 1, 2\}$
Between-study heterogeneity τ	$\{0, 0.6, 1\}$
Baseline event probability μ_{i2}	$\{0.05, 0.1\}$
Number of primary studies m	$\{5, 30, 100\}$
Group size ratio R	$\{0.5, 1, 2\}$
Control group size n_{i2}	50

different values for the parameters n and p according to the respective simulation condition were used. The values for n depended on the simulation conditions as described below. The values for p for study i are (a) the baseline risk μ_{i2} as determined by the simulation condition for the control group, and (b) computed as the product of the baseline risk μ_{i2} and the true RR_i of study i for the treatment group. To model heterogeneous settings, that is, variability in the true RR_i across studies, we sampled m $\log \text{RR}_i$ from a truncated normal distribution (in each trial in each condition; m denotes the number of primary studies in the respective condition), as described below. We then exponentiated the sampled value and multiplied it with the baseline risk, to obtain the event probability in the treatment group, μ_{i1}, which was then used as the parameter p of the binomial distribution from which we generated the data in the treatment group of study i. Based on these generated observations in treatment and control group of the primary studies, we either calculated the log RR and its sampling variance (using one of the two different continuity corrections whenever a primary study displayed zero events in either control or treatment group or both) to be entered into the conventional RE meta-analysis model, or we computed the counts of events per group for each study to be used as the dependent variable in the other models. Based on these data, we estimated all 15 models for every simulation trial in all conditions.

To obtain our simulation conditions, we varied a number of parameters as follows: We used three different underlying true effects ($\text{RR} = \{0.5, 1, 2\}$). We also varied the underlying variability in the true effects, τ^2, or rather τ, so that our simulation scenarios depicted (a) a situation where there is no between-study heterogeneity ($\tau = 0$), (b) a situation where we have as much between-study heterogeneity as we found variation between pooled RR_i in our re-analysis of Cochrane reviews ($\tau = \sqrt{0.36} = 0.6$; which is also roughly in line with Spittal et al., 2015), and (c) a situation with greater heterogeneity, $\tau = 1.0$. We sampled the log RR_i for each study i from a truncated normal distribution. We chose a truncated rather than just a normal distribution (like, e.g., de Rooi, 2008) because as $\text{RR}_i = \mu_{i1}/\mu_{i2}$, where μ_{ij} denote the incidence rate in group j and study i, we cannot sample RR_i (or $\log \text{RR}_i$) completely independently of the baseline risk μ_{i2}. If we were to do so, then especially for larger μ_{i2} and greater heterogeneity, it can happen that $\text{RR}_i \times \mu_{i2} = \mu_{i1} > 1$, even though it should apply that $\mu_{i1} \in [0, 1]$. As we vary the baseline risks systematically but fix them within each condition (see below), we can easily see that this problem would occur when $\text{RR}_i > 1/\mu_{i2}$. Thus, we truncated the normal distribution with an upper boundary of $b = \log(1/\mu_{i2})$ to ensure that we would obtain valid values for μ_{i1} that we computed as $\text{RR}_i \times \mu_{i2}$ in each trial of each condition. Although the truncated normal distribution is still parameterized in terms of parameters μ and σ^2 as well as the lower and upper boundaries, the parameters μ and σ^2 are not equal to the expectation and variance of the truncated normal distribution as they are for the normal distribution. To ensure the validity of our simulation settings, we wanted to make sure that the expectation of the truncated normal distribution was equal to the true log RR we set for the respective condition and that the variance of the truncated normal distribution was equal to the τ^2 we set for the respective condition. The expectation and variance of a truncated normal distribution with a one-sided truncation of the upper tail (as was required for our purposes) are given by

$$E(X \mid X < b) = \mu - \sigma \frac{\phi(z_b)}{\Phi(z_b)} \tag{6}$$

$$\text{Var}(X \mid X < b) = \sigma^2 \left[1 - z_b \frac{\phi(z_b)}{\Phi(z_b)} - \left(\frac{\phi(z_b)}{\Phi(z_b)} \right)^2 \right], \tag{7}$$

with $\phi(.)$ denoting the probability density function of the standard normal distribution, $\Phi(.)$ denoting its cumulative distribution function, and z_b denoting the standardized upper boundary b. We wanted to know which values of μ and σ would result in $E(X \mid X < b) =$ true log RR and $\text{Var}(X \mid X < b) = \tau^2$, so we set Equations (6) and (7) equal to these values and solved them for μ and σ using a nonlinear equation solver (R package nleqslv; Hasselman, 2018). The resulting values (shown in Table S1) were then used together with $b = \log(1/\mu_{i2})$ as parameters of the truncated normal distribution from which the log RR_i for the i studies in each trial in each condition were drawn.

The baseline event probabilities (i.e., the event probabilities in the control group) for the primary studies were systematically varied ($\mu_{i2} = \{0.05, 0.10\}$). The values were chosen to roughly reflect the median (0.06 or 0.07) and the mean (0.11 or 0.13) baseline risk we found in our re-analysis of Cochrane reviews. For the sake of numerical stability, we did not impose any heterogeneity upon the baseline risks. We then used μ_{i2} to compute μ_{i1} as described above. These parameter values were also used in conjunction with the values for the group sizes n_{ij} to generate the observations in treatment and control groups from the binomial distribution. For n_{ij}, we set the values o n_{i2}, that is, the control group size, to 50, and systematically varied the group size ratio R ($= \{0.5, 1, 2\}$), from which we then computed n_{i1} as $n_{i2} \times R$. For n_{i2}, the value $n_{i2} = 50$ was chosen as it had previously been used in other simulation studies as a representative control group size in the medical literature (Cheng et al., 2016; Sweeting et al., 2004). The group size ratio R was varied systematically with values $R = \{0.5, 1, 2\}$.

Finally, we also varied the number of primary studies $m = \{5, 30, 100\}$ included in the meta-analyses. The lowest value of 5 was inspired by the choice of Cheng et al. (2016), Günhan, Röver, and Friede (2018), Sankey et al. (1996), and Spittal et al. (2015), as well as our re-analyses of Cochrane reviews. We included this small number of primary studies as it is common for medical meta-analyses (Cheng et al., 2016). The intermediate value was inspired (yet slightly larger) by choices of other simulation studies (e.g., Bai et al., 2016; Bhaumik et al., 2012), and we chose the largest value based on the maximum number of studies included in the Cochrane reviews we re-analyzed. The inclusion of a greater number of primary studies is also helpful in the assessment of model convergence and performance improvements with increased numbers of primary studies (as suggested by Cai et al., 2010). Altogether, these parameter variations resulted in a total of 162 simulation conditions in which model performance was assessed. Their varied values are summarized in Table 3.

3.2 | Model estimation

All models were fitted in R (R Core Team, 2019). We used the lme4 package (Bates, Mächler, Bolker, & Walker, 2014) to fit the RE Poisson model (poiss; Böhning et al., 2015; Spittal et al., 2015; Stijnen et al., 2010) as well as the conditional binomial model (cond_binom; Böhning et al., 2015; Stijnen et al., 2010). The glmmTMB package (Brooks et al., 2017) served to fit the four variations of the ZIP Model (zip_rifs with random intercept and fixed slope; zip_fifs with fixed intercept and fixed slope; zip_ri with only a random intercept; zip_fi with only a fixed intercept; Böhning et al., 2015) and the parameterization of the conditional beta-binomial model in which the pooled log RR was a parameter of the model (beta_binom). The standard RE meta-analysis models (DL for the standard model using the DL estimator; SJ for the SJ estimator; REML for the REML estimator) were fitted using the metafor package (Viechtbauer, 2010) that also allows for applying the standard continuity correction of adding a constant 0.5 to all cells of the 2 × 2 frequency table in a zero study. We implemented the treatment arm continuity correction (Sweeting et al., 2004) ourselves (standard models in conjunction with the treatment arm continuity correction: DL_tcc, SJ_tcc, REML_tcc). Following other simulation studies (Bradburn et al., 2007; Sweeting et al., 2004), we only applied one of the two continuity corrections in the case of single-zero studies. Double-zero studies were excluded. This is the same treatment (albeit without continuity correction) that is natively administered by the conditional (beta-)binomial models, so it felt most consistent to us. Only in models that were naturally able to take all primary studies into account, even double-zero studies (i.e., the Poisson regression model, the ZIP models, and the beta-binomial model by Kuss, 2015), did we include all primary studies including both single- and double-zero studies. We estimated the beta-binomial models by Kuss (2015) (kuss_binom) and by Cai et al. (2010) (cai_binom) by minimizing their log-likelihoods with the optim function in R, using its default method. For the implementation of the beta-binomial likelihood for kuss_binom, we used the R package extraDistr (Wolodzko, 2019). The kuss_binom model was not fitted with standard R packages as we did not find any package that allowed to fit a beta-binomial regression with a log link. Standard errors for the parameters were obtained from the inverse of the Hessian matrix as computed by optim. To our knowledge, the model by Cai et al. (2010) was also not available in any standard package, and thus, was estimated analogously to kuss_binom. The aforementioned estimators of the pooled RR for cai_binom, zip_rifs, and zip_fifs as well as their derived standard errors were implemented in R using the formulae given in Subsections 2.2 and 2.3, respectively. The variance–covariance matrix used to compute the standard errors was again obtained by taking the inverse of the Hessian matrix computed by optim. Please note that the estimator associated with cai_binom required numerical integration. Our implementation failed to work when trying to approximate the integral over the interval $(-\infty, \infty)$. To still be able to attain the required estimates, we integrated instead over the interval $(-100, 100)$. Please note that this is not entirely accurate (see Section 6 for a reflection upon this).

3.3 | Performance evaluation

As we suspected convergence issues with some of the more complex models in rare events settings (and we also witnessed them in our preliminary trial simulations), we computed the number of trials that did not converge for each condition and each model. We considered models estimated by `metafor`, `lme4` and the likelihoods we minimized using `optim` to have converged when the respective software indicated they had. This was done fully automatically in `R`. In preliminary trial simulations, we found the `glmmTMB` to not have been sensitive enough in this respect, so we chose to consider them converged only when the software returned standard errors for all parameter estimates. Please note that model estimates were not individually screened to check for any unrealistic parameter estimates, but instead we assumed they had converged properly if the criteria described above were met. Model performance was evaluated by assessing the mean and median bias, computed as

$$\text{Bias}_M = \frac{\sum_{i=1}^{n_{trials}} (\hat{\theta}_i - \theta)}{n_{trials}}$$

$$\text{Bias}_{Mdn} = \text{median}(\hat{\boldsymbol{\theta}} - \theta),$$

respectively, where $\hat{\boldsymbol{\theta}}$ denotes the vector of parameter estimates from the n_{trials} simulation trials, the RMSE computed as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n_{trials}} (\hat{\theta}_i - \theta)^2}{n_{trials}}},$$

the mean absolute error (MAE) computed as

$$\text{MAE} = \frac{\sum_{i=1}^{n_{trials}} |\hat{\theta}_i - \theta|}{n_{trials}},$$

and the maximum absolute error (ME) computed as

$$\text{ME} = \max |\hat{\boldsymbol{\theta}} - \theta|,$$

for the point estimator for the pooled $\log \text{RR}$. In all of the equations above, we used n_{trials} to denote the number of converged trials (which differed between models and conditions). There was no parameter as part of the models zip_rifs, zip_fifs, and cai_binom that was directly interpretable as the pooled log RR as was the case for all of the other models. However, we were able to estimate the pooled RR on the basis of the respective model parameters. As these estimates were on the original instead of the log scale, we also computed bias, RMSE, MAE, and ME for these three models on the original RR scale. We also assessed the coverage of the 95% CI for all models. The coverage was computed as the percentage of converged trials in which the 95% CI covered the true $\log \text{RR}$ out of all converged trials (for the respective model in the respective condition). For each model and in each trial i, we computed the 95% CI as $[\hat{\theta}_i - 1.96\widehat{SE}(\hat{\theta}_i), \hat{\theta}_i + 1.96\widehat{SE}(\hat{\theta}_i)]$, with the $\widehat{SE}(\hat{\theta}_i)$ estimates obtained as described in the section above. For models for which estimates of the pooled RR (as opposed to the pooled $\log \text{RR}$) were obtained, the 95% CI were computed on the original scale and the boundaries were subsequently mapped onto the log scale. Then, it was assessed whether the true log RR was covered by these transformed boundaries. The focus of this paper was the evaluation of the models' estimation of the pooled $\log \text{RR}$. In the Supporting Information, we additionally report results regarding the bias (mean and median), the RMSE, the MAE, and the ME of τ^2 for models that provided respective estimates (i.e., the standard meta-analysis models with both continuity corrections, zip_ri, zip_fi, poiss, cond_binom).

4 | RESULTS

We included a plot showcasing the mean relative frequency of zero counts across trials for each condition in the Supporting Information, in case the reader wishes to develop a better impression of the simulated data sets in the conditions. Furthermore, we also display the average number of single- and double-zero studies across trials per condition in the Supporting Information. On average, there were few double-zero studies. The average number of single-zero studies (relative to the number of primary studies) was—as is to be expected—highest in conditions with a baseline risk of 0.05 and an RR of 0.5. In these conditions, there were roughly 50% single-zero studies.

4.1 | Convergence

In this paper, we merely show the results regarding model convergence in the different conditions for conditions with a group size ratio of 0.5, the results for conditions with the other group ratios showed a similar pattern and are shown in the Supporting Information. In Figure 1, we color-coded the amount of trials that failed to converge ranging from black (indicating that none of the 1,000 trials converged for the respective model in the respective condition) to white (indicating that all trials converged for the respective model in the respective condition). We observed that across all conditions, the standard meta-analysis models (with either continuity correction), beta_binom, poiss, and binom converged (nearly) seamlessly. For settings without heterogeneity, the cai_binom model showed considerably more convergence issues than it did in the settings with more heterogeneity (a pattern that also occurred, albeit only very subtly, for kuss_binom). The inverse pattern was observed for the ZIP models that showed overall the poorest convergence rates out of all the models. However, they converged comparatively better in settings without heterogeneity. Figure 1 also shows that in setting without heterogeneity, convergence of the ZIP models was even better the more primary studies were included in the meta-analysis. A greater number of primary studies seems to have aided convergence for cai_binom as well, yet was not able to remedy the convergence problems observed for settings without heterogeneity.

4.2 | Distribution of (pooled) log RR estimates across trials

For a first impression of the results regarding the log RR estimates, we show the distribution of log RR estimates from the 1,000 trials per model and condition in Figure 2 (selectively for conditions with a group size ratio of 0.5; please see the Supporting Information for the figures displaying the results for conditions with group sizes of 1 and 2). As we only wanted to give a rough overview of the models' log RR estimates and also an impression of how the models compare to each other in their performance, we show the results for all models in the same plot, to this end taking the logarithm of estimates of the pooled RR as provided by some models (i.e., zip_rifs, zip_fifs, and cai_binom). We refrain from doing so in all subsequent plots in which we want to scrutinize the results a little more. Please note that naturally an unbiased estimator on the original scale, should show bias on the log scale or vice versa. Figure 2 shows that for all models, log RR estimates varied more across trials in conditions with fewer primary studies and with more heterogeneity. We would like to highlight the very good performance of poiss and kuss_binom (except for in the most extreme conditions in terms of data sparsity). Yet, the comparatively slightly wider distributions for poiss (as opposed to those for some of the other models) also show that for some trials, log RR estimates were still quite far off (as indicated by the narrow but—at least in some conditions—long tails of the distribution; for example, for a very low baseline risk of 0.05, an RR of 0.5 equating to a log RR of −0.67, 30 primary studies, and a group ratio of 0.5: the distribution of estimates indicates some pooled log RR were estimated as small as roughly −2 across different amounts of heterogeneity). However, this becomes much less of a problem as the number of primary studies increases (or alternatively the baseline risk increases). kuss_binom performed very well for lesser degrees of heterogeneity, showing narrower distributions than poiss. However, as heterogeneity increases, the model mean for kuss_binom (but not the one for poiss) did not coincide with the true underlying log RR. We observed a similar pattern for beta_binom and cond_binom. The distributions for the ZIP models were overall slightly wider spread than those of the other models but the respective model means of the ZIP models usually came very close to the true underlying log RR (indicated by the gray vertical lines), but please note the conditions where this was not the case. cai_binom exhibited slightly wider distributions that—as for all models—narrowed with increased numbers of primary studies. Yet, for cai_binom we observed model means far off from the center of the distribution relatively more

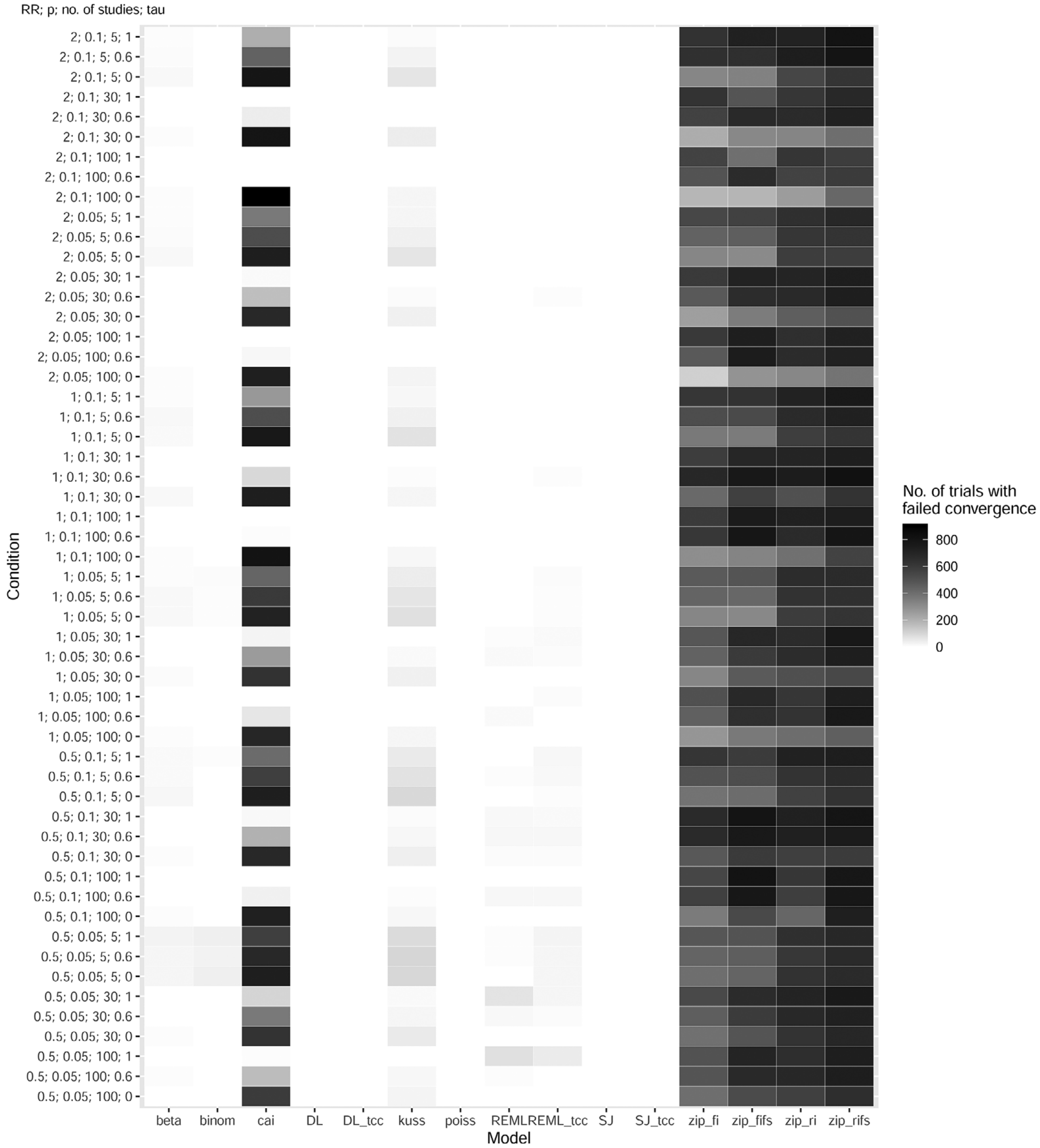

FIGURE 1 Number of trials for which convergence failed shown per model (shown on *x*-axis) and condition (shown on *y*-axis), selectively for conditions with a group size ratio of 0.5

Note. The values are color-coded with red indicating that all trials failed to converge for the respective model and condition and green indicating convergence in all trials. Abbreviations on the *x*-axis: beta, beta_binom; binom, cond_binom; cai, cai_binom; kuss, kuss_binom. Conditions on *y*-axis are described in terms of true RR (RR), baseline risk (p), number of primary studies (no. of studies), and amount of heterogeneity (tau); the values of these parameters together with these abbreviations are shown on the *y*-axis

often than for the other models. It should be noted that this may have also resulted from or rather been aggravated by our transformation of the values onto the log scale for this plot, but as we are going to see below, bias, RMSE, MAE, and ME paint a similar picture. The standard meta-analysis models showed the (relative to the other models in the respective conditions) narrowest distribution. However, their model means were also consistently off, even for settings without heterogeneity. The application of the alternative treatment-arm correction tended to lead to further off distributions than that of the standard continuity correction. It seems also worth noting that the cases in which the model means (displayed as red triangles) do not fall into the center of the distribution indicate that there might have been outliers for the respective model in the respective condition that considerably distorted the average pooled log RR. Results were overall similar also for group size ratios of 1 and 2.

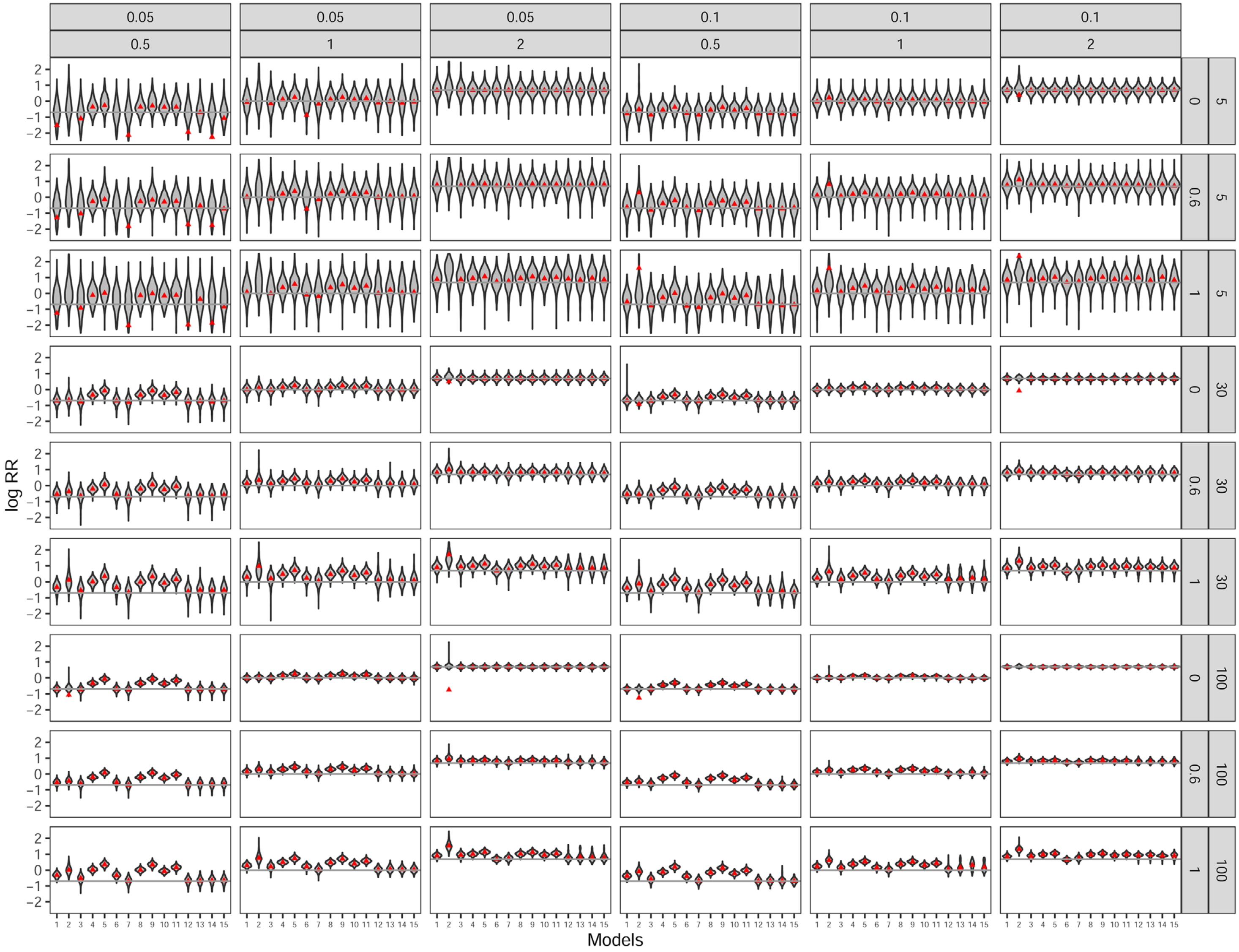

FIGURE 2 Violin plots showing the distribution of the pooled log RR estimates (displayed on the *y*-axis) across simulation trials for a group size ratio of 0.5. Different models (displayed on the *x*-axis) are shown in different colors (see legend on the right-hand side)

Note. Note that for zip_rifs, zip_fifs and cai_binom, estimates obtained are for the pooled RR and were converted to the log scale for this plot. Gray panels along *y*-axis indicate the number of primary studies (5, 30, or 100) in the upper and degree of heterogeneity (as the standard deviation on the log scale: 0, 0.6, or 1.0) in the lower row. Gray panels along the *x*-axis indicate the true baseline risk (0.05 or 0.10) in the upper and the true RR (0.5, 1, or 2) in the lower row. Gray horizontal lines indicate the true log relative risk. The red triangle denotes the mean pooled log RR across simulation trials per model and condition. 1 = beta_binom, 2 = cai_binom, 3 = cond_binom, 4 = DL, 5 = DL_tcc, 6 = kuss_binom, 7 = poiss, 8 = REML, 9 = REML_tcc, 10 = SJ, 11 = SJ_tcc, 12 = zip_fi, 13 = zip_fifs, 14 = zip_ri, 15 = zip_rifs

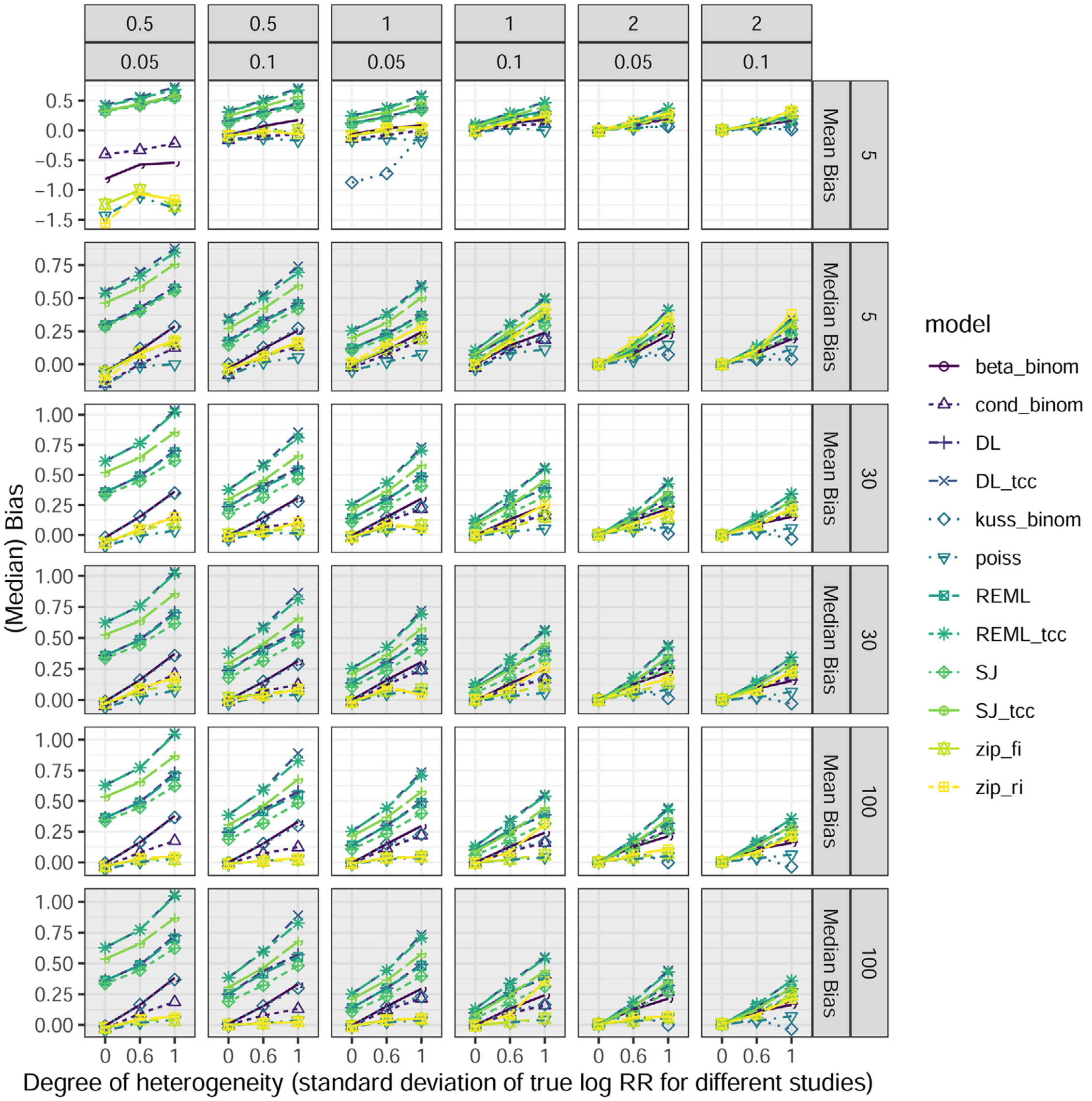

FIGURE 3 Median (gray panels) and mean (white panels) biases of the estimate for pooled log RR (displayed on *y*-axis) shown for different degrees of heterogeneity (displayed on *x*-axis), shown selectively only for conditions with a group size ratio of 0.5
Note. Results for different models are shown in different colors, shapes and line types (see legend on right-hand side). Note that cai_binom, zip_fifs, and zip_rifs are not shown here. Note that the values range of the *y*-axis varies between panels. Only values > −2 are shown. Gray panels along *x*-axis indicate true RR (0.5, 1, or 2) in the upper and true baseline risk (0.05 or 0.10) in the lower row. Gray panels along the *y*-axis indicate type of bias (mean or median) in the lower and number of primary studies (5, 30, or 100) in the upper row

4.3 | Bias

Mean and median biases in the estimated pooled log RR are displayed for a group size ratio of 0.5 and a restricted value range (only bias values above −2 are shown) are displayed in Figure 3 (please see the Supporting Information for the results regarding group size ratios of 1 and 2 as well as complete value ranges). Please note that if for any value of between-study heterogeneity (displayed on *x*-axis) for any model, no bias is shown in the plot, than that is due to the bias being too great to be displayed on the scale of the *y*-axis in the plot. Such a bias would be unacceptably large regardless of the exact value. Please note that results for cai_binom, zip_fifs, and zip_rifs are not shown in Figure 3, but instead in Figure 4 in order to be able to show their bias on the scale of the estimator (i.e., the original scale of the RR, not the log scale as for the models shown in Figure 3). Overall, the condition with a baseline risk of 0.05, a group ratio of 0.5, an RR of 0.5 and five primary studies resulted in the most bias for the models. For this condition, mean biases (displayed on white

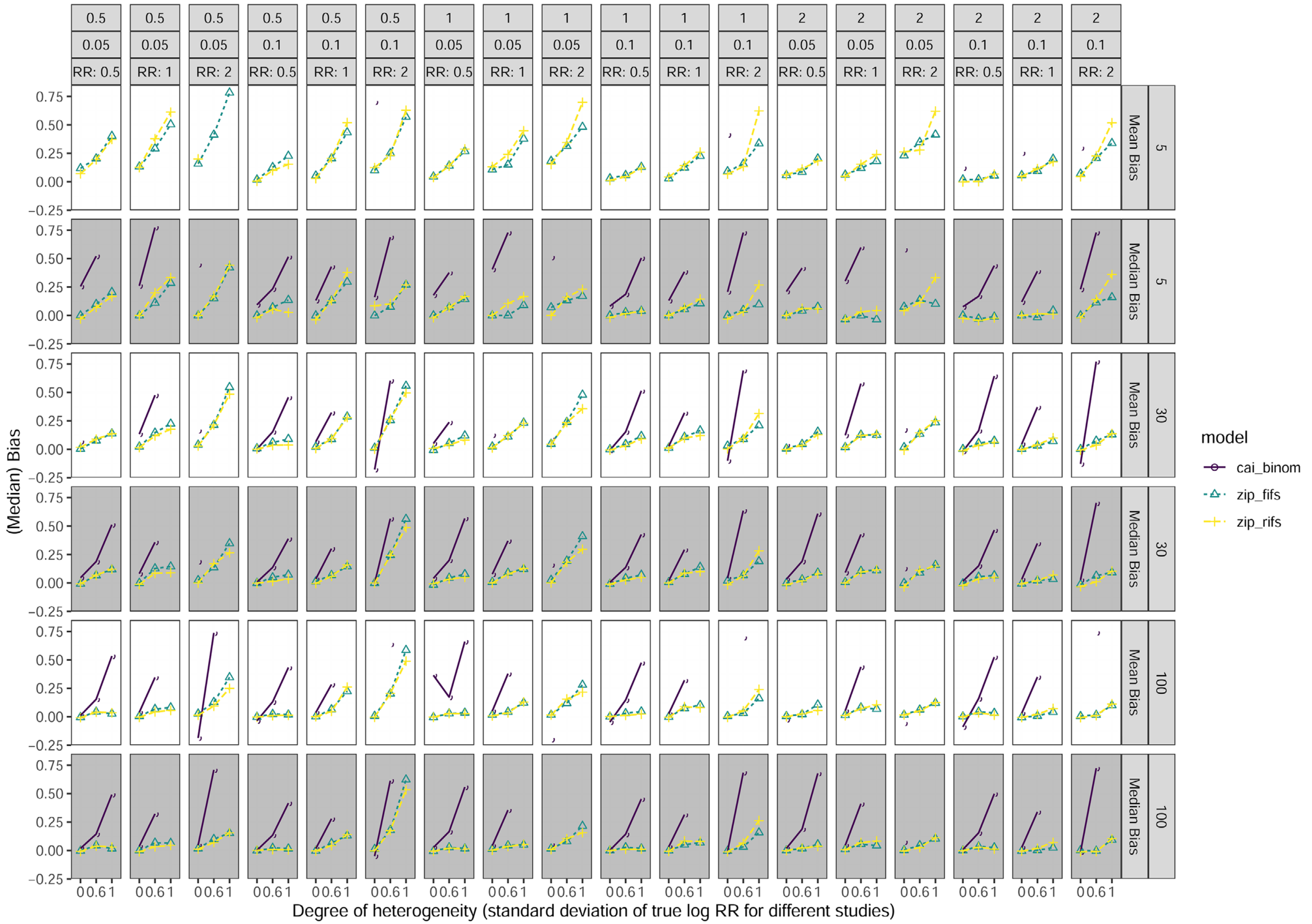

FIGURE 4 Median (gray panels) and mean (white panels) biases of the estimate for pooled RR (displayed on y-axis) for the models cai_binom, zip_fifs and zip_rifs, shown for different degrees of heterogeneity (displayed on x-axis), shown for all conditions

Note. Results for different models are shown in different colors, shapes, and line types (see legend on right-hand side). Note that the values range of the y-axis varies between panels. Only values between −0.2 and 0.8 are shown. Gray panels along x-axis indicate true RR (RR: 0.5, RR: 1, or RR: 2) in the lower, true baseline risk (0.05 or 0.10) in the middle and group size ratio (0.5, 1, or 2) in the upper row. Gray panels along the y-axis indicate type of bias (mean or median) in the lower and number of primary studies (5, 30, or 100) in the upper row panels) differed considerably from median biases (displayed on gray panels), suggesting the presence of outliers for the respective models in that condition. Notably, poiss and kuss_binom (too large to be shown in Figure 3 in the respective condition) showed very large biases in that condition, yet, for most other conditions, poiss and kuss_binom showed the least amount of bias out of all the models. (Mean and median) bias for poiss and kuss_binom is very close to zero in most condition. (Mean and median) bias for poiss tended to be slightly negative in conditions without heterogeneity and slightly positive for settings with a lot of heterogeneity. Especially for an RR of 0.5 (so when zero-events were most likely in the treatment group), poiss performed better in terms of bias than kuss_binom, in particular for settings with more heterogeneity. In these settings, the ZIP models also tended to outperform kuss_binom in terms of bias (but note how much the ZIP models struggled in settings with an RR of 2, especially in conjunction with only five primary studies). This pattern is reversed for an RR of 2, but in these conditions, bias for kuss_binom was not visibly smaller in absolute magnitude than that for poiss, just negative (while the bias for poiss was slightly positive) for great heterogeneity. Performance of the standard models of meta-analysis was overall poor in terms of bias, regardless of the condition. cond_binom and beta_binom tended to exhibit bias that was somewhere in between the bias observed for the well and the poorly performing models. For control group size ratios of 1 and 2 (see the Supporting Information), the overall patterns of results were similar. The most notable difference was that poiss showed much more clearly superior performance in terms of bias, especially for larger numbers of primary studies (although the ZIP sometimes came close again). For these greater group sizes ratios, kuss_binom tended to show rather noticeable negative biases in settings with great heterogeneity. Performance of cond_binom was noticeably poorer for group size ratios of 1 and 2, in particular for an RR of 2. It might also be interesting to note that for greater group size ratios, in particular in conjunction with greater RR, the standard models of meta-analyses tended to show negative bias in settings without heterogeneity. Mean and median biases for the remaining models, that is, cai_binom, zip_fifs, and zip_rifs, are shown on their original scale in Figure 4. For this figure, we restricted the value range in order to be able to make out relevant differences. Only the display of biases for cai_binom was affected by this choice and the respective complete value range is displayed in a table in the Supporting Information. Overall, performance of cai_binom in terms of bias can be summarized as very poor with better performance for settings without heterogeneity (but please note that for these conditions, cai_binom demonstrated the greatest convergence problems) and settings with more primary studies. Performance of zip_fifs and zip_rifs was similar to each other and much better than that of cai_binom. Bias was smaller for more primary studies and noticably larger for more heterogeneity and an RR of 2. Please consult the Supporting Information for Monte Carlo standard errors for the (mean) bias estimates.

4.4 | RMSE, MAE, and ME

RMSE for all models except cai_binom, zip_fifs and zip_rifs (which again provided estimates for the pooled RR instead of the pooled log RR, resulting in RMSE values on a different scale; see Figures S14 and S15) are shown in Figure 5. MAE and ME are shown in Figures 6 and 7 for all models that provide an estimate of the pooled log RR, respectively (please see Figures S17 and S18 as well as S21 and S22 for the results for the other models, namely, cai_binom, zip_fifs, and zip_rifs). All these plots only show restricted value ranges for which the limits are always indicated in the figure description. Please consult the Supporting Information for plots and tables (depending on the model) showing the complete value range. In Figures 5–7, we can see that overall, RMSE, MAE, and ME tended to be higher across conditions and models for settings with more heterogeneity. As we have seen before for other measures, while performance in terms of RMSE and MAE was remarkably poor for poiss and especially kuss_binom in the most extreme conditions in terms of data sparsity, both models showed good (either absolutely or at least compared to the other models) performance in the majority of conditions. Figure 7 as well as the corresponding plots in the Supporting Information that show the complete value range highlight how poor performance of these models was in the worst cases with maximum errors for poiss exceeding 30 and kuss_binom exceeding 600 on the log scale. Also in line with previous observations, poiss tended to perform best for conditions with an RR of 0.5 and kuss_binom tended to perform best for conditions with an RR of 2. These differences were especially noticeable for conditions with great heterogeneity. For an RR of 1, the models' performance was similar, with some conditions where poiss was better than kuss_binom and some conditions where the opposite was the case. Even though not as pronounced, the ZIP models showed a similar pattern of performance across conditions as poiss, however, poiss tended to outperform them slightly in terms of RMSE and MAE in most conditions in which poiss performed very well. In terms of ME, the ZIP models even tended to show slightly smaller ME than poiss on occasion. With regard to the remaining models, the results regarding the RMSE and MAE also hardly revealed any new patterns so that due to space limitations, we will not reiterate the descriptions of the patterns in the results for the RMSE and the MAE that were similar to those in the results for the bias. However, we did notice that in terms of RMSE that while usually the RMSE tended to be larger

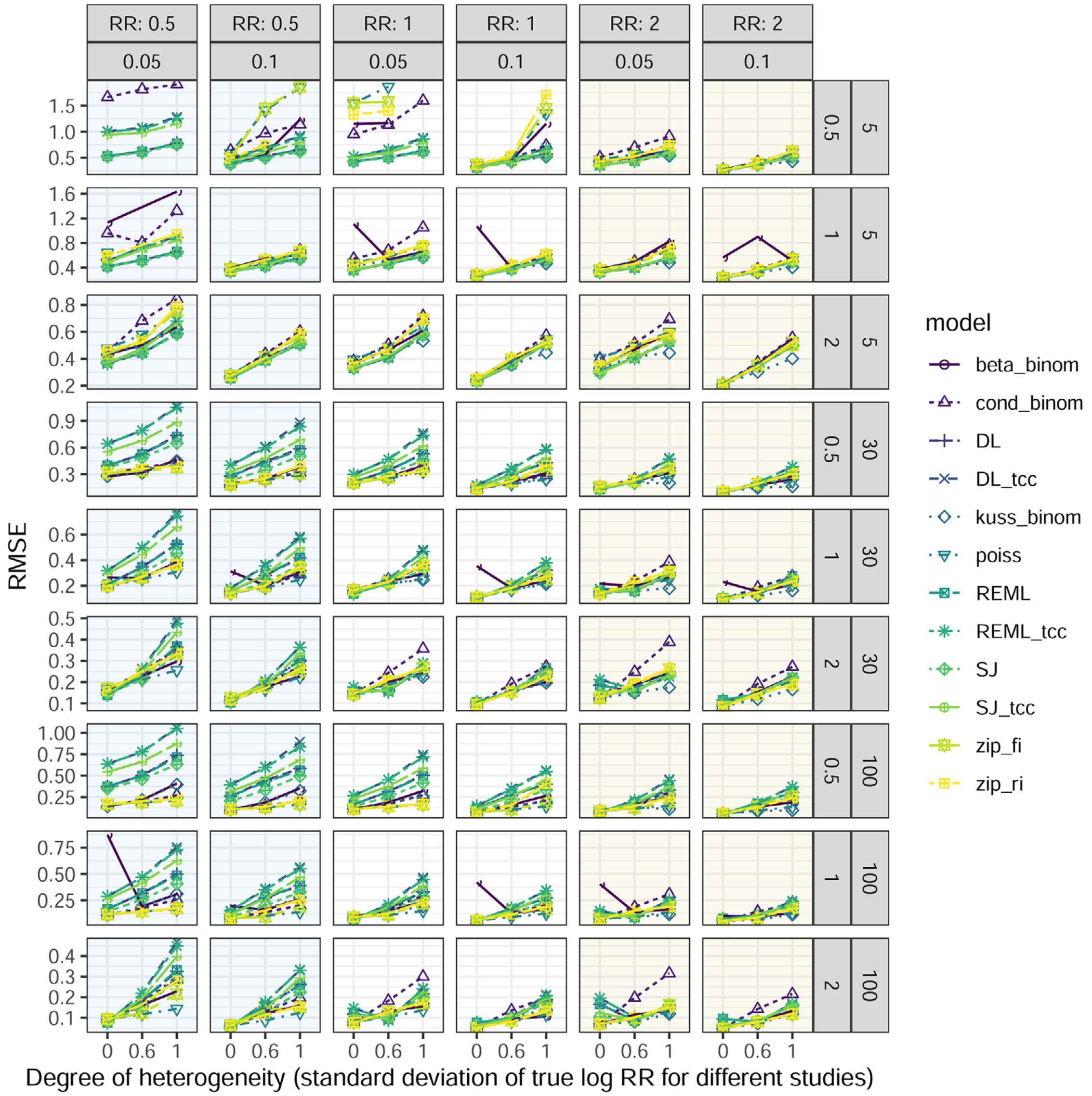

FIGURE 5 RMSE of log RR (displayed on *y*-axis) for different degrees of heterogeneity (displayed on *x*-axis)
Note. The underlying true RR is indicated by the upper gray panels along the *x*-axis as well as by color-coded background panels (RR of 0.5 in blue, RR of 1 in white, RR of 2 in yellow). Results for different models are shown in different colors, shapes, and line types (see legend on right-hand side). Please note that cai_binom, zip_fifs, and zip_rifs are not shown here. Note that the scale of the *y*-axis varies between rows. Only RMSE values smaller than 2 are shown. Lower gray panels along *x*-axis indicate true baseline risk (0.05 or 0.10). Gray panels along the *y*-axis indicate the number of primary studies (5, 30, or 100) in the upper row and group size ratio (0.5, 1, or 2) in the lower row

for settings with more heterogeneity, it was noticeably larger for beta_binom in settings without heterogeneity in some conditions. We did not see the same pattern for beta_binom in terms of the MAE, suggesting that these large RMSE values might have arisen due to single outliers that are weighted more heavily by the RMSE than the MAE due squaring the deviations from the true parameter value as opposed to taking the absolute difference. This also seems plausible in light of the very high ME we observed (see Figure 7) for beta_binom in those conditions. We also observed on occasion very large ME for cond_binom (see Figure 7). Figures S14, S15, S17, S18, S21, and S22 show that much like in terms of bias, cai_binom also performed very poorly in terms of RMSE, MAE, and ME, respectively, even more dramatically so in conditions with fewer primary studies. Again, we were not able to show the whole value range for cai_binom in terms of RMSE, MAE, and ME, but we included tables in the Supporting Information in which the whole value range is presented. Performance of zip_rifs and zip_fifs was overall much better and was in particular very good for large numbers of primary studies and less heterogeneity.

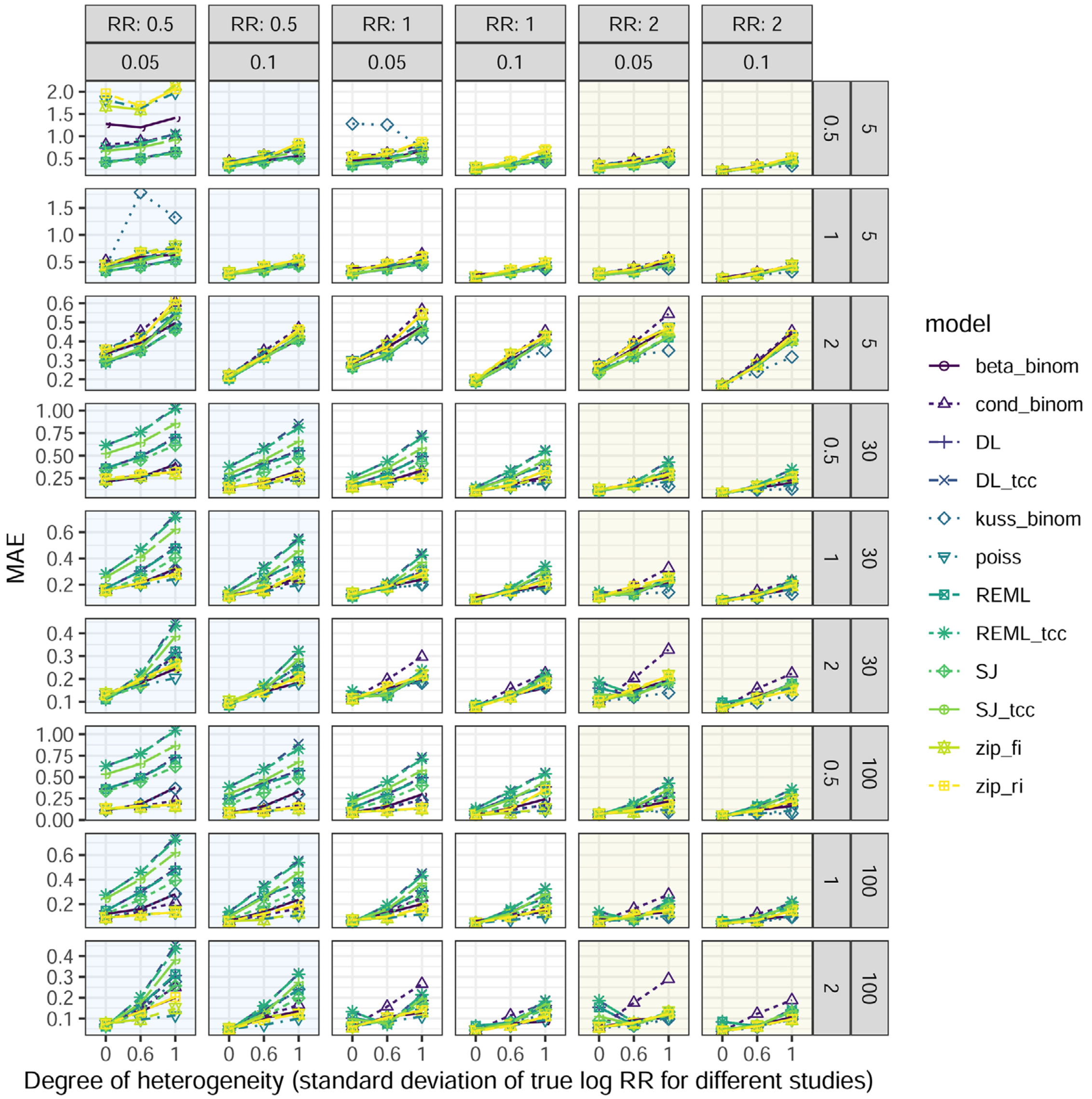

FIGURE 6 Mean absolute error (MAE) of log RR (displayed on *y*-axis) for different degrees of heterogeneity (displayed on *x*-axis) *Note*. The underlying true RR is indicated by the upper gray panels along the *x*-axis as well as by color-coded background panels (RR of 0.5 in blue, RR of 1 in white, RR of 2 in yellow). Results for different models are shown in different colors, shapes, and line types (see legend on right-hand side). Please note that cai_binom, zip_fifs, and zip_rifs are not shown here. Note that the scale of the *y*-axis varies between rows. Only MAE values smaller than 2 are shown. Lower gray panels along *x*-axis indicate true baseline risk (0.05 or 0.10). Gray panels along the *y*-axis indicate the number of primary studies (5, 30, or 100) in the upper row and group size ratio (0.5, 1, or 2) in the lower row

4.5 | Coverage

Coverage of the 95% CI for the pooled log RR is displayed in Figure 8. Please see Subsection 3.3 on why we display the results for all models in one plot. The nominal level of 95% is marked with a black horizontal line. It is immediately visible by looking at Figure 8 that coverage was an issue for the overwhelming majority of the models in most conditions. As a general tendency for all values of the RR, coverage for most models tended to be better for conditions with less heterogeneity. Coverage was closest to nominal level in those conditions in which the number of primary studies was smaller. As the number of primary studies increases, the coverage of the 95% CI decreased for several models, especially in conditions with smaller group size ratios—and that in part drastically so. For instance, for 100 primary studies, the actual coverage for REML_tcc, REML, SJ, or SJ_tcc sometimes dipped as low as 0%. As precision increases with higher numbers of primary studies (i.e., CI grow more narrow), it is likely that the bias we have observed above, in conjunction with those increasingly more narrow CI, resulted in considerable larger numbers of CI than expected that do not cover the true parameter. To corroborate this suspicion of ours, we inspected the ratio of the average estimated standard error

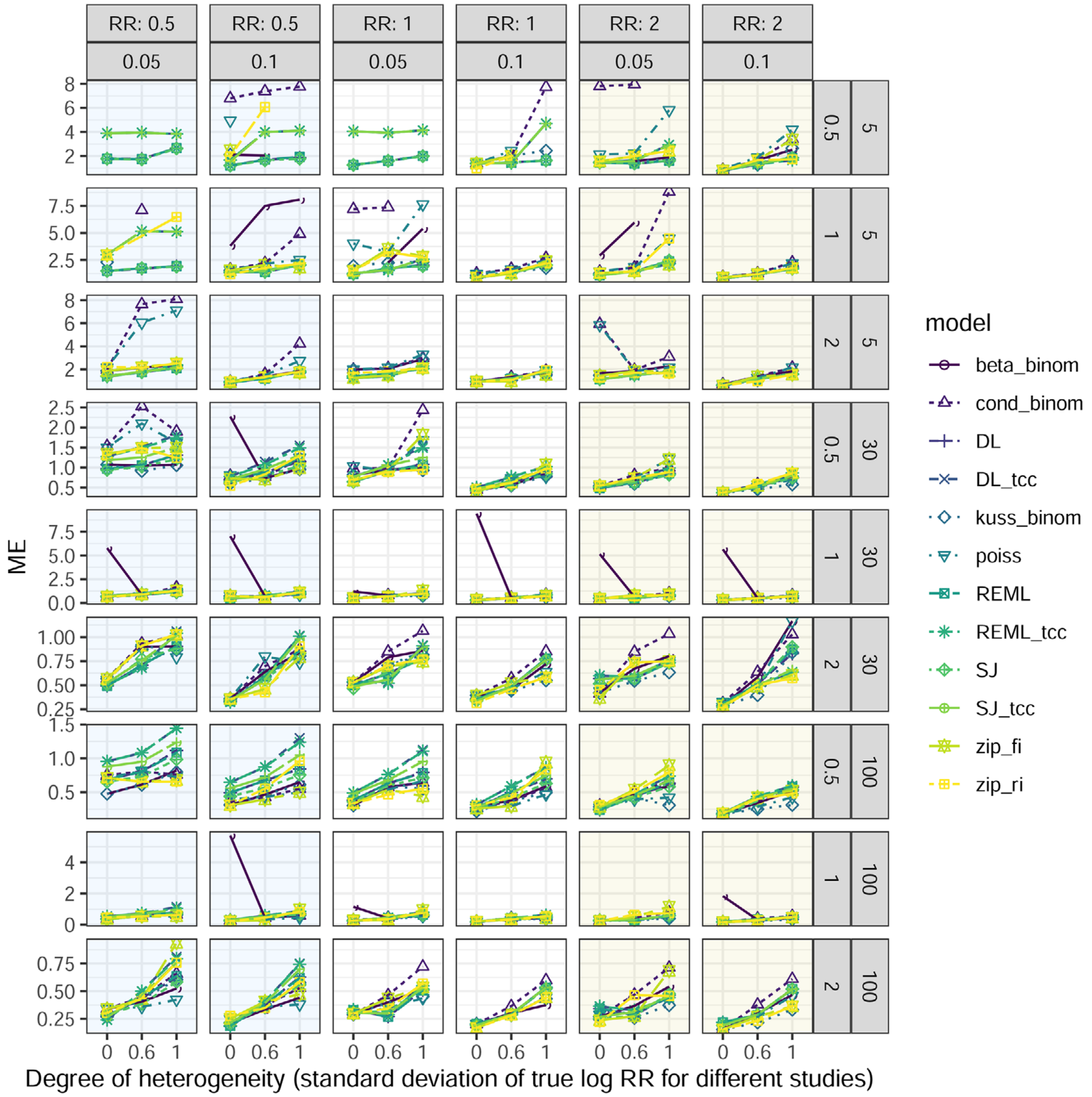

FIGURE 7 Maximum absolute error (ME) of log RR (displayed on y-axis) for different degrees of heterogeneity (displayed on x-axis) *Note*. The underlying true RR is indicated by the upper gray panels along the x-axis as well as by color-coded background panels (RR of 0.5 in blue, RR of 1 in white, RR of 2 in yellow). Results for different models are shown in different colors, shapes, and line types (see legend on right-hand side). Please note that cai_binom, zip_fifs, and zip_rifs are not shown here. Note that the scale of the y-axis varies between rows. Only ME values smaller than 10 are shown. Lower gray panels along x-axis indicate true baseline risk (0.05 or 0.10). Gray panels along the y-axis indicate the number of primary studies (5, 30, or 100) in the upper row and group size ratio (0.5, 1, or 2) in the lower row

of the parameter estimate to the observed standard deviation of parameter estimates across trials, for each model in each condition (i.e., mean($\widehat{\mathrm{SE}}(\hat{\theta})$)/ SD($\hat{\theta}$), with θ denoting the pooled (log) RR; referred to as "SE ratio" in the following). The results are visualized in the Supporting Information. Ratios below 1 indicate that the SE tended to be underestimated, suggesting that the low coverage might be caused by this instead of the bias, as we have speculated. We found that the standard models tended to overestimate the SE, speaking for our explanation of the low coverage of these models. At the same time, for a small number of primary studies, SJ and SJ_tcc were among the, if not the, models with the highest coverage, which was sometimes—especially for no heterogeneity—higher than the nominal level. In fact, for a small number of primary studies, we observed that most models were too conservative in settings without heterogeneity. For only five primary studies, cai_binom also showed good coverage, but the performance dropped with an increase in the number of primary studies so much so that for 100 primary studies, cai_binom performed as one of the worst models. Again, this could be due to the increased narrowness of the CI around very biased estimates. Our inspection of the SE ratios (see the

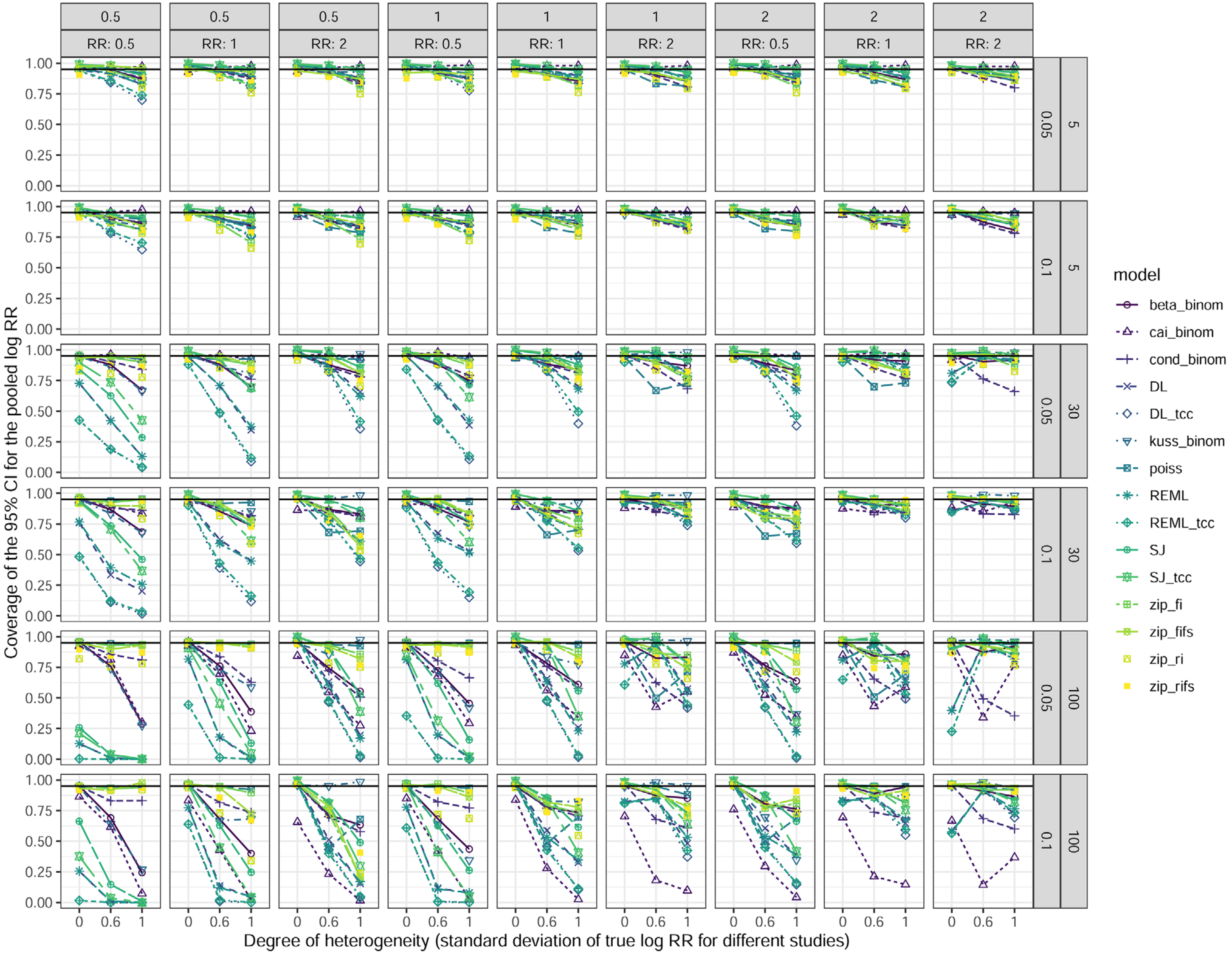

FIGURE 8 Coverage of the 95% CI for the pooled log RR (y-axis) for different degrees of heterogeneity (x-axis)

Note. Coverage for different models is shown in different colors, line types, and shapes (see legend on right-hand side). The black horizontal lines indicate the nominal level of coverage (95%). Gray panels along x-axis indicate group size ratio (0.5, 1, or 2) in the upper and true RR (RR: 0.5, RR: 1, or RR: 2) in the lower row. Gray panels along the y-axis indicate true baseline risk (0.05 or 0.10) in the lower and number of primary studies (5, 30, or 100) in the upper row

Supporting Information) revealed that while for conditions with more primary studies, there was a slight tendency that SE were underestimated, speaking against bias as the only explanation for the low coverage of these models. Nonetheless, both factors could be at work here. For settings with little heterogeneity, SE for cai_binom tended to have been overestimated Coverage for models that demonstrated less bias, especially with increased number of primary studies, also showed better coverage with increased number of primary studies. For instance, for the majority of conditions with larger number of primary studies, albeit there are exceptions, the coverage for poiss is comparatively the closest to the nominal level and often not very far off. In a considerable number of conditions, coverage for kuss_binom was also quite good, however, it dropped to alarmingly low levels for smaller group size ratios, in particular in conjunction with greater amounts of heterogeneity and more primary studies. Coverage for the ZIP models, in particular also the ZIP with a slope in the zero-inflation arm of the model, was quite consistent across conditions at very close to nominal level for no heterogeneity and—at least in some conditions—further below nominal level for more heterogeneity. This pattern of diminished coverage for settings with more heterogeneity is accompanied by a tendency of poiss and the ZIP models of underestimating the SE in these settings (see the Supporting Information). However, for kuss_binom, SE tended to be underestimated for settings without heterogeneity (especially in conjunction with an RR of 0.5) and overestimated for settings with more heterogeneity (especially in conjunction with an RR of 1 and 2). In most conditions, coverage for cond_binom and beta_binom was not notably better than for the relatively well performing models, in fact, it was often rather extremely poor. Both models also exhibited SE ratios indicating that underestimation of SE has occurred in some settings, most notably so for settings without heterogeneity for beta_binom and for settings with more heterogeneity for cond_binom.

4.6 | Additional results

Additional results mainly regarding the models' performance in estimating the between-study variance in the underlying effect, τ^2, are available in the Supporting Information. As this was not the major focus of this paper and space is limited, we are not going to go into detail about these results in this paper. However, we would like to point out that the model that performed (among the) best in terms of estimating the pooled effect across studies, that is, poiss, also performed the best in terms of estimating the between-study variance in the effects (as indicated by bias, RMSE, MAE, and ME) in the majority of conditions. However, it should also be highlighted that poiss also failed to provide unbiased estimates of the between-study heterogeneity for the most extreme conditions in terms of data sparsity, in particular for settings with only five primary studies. Additionally, it is also important to note that performance increased visibly with larger numbers of primary studies. For 30 and 100 primary studies, cond_binom, zip_fi, and zip_ri showed similar performance as poiss (especially cond_binom; but note that cond_binom also exhibited the largest ME in some conditions, especially in those with greater numbers of zeroes and in particular in settings with only five primary studies), but provided slightly more biased estimates for settings with great heterogeneity (especially zip_fi and zip_ri). The standard models of meta-analysis showed noticeably larger amounts of bias. Yet, while the standard models were consistently off in their estimates, the alternative models tended to exhibit considerably higher ME, especially in the more extreme conditions in terms of data sparsity. Please note that not all models examined in this paper provide estimates for τ^2, even though they do all take between-study variability in the effect into account.

5 | ILLUSTRATIVE EXAMPLES

To illustrate how the models can differ in their effect estimates when applied to one single data set, we used the data from two recent Cochrane reviews (Hemkens et al., 2016; Squizzato et al., 2017) for which the data are available through the Cochrane Library (https://www.cochranelibrary.com). Excerpts of the data are shown in Table 1 (four studies for each outcome). Squizzato et al. (2017) summarized results on the effects (both beneficial and adverse) of using aspirin as a prophylactic antiplatelet drug in conjunction with another antiplatelet drug, namely clopdiogrel, for cardiovascular disease patients. Hemkens et al. (2016) pooled effects (both beneficial and adverse) of the anti-inflammatory drug Colchicine on cardiovascular outcomes. Please see the respective reviews for the detailed descriptions of the respective object of investigation as well as details on the methodological approaches of the papers, which followed Cochrane guidelines. We chose to re-analyze some of the data subsets from these reviews in which the outcome was an adverse event, specifically cardiovascular mortality (Hemkens et al., 2016; Squizzato et al., 2017) and fatal strokes (Hemkens et al., 2016). The examples chosen and presented here are supposed to illustrate how difficult meta-analysis of rare events can present itself in real-life

TABLE 4 Results obtained with the different models for different outcomes from three data examples

	Cardiovascular mortality				Fatal stroke	
	Squizzato et al.		Hemkens et al.		Hemkens et al.	
Model	**RR**	**95% CI**	**RR**	**95% CI**	**RR**	**95% CI**
REML	0.983	[0.879, 1.099]	0.346	[0.091, 1.318]	2.947	[0.121, 71.567]
REML_tcc	0.984	[0.879, 1.101]	0.243	[0.029, 2.050]	217.004	[0.000, 9.770e+14]
DL	0.983	[0.879, 1.100]	0.335	[0.094, 1.185]	2.947	[0.121, 71.567]
DL_tcc	0.984	[0.880, 1.100]	0.211	[0.042, 1.067]	217.004	[0.000, 9.770e+14]
SJ	1.018	[0.779, 1.330]	0.366	[0.082, 1.625]	2.947	[0.121, 71.567]
SJ_tcc	1.039	[0.654, 1.651]	0.263	[0.018, 3.819]	217.004	[0.000, 9.770e+14]
poiss	1.130[a]	[0.865, 1.478][a]	0.270	[0.052, 1.409]	1.002e+13[a]	[0.000, Inf][a]
zip_rifs	–	–	–	–	–	–
zip_fifs	–	–	–	–	–	–
zip_ri	–	–	–	–	–	–
zip_fi	–	–	–	–	–	–
cond_binom	0.983[a]	[0.877, 1.102][a]	0.195[a]	[0.056, 0.679][a]	–	–
beta_binom	–	–	–	–	–	–
kuss_binom	0.935	[0.451, 1.938]	0.200	[0.058, 0.691]	–	–
cai_binom	–	–	–	–	–	–

Note. When no values are provided (–), the model did not converge.
Models for which warnings were given are marked with[a], warnings occurred when model fit was singular.

settings in which only few primary studies are available. We wanted to highlight that what is an unpleasant percentage of convergence issues in our simulations, means not being able to compute an analysis in real-life settings. This point has been made before (Jackson et al., 2018): most of the alternative methods are quite complex and thus prone to convergence difficulties, especially in very sparse data settings. As we have seen in our simulation and has been illustrated elsewhere (see, e.g., Böhning et al., 2015), basing the meta-analysis on more primary studies provides relief to this issue. With regard to our examples, for one outcome (i.e., "serious adverse events") in the review by Hemkens et al. (2016), all four studies that had fit the inclusion criteria were double-zero studies. It is trivial that none of the models were able to provide estimates of the pooled RR, but we wanted to highlight that these kinds of data settings are realistic when investigating rare events. The results for the other selected outcomes are shown in Table 4. Whenever we provided no values in Table 4, model computation failed. That is, the model did not converge properly (as indicated by missing standard errors or warnings provided by the software). Like for the simulation study, the code for the illustrative analyses is also available on the GitHub repository of this paper (`https://github.com/mariebeisemann/metaanalysis_for_rare_events_simulation`); the data are available through the Cochrane Library.

The data for the outcome "cardiovascular mortality" in the review by Squizzato et al. (2017) consisted of seven studies, out of which one was a single-zero and none were double-zero studies. The average sample size across studies was 4,557.6 ($SD = 6667.5, \text{min} = 86, \text{max} = 15,603$). We can see in Table 4 that even though only seven primary studies were included—but in conjunction with in part very large sample sizes—the models that converged provided rather similar estimates of the pooled RR, all indicating no effect. The effects estimated by SJ, SJ_tcc, and in particular poiss were slightly higher than those estimated by the remaining models. 95% CIs were narrower for the standard models and particularly large for kuss_binom. The model fit of poiss and cond_binom was singular, more specifically, the RE variance of the respective model parameter representing the log RR was estimated (close to) 0. The singular fit suggests that assuming a random effect of the log RR is too complex and not supported by the data. This is supported by the standard models all exhibiting insignificant tests for heterogeneity and mostly (except for SJ_tcc) estimating the between-study variance (very close to) 0. We discuss this point further in Section 6. For the same outcome but in the review by Hemkens et al. (2016), we also re-analyzed seven studies, out of which three were single-zero and two were double-zero studies. Average sample size was lower here with 161.7 ($SD = 174.3, \text{min} = 41, \text{max} = 532$). In this sparser data situation, differences between models emerged. We can see in Table 4, that while cond_binom and kuss_binom indicated a significant effect (in the direction of a lower risk in the treatment group; but note here that cond_binom exhibited singular fit), the remaining converged models yielded insignificant results (which aligns with the results in the review by Hemkens et al., 2016, who used an MH

RE model). Point estimates also differed slightly more between models. Finally, we re-analyzed the data for the outcome "fatal stroke," also from the review by Hemkens et al. (2016), for which only four studies had met the inclusion criteria of the review authors, out of which one was a single-zero study and three were double-zero studies (the latter of which we, in alignment with the simulation study, excluded for the standard and the binomial models except for kuss_binom) with an average sample size of 228.8 ($SD = 215.3$, min = 41, max = 532). Here, we saw how pathological results for meta-analyses with rare events can be (see Table 4). Please also note how our examples illustrate the practical challenges faced in the application of the zero-inflation Poisson models, beta_binom and cai_binom in sparse data situations especially with only small numbers of primary studies: these models did not converge for any of the exemplary data sets. Our examples also illustrate the unreliability of the available methods when only few primary studies with a number of zero studies among them are included in a meta-analysis.

6 | DISCUSSION

In the present simulation study, we investigated the performance of different meta-analysis models in terms of estimating the pooled (log) RR across different primary studies in an RE setting in which the event of interest was rare. We examined three recently proposed or discussed models by Böhning et al. (2015) (an RE Poisson regression model, an RE ZIP regression model, and an RE binomial model) as well as compared them to other similar recommended models (beta-binomial models by Cai et al., 2010; Kuss, 2015) and standard RE meta-analysis models. We investigated and compared the models' performance under 162 different simulation conditions distinguished by different underlying true RR, different event occurrence probabilities in the control group (i.e., baseline probabilities), different degrees of heterogeneity, different group size ratios between treatment and control group, and different numbers of primary studies. The data were simulated from two binomial distributions (one per study arm) within each study, and the true underlying RR was drawn as a log RR from a (truncated) normal distribution. Out of all the models, we found the Poisson regression model to have shown the best performance, both in terms of estimating the pooled log RR and also—though not the main focus of this paper—in terms of estimating the between-study heterogeneity as well as in terms of convergence rates. Another model that performed very well, often close to the Poisson regression model and on occasion even better was the beta-binomial model suggested by Kuss (2015). Yet, it is important to point out that settings with only very few primary studies (especially in conjunction with very low baseline risks and few observations) proved difficult for any of the examined models to perform in, also for the Poisson model and the beta-binomial model by Kuss (2015) both of which exhibited considerable bias in these settings. In particular when samples were not extremely large and a lot of zero studies were included in the meta-analysis, our illustrative examples also highlighted this issue: The models showed notable differences in their estimates, a considerable number of them failed to converge, and others indicated singular fit in some cases. Singular fit suggests that the assumption of heterogeneity in the underlying effect is not supported by the data. Yet, in a lot of theoretical settings, the assumption of heterogeneous underlying effects is theoretically valid and called for. When we only have few studies (as does happen and is illustrated by our examples) and especially when a number of them are single-zero or maybe even double-zero studies, modeling such RE might nonetheless be very difficult because the data do not allow to estimate such a complex model structure. This is a challenge in the application of meta-analysis models for rare events that we wanted to highlight with our examples. In the following, we are going to discuss and reflect upon our findings in more detail as well as point out limitations of the present simulation study and give ideas for future research.

6.1 | Convergence

Before even discussing our findings regarding the models' performance in estimating the pooled (log) RR, we need to address the high rates of trials with convergence failures for some models. This concerns mostly the ZIP models (of which we investigated four different variants, differing in the linear combination in the zero-inflation arm of the model), but also the beta-binomial model by Cai et al. (2010). This finding was also reflected in the failure of these models to converge for any of our illustrative examples. The practical issue of convergence in real-life applications of alternative meta-analyses models in rare event settings was also observed and discussed in Jackson et al. (2018), but there in respect to the OR. The former showed relatively better convergence rates for settings without heterogeneity and within those better convergence for meta-analyses with more primary studies. Yet, this comparatively better convergence is of little practical consequence as (a) truly homogeneous settings are unrealistic in real-life applications, and (b) if one were to conduct a meta-analysis in fact in a truly homogeneous setting, one might prefer to revert to actual fixed-effects methods, that is, modeling no RE in the ZIP models. These could likely improve convergence beyond the rates we have observed for the RE models in homogeneous settings. Moreover, convergence rates for the ZIP models were quite poor even in settings where models such as the Poisson model showed only very little bias, for example, in scenarios with an RR of 0.5 or 1, a baseline risk of 0.05 or 0.1, and 100 primary studies (for a group size ratio of 0.5), suggesting that the convergence of these models (or failure thereof) does unfortunately not serve as somewhat of a warning sign that should raise concerns with regard to other models' performance. However, as we are going to discuss below, when the zero-inflated models did converge, their performance was quite good. This may be motivating to investigate possibilities to improve convergence for these models, for example, by employing Bayesian methods with weakly informative prior distributions or using frequentist methods of regularization. Out of the four variants of the ZIP models examined here (see Subsection 2.2), the least complex, that is, the ZIP model with only a fixed intercept in the zero-inflation arm of the model, tended to show the least amount of convergence issues. The beta-binomial model proposed by Cai et al. (2010) showed considerably better performance rates for settings with more heterogeneity. It is parameterized in terms of two parameters, γ and ψ (see Subsection 2.3), none of which may be directly interpreted as the parameter of interest, namely the pooled log RR. The parameter ψ provides a way to adjust the shape of the beta-binomial distribution according to the between-study variation in the treatment effects. The less between-study variation, the larger is ψ (Cai et al., 2010), tending toward infinity for a setting without heterogeneity. If one were interesting in remedying the convergence issues exhibited by the model of Cai et al. (2010)—which we would not necessarily place the utmost importance on, based on the overall rather unsatisfactory performance of this model—one could explore whether other parameterizations might relieve the observed convergence difficulties in settings without heterogeneity. One could of course also argue that this being an RE model, its performance in fixed-effects settings is less relevant and it simply should not be used in such settings.

6.2 | Estimating the pooled (log) RR

The model that performed most consistently well (and often also best in the model comparison) across most measures of performance and most simulation settings was the Poisson regression model, closely followed by the beta-binomial model by Kuss (2015)—a finding that aligns well with previous research (Kuss, 2015; Spittal et al., 2015). For an RR of 0.5, the Poisson regression model tended to perform best and for an RR of 2, the beta-binomial model by Kuss (2015) tended to perform better overall, yet not necessarily in terms of bias. This is in line with findings by Kuss (2015), who also did not find the beta-binomial model to be free of bias in RE settings. However, it should be noted that even these overall very well performing model exhibited extremely problematic performance in very difficult data settings, particularly in meta-analyses with only five primary studies in which the event of interest had a very low baseline probability and even lower occurrence probability in the event group. In fact, a very important observation that we made is that in settings with only five primary studies, in particular in conjunction with smaller treatment groups (i.e., 25 participants in our simulations) and very rare events, all models showed unacceptably large biases. Furthermore, it is also important to pay attention to the value range in the distribution of the estimated log RR. Please note that estimated log RR with an opposite sign to the true effect, indicate that in some trials, the respective model estimated the pooled effect in the opposite direction of the true effect. Looking at the distributions of the models' pooled log RR estimates, we can see that especially for only five primary studies and smaller experimental groups (i.e., 25 participants in our simulations), this is an issue for all of the models that we examined here. When conducting just one meta-analysis in a practical application, this might be a very relevant problem. This should show us quite plainly how even usually well performing and recommended models are not well equipped to be used to conduct a meta-analysis in such a data setting. More importantly, these problems can be remedied rather simply by holding off on conducting the meta-analysis until more primary studies are available. At the very least more than five, even though the present study cannot make more precise recommendations based on our simulation conditions. Unsurprisingly on the basis of prior research (Bradburn et al., 2007; Sweeting et al., 2004; Tang, 2000), the overall performance of the standard meta-analysis models was unsatisfactory, regardless of the estimator for τ^2 used. The alternative continuity correction proposed by Sweeting et al. (2004) tended to lead to even worse performance. The beta-binomial model proposed by Cai et al. (2010) showed overall rather poor performance as well. Especially problematic was the instability of this model, even though, ironically, Cai et al. (2010) proposed their model to alleviate issues with instability due to numerical integration in models such as the binomial model with normal RE. Yet, albeit their model does not rely on numerical integration, their estimator for the pooled RR—which applied researchers will likely be most interested in—does and was, in fact, rather unstable. To make matters worse, we were not able to approximate the integral over the interval $(-\infty, \infty)$ (as intended), so that we instead integrated over the interval $(-100, 100)$ in order to be able to obtain any estimates at all. Clearly, this is only an approximation of the integral we actually wanted to compute. How good this approximation was depends on the distribution of values across the value range. These issues are not addressed in Cai et al. (2010), as they only evaluated the performance of the model in terms of estimating the model parameters γ and ψ. Especially in light of the considerably better and more stable performances of other models, this generally speaks against the beta-binomial model proposed by Cai et al. (2010). The third beta-binomial and the binomial model (Böhning et al., 2015; Stijnen et al., 2010) tended to perform overall less well than the Poisson regression and the beta-binomial model by Kuss (2015) but usually better than the standard meta-analysis models. A conceptual drawback of these models, as discussed in Kuss (2015), is their inability to include double-zero studies that might have contributed to their inferior performance compared to models that are able to include double-zero studies, such as the Poisson regression model or the beta-binomial model recommended by Kuss (2015). For the ZIP model, we found—insofar they actually converged—rather good results. For the Poisson models with a (fixed) treatment effect in the zero-inflation arm of the model, this speaks for the estimator for the pooled RR we suggested based on previous work by Dong et al. (2019). Yet, the model complexity for these models is quite high. This certainly offers an explanation for the high rates of failed convergence. This issue in turn makes our estimates of their performance less reliable and also poses a challenge to the practical application of these models, as highlighted by our illustrative examples.

6.3 | Coverage of the 95% CI for the pooled log RR

The assessment of the coverage of the 95% CI for the pooled log RR did not throw a great light on most models' performance, an observation that has also previously been made in other simulation studies on pooling the RR in a rare-events RE setting (Kuss, 2015). Especially poor was the coverage of the 95% CI for the standard models of meta-analyses, but (at least in some conditions) also for the beta-binomial models except for the beta-binomial model by Kuss (2015). With increasing numbers of primary studies, the coverage was extremely low, in extreme cases even dipping as low as 0%. With larger numbers of primary studies the 95% CIs should become increasingly narrower and at the same time, the standard models consistently show bias. Together, these two tendencies might explain the extremely poor coverage observed for the standard models of meta-analysis. We did not find evidence for the alternative explanation that standard errors were underestimated the standard models, leading to the observed low coverage. In fact, standard errors tended to be overestimated by the standard models. Better results in terms of coverage (at least in some conditions close to the nominal level of 95%.) were exhibited by the ZIP regression models, the Poisson regression model, and the beta-binomial model by Kuss (2015). This is particularly good considering that we newly derived the standard errors for the ZIP models with a (fixed) treatment effect in the zero-inflation arm of the model. Again, there was a tendency of better performance of the beta-binomial model by Kuss (2015) for a true RR of 2, and better performance of the Poisson regression model otherwise. Overall, it is important to highlight that to some degree and with different frequencies, all models diverged notably from nominal level coverage and that in an anticonservative manner, that is, the 95% CI covered the true effect less often than the nominal 95% of the cases. Clearly, this is undesirable and in our opinion, the choice of the meta-analytical model should at least reflect that we can do considerably better in terms of coverage than the standard meta-analytical models. Yet, these findings for the RR stand in contrast to conclusions drawn for the OR by Jackson et al. (2018), who found much better coverage for standard models using the DerSimonian–Laird and the REML estimator. However, in light of these discrepancies between the results for these two effect measures, it should be noted that Jackson et al. (2018) used simulation settings with far fewer primary studies than we did in some of our conditions, with the highest number of primary studies in their simulation being merely 20. Yet, it is only for considerably more primary studies that we see those drastic drops in coverage in our simulations. It might be interesting to explore whether our findings replicate for the OR as an effect measure, especially as Jackson et al. (2018) deem the standard model of meta-analysis as remaining an adequate model choice for meta-analysis of rare events.

6.4 | Limitations and directions for future research

Any implications, conclusions, and recommendations derived from our findings should always be considered carefully in light of the limitations of this study. First of all, as we have already touched upon above, the in part very poor convergence rates for some models, in particular, the ZIP models, resulted in differing precision of simulation estimates between models and conditions. This leads us to one direction for future research and that is exploring whether convergence of the ZIP models can be improved upon, for example, with the help of Bayesian methods. Previous work, which focused on the OR as an effect measure and different models, showed promising results regarding the development of weakly informative priors that improve model performance (Friede et al., 2017; Günhan et al., 2018). Another limitation of our work is constituted by our inability to attain a working implementation in `R` of the beta-binomial model with a log link by Kuss (2015) using the parameterization given in his paper and implemented in `SAS`. Yet, our implementation in `R` did produce results that align well with the results in Kuss (2015) and might constitute a viable alternative for applied researchers using `R`. We did not explore different parameterizations and their comparative performance that might be an interesting endeavor for future research. Further, we did not dedicate a separate simulation to our derived standard errors. For the model by Cai et al. (2010), which we overall do not recommend, this is less important than it might be for the ZIP regression models if their application is further pursued (under the assumption that their convergence rates can be improved). It might also be interesting to explore bootstrapped standard errors.

Based on their simulation study for the OR with simulation conditions assuming a maximum of 20 primary studies, Jackson et al. (2018) concluded that the standard method of meta-analysis remains a viable option. Yet, our findings for the RR of drastically impaired coverage for the standard model originated predominantly in settings with 100 primary studies. Thus, we would recommend also evaluating the standard model's performance with the OR as an effect measure in such settings. We would also like to point out that based on our Cochrane review re-analysis, these settings are not unrealistic. On a similar note, we have overall seen the importance of the number of primary studies included in the meta-analysis for the model performance. Notably, we have observed that even though some of the alternative models and in particular the Poisson regression model, perform very well, they show in part very poor performance for only five primary studies. Yet, our re-analysis of Cochrane reviews as well as the review by Warren et al. (2012) indicated that such low numbers of primary studies included in one meta-analysis are not uncommon. Although we can advise against conducting RE meta-analysis in a rare events settings on the basis of merely five primary studies, our simulations do not enable us to make a recommendation regarding the minimum number of primary studies required to ensure sufficiently stable performance of the employed model. Please note that such a recommendation of a minimum number of primary studies might also vary between models and is not independent of the remaining meta-analytical setting. Such an investigation is especially warranted for the models we wish to recommend, first and foremost the Poisson regression model, but also the beta-binomial model by Kuss (2015).

In our simulation scenarios, we assumed that we had access to all primary studies we generated for the respective meta-analysis. In real life, it is unlikely that this is actually the case. A general challenge for meta-analysis is constituted by a tendency of publishing only studies that yielded significant or otherwise favorable results—a phenomenon that is referred to as publication bias (Harris, Hedges, & Valentine, 2009). This is a problem for meta-analysis as it obviously distorts the estimate of the pooled effect. In a setting of rare events, where it is not unlikely to observe no events at all in any one primary study, we suspect this issue could be particularly relevant. That is, the actual number of single- and double-zero studies could be higher than the number of published single- and double-zero studies. Several other simulation studies have also taken our approach of not considering the publication bias in the design of their simulation conditions (e.g., Bakbergenuly & Kulinskaya, 2018; Jackson et al., 2018; Kuss, 2015), and to our knowledge, no methodological studies have been entirely devoted to study the effect of publication bias in rare-event settings. With regard to real-life meta-analyses, for instance on adverse effects, it has been pointed out that only around 50% considered publication bias at all, and only 20% conducted a quantitative examination to this end (Warren et al., 2012). Only two of 166 meta-analysis reviewed by Warren et al. (2012) corrected for publication bias. It would therefore be important for future research to (a) examine the effect of publication bias in rare-events settings as well as the effect of adjusting for publication bias, and (b) examine recommended models in more realistic simulation settings, including such with different degrees of publication bias. To study the prevalence of (undetected) publication bias in published meta-analyses, the consideration of clinical trials with mandatory preregistration may prove useful. On a similar note, we also simplified our simulation design by not considering different observation times in different studies, which some of the models (e.g., the Poisson model, see Subsection 2.1 for details) can also take into account. It might be interesting to explore if the Poisson model shows a more notably superior performance in settings where observations time vary. Furthermore, while our conditions yielded considerable numbers of single-zero studies, they only included a small number of double-zero studies on average. As we have already seen poor performance in several settings with our simulation study, we would expect to model performance deteriorate further with higher numbers of double-zero studies. We also restricted our simulation conditions to primary study group sizes of 25, 50, or 100. Although these values align well with previous simulation studies that chose values that are representative of the medical literature (e.g., Cheng et al., 2016; Kuss, 2015; Sweeting et al., 2004), studies with much

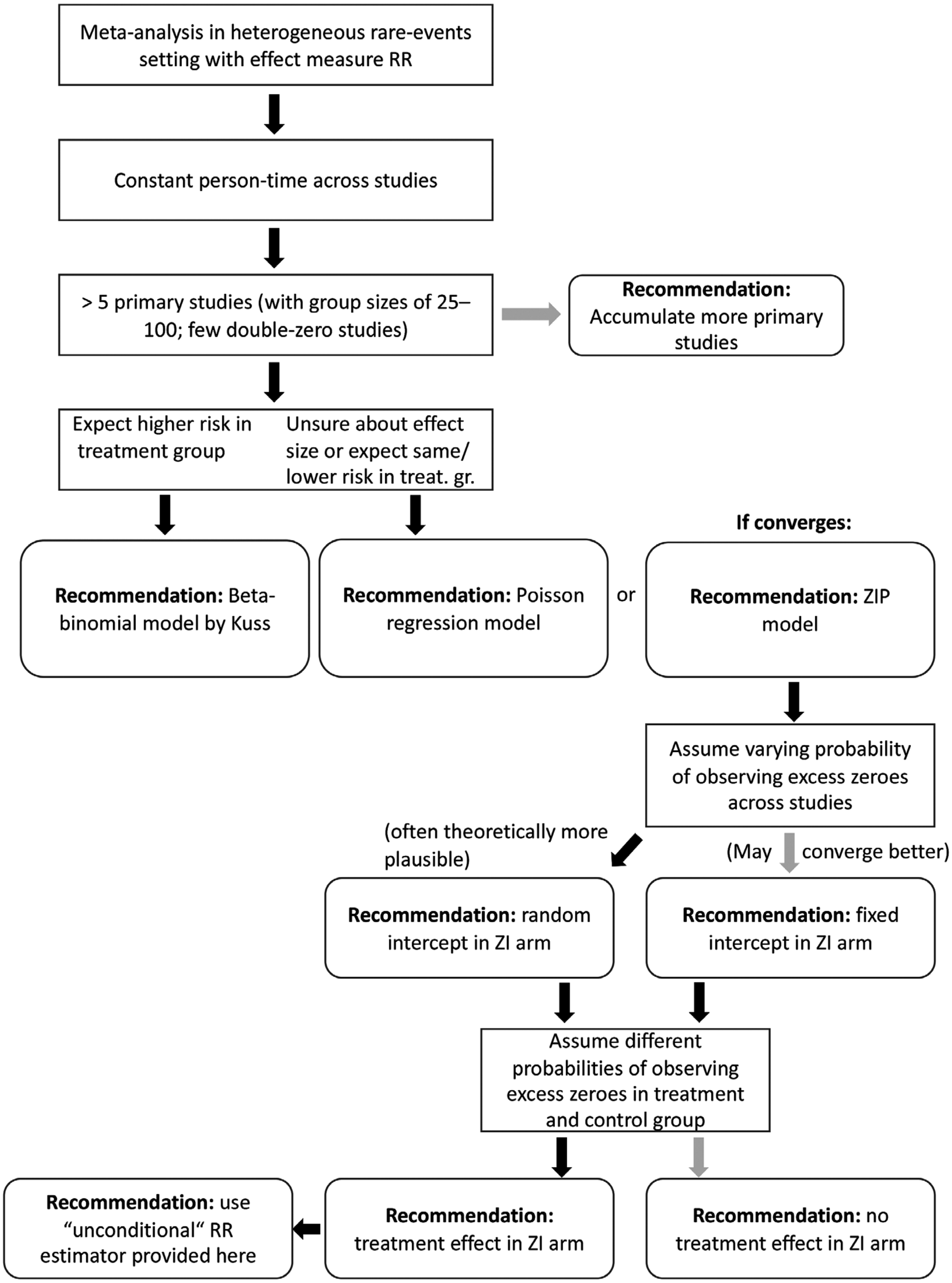
Meta-analysis in heterogeneous rare-events setting with effect measure RR
Constant person-time across studies
> 5 primary studies (with group sizes of 25–100; few double-zero studies)
Recommendation: Accumulate more primary studies
Expect higher risk in treatment group
Unsure about effect size or expect same/ lower risk in treat. gr.
Recommendation: Beta-binomial model by Kuss
Recommendation: Poisson regression model
or
If converges:
Recommendation: ZIP model
Assume varying probability of observing excess zeroes across studies
(often theoretically more plausible)
(May converge better)
Recommendation: random intercept in ZI arm
Recommendation: fixed intercept in ZI arm
Assume different probabilities of observing excess zeroes in treatment and control group
Recommendation: use "unconditional" RR estimator provided here
Recommendation: treatment effect in ZI arm
Recommendation: no treatment effect in ZI arm

FIGURE 9 Flowchart summarizing the practical implications of this simulation study.
Note. Decisions in the data analysis process relevant to our findings are shown in rectangular boxes. Black arrows indicate the answer "yes," gray arrows indicate the answer "no." Recommendations are shown in rectangular boxes with rounded edges (marked as "Recommendation"). Abbreviations: ZIP, zero-inflated Poisson; ZI, zero inflation; treat. gr., treatment group

larger samples do occur. Judging from our illustrative examples, larger samples appears to have aided in the stabilization of the models. It may thus be worthwhile to explore in future research in how far considerably larger sample sizes in the primary studies affect the models' performance.

The choice of the models compared in this simulation study was based on both a literature search of the methodological literature on the topic of meta-analysis in the case of rare events and on findings from a systematic literature review by

Warren et al. (2012). In the latter, Warren et al. (2012) examined which methods of meta-analyses are employed in the medical literature on adverse events, a common example for rare events. We would like to highlight that it would have been preferable to base our model selection on an even more recent search of the medical literature, however, this would have been an undertaking too extensive in scope for the present study. Yet, future research might wish to update the review by Warren et al. (2012) and extend our simulative comparison to methods that are employed in practice but were not included in the present simulation study. Finally, we would also like to stress that our conclusions and recommendations naturally are limited to the models we selected to compare in our simulations. We did not examine the performance of Bayesian models, models from a line of research most notably spear-headed by Shuster and colleagues who distinguish between different variants of RE and developed respective methods (Shuster et al., 2012; Shuster & Walker, 2016), or exact methods. The latter have shown very promising performance for fixed-effects settings (Liu, Liu, & Xie, 2014; Tian et al., 2008)—examined for the OR in those studies but generally also applicable to the RR—but are not as easily extendable to RE settings. If future research extends these methods to RE settings, a comparison of their performance to the best performing models from our simulations would also be very interesting. All of these limitations may constitute avenues for future research.

6.5 | Conclusions and practical recommendations

In summary, on the basis of our simulations, we would overall recommend the Poisson regression model for the meta-analysis of rare events in an RE setting when using the RR as an effect measure. In situations where one expects larger event probabilities in the treatment than in the control group (i.e., an RR greater than 1), we would also recommend using the beta-binomial model by Kuss (2015). The ZIP regression models also performed well, even though the practical drawbacks, that is, the considerable convergence issues, of these models as discussed above should be considered carefully. In agreement with Kuss (2015), we would strongly recommend against employing the standard methods of meta-analysis as there are other methods available with which the performance of the standard models can be considerably improved upon. Another important finding of our simulations is that we would not recommend conducting a meta-analysis with the RR as an effect measure in a rare-events RE setting on the basis of as few as five primary studies, as even our recommended models performed very poorly in such a setting (especially in conjunction with smaller experimental group, that is, 25 participants in our simulations). Poor performance in this context refers to not only atrociously large biases, but also estimates of the pooled effect in the opposite direction of the true effect. The exact minimal number of primary studies necessary has yet to be determined, but our results suggests it is more than 5 and below 30. In accordance with Kuss (2015), our findings have clearly shown that recommendations regarding meta-analysis of the RR in rare-events RE settings by the Cochrane collaboration (Higgins & Green, 2011) are outdated and can be improved upon. We have summarized our recommendations in a flowchart (see Figure 9) to aid practitioners in their decision process when conducting a meta-analysis of rare events. This flowchart may also be helpful for researchers deriving more comprehensive method recommendations.

ACKNOWLEDGMENTS

This work was supported by grant HO 1286/16-1 of the German Research Foundation (DFG) to Heinz Holling. We would like to thank three anonymous reviewers and an associate editor as well as the editor, Dr. Dankmar Boehning, for their helpful and valuable suggestions and comments on earlier drafts of this manuscript.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

ORCID

Marie Beisemann https://orcid.org/0000-0001-6977-167X
Philipp Doebler https://orcid.org/0000-0002-2946-8526
Heinz Holling https://orcid.org/0000-0002-0311-3970

REFERENCES

Bai, O., Chen, M., & Wang, X. (2016). Bayesian estimation and testing in random effects meta-analysis of rare binary adverse events. *Statistics in Biopharmaceutical Research*, *8*, 49–59. https://doi.org/10.1080/19466315.2015.1096823

Bakbergenuly, I., & Kulinskaya, E. (2018). Meta-analysis of binary outcomes via generalized linear mixed models: A simulation study. *BMC Medical Research Methodology*, *18*, 70. https://doi.org/10.1186/s12874-018-0531-9

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bhaumik, D. K., Amatya, A., Normand, S.-L. T., Greenhouse, J., Kaizar, E., Neelon, B., & Gibbons, R. D. (2012). Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association*, *107*, 555–567. https://doi.org/10.1080/01621459.2012.664484

Böhning, D., Mylona, K., & Kimber, A. (2015). Meta-analysis of clinical trials with rare events. *Biometrical Journal*, *57*, 633–648. https://doi.org/10.1002/bimj.201400184

Bradburn, M. J., Deeks, J. J., Berlin, J., & Russell Localio, A. (2007). Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, *26*, 53–77. https://doi.org/10.1002/sim.2528

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., … Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*, 378–400. https://doi.org/10.3929/ethz-b-000240890

Cai, T., Parast, L., & Ryan, L. (2010). Meta-analysis for rare events. *Statistics in Medicine*, *29*, 2078–2089. https://doi.org/10.1002/sim.3964

Cheng, J., Pullenayegum, E., Marshall, J. K., Iorio, A., & Thabane, L. (2016). Impact of including or excluding both-armed zero-event studies on using standard meta-analysis methods for rare event outcome: A simulation study. *BMJ Open*, *6*, e010983. https://doi.org/10.1136/bmjopen-2015-010983

Cox, D. R., (1970). The continuity correction. *Biometrika*, *57*, 217–219.

De Rooi, J. J. (2008). *Smoothing zeros and small counts in meta-analysis of clinical trials* (Master's thesis, Utrecht University, Utrecht, the Netherlands).

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Journal of Statistical Software*, *7*, 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Dong, C., Zhao, Y., & Tiwari, R. (2019). Meta-analysis of clinical trials with sparse binary outcomes using zero-inflated binomial (ZIB) models. *Statistics in Biopharmaceutical Research*, *11*, 1–17. https://doi.org/10.1080/19466315.2018.1537885

Efthimiou, O. (2018). Practical guide to the meta-analysis of rare events. *Evidence-Based Mental Health*, *21*, 72–76. https://doi.org/10.1136/eb-2018-102911

Friede, T., Röver, C., Wandel, S., & Neuenschwander, B. (2017). Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical Journal*, *59*, 658–671. https://doi.org/10.1002/bimj.201500236

Friedrich, J. O., Adhikari, N. K. J., & Beyene, J. (2007). Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Medical Research Methodology*, *7*, 5. https://doi.org/10.1186/1471-2288-7-5

Guevara, J. P., Berlin, J. A., & Wolf, F. M. (2004). Meta-analytic methods for pooling rates when follow-up duration varies: A case study. *BMC Medical Research Methodology*, *4*, 17. https://doi.org/10.1186/1471-2288-4-17

Günhan, B. K., Röver, C., & Friede, T. (2018). Meta-analysis of few studies involving rare events. Preprint, arXiv:1809.04407.

Harris, C., Hedges, L., & Valentine, J., (2009). *Handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Hasselman, B. (2018). nleqslv: Solve Systems of Nonlinear Equations [R package].

Hemkens, L. G., Ewald, H., Gloy, V. L., Arpagaus, A., Olu, K. K., Nidorf, M., … Briel, M. (2016). Colchicine for prevention of cardiovascular events. *Cochrane Database of Systematic Reviews 2016*, *1*, CD011047. https://doi.org/10.1002/14651858.CD011047.pub2

Higgins, J.P.T., & Green, S., (2011). *Cochrane handbook for systematic reviews of interventions*. New York, NY: Wiley.

Jackson, D., Law, M., Stijnen, T., Viechtbauer, W., & White, I. R. (2018). A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine*, *37*, 1059–1085. https://doi.org/10.1002/sim.7588

Jackson, D., & White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, *60*, 1040–1058. https://doi.org/10.1002/bimj.201800071

Keus, F., Wetterslev, J., Gluud, C., Gooszen, H. G., & Van Laarhoven, C. J. H. M. (2009). Robustness assessments are needed to reduce bias in meta-analyses that include zero-event randomized trials. *The American Journal of Gastroenterology*, *104*, 546–551. https://doi.org/10.1038/ajg.2008.22

Kuss, O. (2015). Statistical methods for meta-analyses including information from studies without any events—Add nothing to nothing and succeed nevertheless. *Statistics in Medicine*, *34*, 1097–1116. https://doi.org/10.1002/sim.6383

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1–14. https://doi.org/10.2307/1269547

Lane, P. W. (2013). Meta-analysis of incidence of rare events. *Statistical Methods in Medical Research*, *22*, 117–132. https://doi.org/10.1177/0962280211432218

Li, L., Bai, O., & Wang, X. (2018). An integrative shrinkage estimator for random-effects meta-analysis of rare binary events. *Contemporary Clinical Trials Communications*, *10*, 141–147. https://doi.org/10.1016/j.conctc.2018.04.004

Liu, D., Liu, R. Y., & Xie, M.-G. (2014). Exact meta-analysis approach for discrete data and its application to 2 × 2 tables with rare events. *Journal of the American Statistical Association*, *109*, 1450–1465. https://doi.org/10.1080/01621459.2014.946318

Malzahn, U., Böhning, D., & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*, *87*, 619–632. https://doi.org/10.1093/biomet/87.3.619

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748. https://doi.org/10.1093/jnci/22.4.719

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). truncnorm: Truncated Normal Distribution [R package version 1.0-8].

Microsoft Corporation and Watson, S. (2018). doParallel: Foreach Parallel Adaptor for the 'Parallel' Package [R package version 1.0.14].

Pateras, K., Nikolakopoulos, S., Mavridis, D., & Roes, K. C. B. (2018). Interval estimation of the overall treatment effect in a meta-analysis of a few small studies with zero events. *Contemporary Clinical Trials Communications*, *9*, 98–107. https://doi.org/10.1016/j.conctc.2017.11.012

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rücker, G., Schwarzer, G., Carpenter, J., & Olkin, I. (2009). Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*, *5*, 721–738. https://doi.org/10.1002/sim.3511

Sankey, S. S., Weissfeld, L. A., Fine, M. J., & Kapoor, W. (1996). An assessment of the use of the continuity correction for sparse data in meta-analysis. *Communications in Statistics-Simulation and Computation*, *25*, 1031–1056. https://doi.org/10.1080/03610919608813357

Shuster, J. J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine*, *19*, 1259–1265. https://doi.org/10.1002/sim.3607

Shuster, J. J., Guo, J. D., & Skyler, J. S. (2012). Meta-analysis of safety for low event-rate binomial trials. *Research Synthesis Methods*, *3*, 30–50. https://doi.org/10.1002/jrsm.1039

Shuster, J. J., & Walker, M. A. (2016). Low-event-rate meta-analyses of clinical trials: Implementing good practices. *Statistics in Medicine*, *35*, 2467–2478. https://doi.org/10.1002/sim.6844

Sidik, K., & Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*, 367–384. https://doi.org/10.1111/j.1467-9876.2005.00489.x

Spittal, M. J., Pirkis, J., & Gurrin, L. C. (2015). Meta-analysis of incidence rate data in the presence of zero events. *BMC Medical Research Methodology*, *15*, 42. https://doi.org/10.1186/s12874-015-0031-0

Squizzato, A., Bellesini, M., Takeda, A., Middeldorp, S., & Donadini, M. P. (2017). Clopidogrel plus aspirin versus aspirin alone for preventing cardiovascular events. *Cochrane Database of Systematic Reviews 2017*, *12*, CD005158. https://doi.org/10.1002/14651858.CD005158.pub4

Stijnen, T., Hamza, T. H., & Özdemir, P. (2010). Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*, *29*, 3046–3067. https://doi.org/10.1002/sim.4040

Sweeting, M. J., Sutton, A. J., & Lambert, P. C. (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, *23*, 1351–1375. https://doi.org/10.1002/sim.1761

Tang (2000). Weighting bias in meta-analysis of binary outcomes. *Journal of Clinical Epidemiology*, *53*, 1130–1136. https://doi.org/10.1016/S0895-4356(00)00237-7

Tian, L., Cai, T., Pfeffer, M. A., Piankov, N., Cremieux, P.-Y., & Wei, L. J. (2008). Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 × 2 tables with all available data but without artificial continuity correction. *Biostatistics*, *10*, 275–281. https://doi.org/10.1093/biostatistics/kxn034

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *3*, 1–48. https://doi.org/10.18637/jss.v036.i03

Warren, F. C., Abrams, K. R., Golder, S., & Sutton, A. J. (2012). Systematic review of methods used in meta-analyses where a primary outcome is an adverse or unintended event. *BMC Medical Research Methodology*, *12*, 64. https://doi.org/10.1186/1471-2288-12-64

Wickham, H., François, R., Henry, L., & Müller, K. (2019). dplyr: A Grammar of Data Manipulation [R package version 0.8.1].

Wickham, H., & Henry, L. (2019). tidyr: Easily Tidy Data with 'Spread()' and 'Gather()' Functions [R package version 0.8.3].

Wolodzko, T. (2019). extraDistr: Additional Univariate and Multivariate Distributions [R package version 1.8.11].

SUPPORTING INFORMATION

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

How to cite this article: Beisemann M, Doebler P, Holling H. Comparison of random-effects meta-analysis models for the relative risk in the case of rare events: A simulation study. *Biometrical Journal*. 2020;62:1597–1630. https://doi.org/10.1002/bimj.201900379

APPENDIX

A.1 | First Derivatives of Pooled RR Estimator for cai_binom

$$
\begin{aligned}
\frac{\delta g(\psi,\gamma)}{\delta\psi} = {} & \exp(\theta)(W/(W+\exp(\theta)))^{W\psi}(\exp(\theta)/(W+\exp(\theta)))^{\psi\gamma} \times \\
& \frac{\gamma\log(\exp(\theta)/(W+\exp(\theta)))}{B(\psi\gamma,W\psi)} + \\
& \frac{W\log(W/(W+\exp(\theta))) + W\psi^*(\psi(W+\gamma))}{B(\psi\gamma,W\psi)} + \\
& \frac{\gamma\psi^*(\psi(W+\gamma)) - W\psi^*(W\psi) - \gamma\psi^*(\psi\gamma)}{B(\psi\gamma,W\psi)}
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\delta g(\psi,\gamma)}{\delta\gamma} = {} & \exp(\theta)\psi(W/(W+\exp(\theta)))^{W\psi} \times \\
& (\exp(\theta)/(W+\exp(\theta)))^{\gamma\psi} \times \\
& \frac{(\log(\exp(\theta)/(W+\exp(\theta)))B(\gamma\psi,W\psi) - B(\gamma\psi,W\psi)(\psi^*(\gamma\psi) - \psi^*(\gamma\psi+W\psi)))}{B(\gamma\psi,W\psi)^2},
\end{aligned}
$$

where B(a,b) denotes the beta function. Please note that we denote the digamma function (which is commonly denoted as $\psi(x)$) as $\psi^*(x)$ in the equations above, as we already use ψ to denote the respective parameter used in the model proposed by Cai et al. (2010).