Year : 2020

Meta-analysis of rare events: the challenge of combinïng the lack of information

Piaget-Rossel Romain

Piaget-Rossel Romain, 2020, Meta-analysis of rare events: the challenge of combinïng the lack of information

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_6FFCED4133018

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Centre universitaire de médecine générale et santé publique

Meta-analysis of rare events: the challenge of combining the lack of information

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Romain PIAGET-ROSSEL

Economiste diplômé de la University of British Columbia (UBC)

Jury

Prof. Joachim Marti, Président
Dr. PD Patrick Taffé, Directeur de thèse
Prof. Valentin Rousson, Co-directeur de thèse
Prof. Elena Kulinskaya, Experte
Prof. Martin Huber, Expert

Lausanne, 2020

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Joachim	**Marti**
Directeur·trice de thèse	Monsieur	Dr	Patrick	**Taffé**
Co-directeur·trice	Monsieur	Prof.	Valentin	**Rousson**
Expert·e·s	Madame	Prof.	Elena	**Kulinskaya**
	Monsieur	Prof.	Martin	**Huber**

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Romain Piaget-Rossel

Master of Arts in Economics, University of British Columbia, Canada

intitulée

Meta-analysis of rare events: the challenge of combining the lack of information

Lausanne, le 13 février 2020

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Joachim Marti

Remerciements

Je voudrais tout d'abord témoigner toute ma gratitude à mes directeurs de thèse, qui m'ont, chacun à leur manière, si bien accompagné tout au long de ces trois années. Patrick, merci de m'avoir donné accès à ton savoir encyclopédique et d'avoir su accepter l'émancipation de ton Padawan. Valentin, merci pour toutes ces minutes passées à discuter dans ton bureau de sujets jamais trop éloignés de la statistique et toujours empreints du panache dumasien.

Je désire également remercier Yves Eggli, qui fut le premier à m'accorder sa confiance lors de mes débuts au désormais ex-IUMSP et qui m'a fortement soutenu lors de mes démarches pour obtenir ce post de doctorant, comme à bien d'autres reprises tout au long de mon parcours.

Un projet de longue haleine tel que cette thèse n'aurait sans doute pas été possible sans un excellent environnement de travail. Je tiens donc à remercier Isabella et Aziz, mes deux collègues de bureau, pour leur présence discrète, bienveillante et inspirante.

Mes remerciements vont bien sûr également à ma famille et à mes proches, auprès desquels j'ai toujours pu puiser ma force. Je remercie en particulier ma sœur Marie pour ses innombrables relectures et améliorations de textes tout au long de mon parcours académique. Merci également à mon grand-père, le Professeur Yves Baer, dont la carrière académique a été une grande source d'inspiration.

Merci enfin – et surtout – à ma femme, Lamia Piaget-Rossel, mon éternel soutien dans toutes les étapes de ma vie. Je te dédie ce travail et te remercie d'avoir accepté que je te délaisse parfois quelque peu pour « additionner des zéros et des uns ».

Résumé

Avec des données de comptage ou de taux d'incidence, l'inférence statistique tend à manquer de fiabilité lorsque le nombre d'événements observés est trop faible. Dans ce contexte, il semble donc pertinent de chercher à regrouper plusieurs études afin d'augmenter l'information disponible. Malheureusement, les méthodes standards de méta-analyse ne sont plus valides en présence d'événements rares. Dans cette thèse composée de quatre articles, nous nous sommes intéressés à la difficulté de combiner le manque d'information. À l'aide de simulations, nous avons comparé plusieurs méthodes théoriquement mieux adaptées au phénomène d'événements rares. Nous avons à la fois considéré des méthodes existantes et développé des méthodes innovantes pour des données de comptage et de taux d'incidences. Les résultats obtenus nous ont permis de tirer plusieurs conclusions. Avec des données de comptage et sous l'hypothèse d'homogénéité de l'effet du traitement, la méthode Mantel-Haenszel peut être utilisée quel que soit le niveau de rareté. Une nouvelle méthode basée sur une pseudo-vraisemblance offre des performances similaires à Mantel-Haenszel tout en permettant un gain de précision en présence d'études avec un seul bras de traitement. De plus, contrairement à Mantel-Haenszel, cette méthode de pseudo-vraisemblance peut être étendue au cas d'effet hétérogène du traitement et fournir une bonne estimation de l'effet moyen du traitement ainsi que des intervalles de prédiction informatifs, même en cas d'extrême rareté. Pour ce qui concerne la méta-analyse de taux d'incidence, nous avons démontré que le fait de tenir compte de la sur-dispersion à l'aide d'un modèle binomial-négatif permettait d'améliorer la performance du modèle Poisson, même en présence d'études ne rapportant aucun événement et/ou seulement un bras de traitement.

Abstract

For both count and incidence rate data, it is complicated to provide reliable inference of a treatment effect when the number of observed events is too low. Therefore, the idea of regrouping several studies to increase the amount of available information seems particularly appealing in such settings. Unfortunately, standard meta-analysis methods break down with rare events. This thesis aimed at studying the challenge of combining the lack of information. Throughout four articles, we assessed, via simulations, the performance of several alternative meta-analysis methods that better accommodate rare events. Not only did we consider existing methods, but we also designed innovative methods for both count and incidence rate data. Based on the results obtained in these different papers, we were able to draw several recommendations for applied researchers. With count data, and under the assumption of a homogeneous treatment effect, the Mantel-Haenszel method can be used safely, no matter the scarcity level considered. A newly designed pseudo-likelihood approach performed as well as the Mantel-Haenszel method and allowed a gain of precision when the meta-analysis included studies with missing treatment arms. Moreover, unlike Mantel-Haenszel, this pseudo-likelihood approach could be extended to settings with treatment effect heterogeneity and was shown to provide good estimates of the mean treatment effect and informative prediction intervals, even in extremely rare event settings. As for the meta-analysis of incidence rate data, we found that accounting for over-dispersion using a negative-binomial model allowed for improvements in the performance of the classical Poisson model, even in the presence of studies reporting zero event and/or only one treatment arm.

Introduction

In Biostatistics, one is often interested in the effect of a new treatment on a given population of patients. Since it is generally not timely, costly, and ethically feasible to conduct a randomized controlled trial (RCT) on the whole population of interest, researchers carry out their investigations on a sample drawn from this population, randomly allocating half of their sample to the new treatment and the other half to the usual treatment (or placebo). An estimate of the treatment effect can be obtained by contrasting the outcome obtained in the treated arm (i.e. group of patients receiving the new treatment) with the one obtained in the control arm (i.e. group of patients receiving the usual treatment or a placebo) using a so-called "effect size" (ES). The objective of statistical inference is, then, to generalize this estimate to the population of interest. Provided that the information contained in the sample size is sufficient and that the appropriate statistical techniques are applied, estimates computed in one given sample should be close to the true population parameter value, and confidence intervals (CIs) should have nominal coverage rates.

When analyzing count or incidence rate (IR) data, one extracts the information from the number of observed events. Therefore, it is important to observe a sufficient number of events in order to provide reliable inference. With rare events (i.e. when the probability of experiencing the event in the population of interest is very low), it is often not feasible to conduct a study with appropriate sample size. A priori, running a meta-analysis (MA) [1] in rare event settings seems, thus, particularly sensible since combining the results of several studies (referred to as "primary studies" in the context of a MA) allows to increase the overall sample size and, consequently, the number of observed events. Unfortunately, this is precisely in such settings that the classic inverse-variance-weighting MA method breaks down and statistical issues arise [2].

In practice, researchers meta-analyzing count data often come across rare events [3-5], which explains why many alternative MA methods that accommodate rare events in a better way have been developed. Most of these methods rely on the assumption of a homogeneous treatment effect across the primary studies (i.e. fixed-effect methods). In our first paper, we ran extensive simulations to compare the performance of various existing fixed-effect (FE) methods for count data [6]. Besides the classical inverse-variance-weighting methods, we considered the Mantel-Haenszel (MH) method [7], the Peto method [8], the continuity correction (CC) method [9], the median unbiased estimator method [10], and the binomial regression method. Except for the binomial regression and Peto methods, which are restricted to the odds ratio (OR), all the other methods were used to estimate the OR, the relative risk (RR), and the risk difference (RD), which are the three most commonly used ESs for count data. Compared to previous simulations studies on this subject [11-12], our simulations covered several additional scenarios.

Based on the results of our simulations[i], we concluded our first paper by warmly recommending the use of the MH method without CC for the MA of count data with rare events and homogeneous treatment effect, no matter the ES of interest. Nevertheless, the MH method suffers from theoretical weaknesses related to the formulae used to compute the different ESs' estimators. Indeed, double-zero (DZ) studies (i.e. studies reporting zero event in both treatment arms) do not contribute to the computation of the MH's OR and RR estimators even if they might carry some useful pieces of information regarding the treatment effect. Similarly, studies with only one treatment

[i] To understand the genesis of our second paper, we provided here a little preview of the results obtained in our first paper. Nevertheless, the reader will find all results – from this paper and the three others – in the dedicated Results Section and in the articles themselves.

arm, which we referred to as "single-arm (SA)" studies, are not taken into account by the MH method when computing the OR, the RR, or the RD.

To assess the impact of these exclusions on the MH method's performance, we needed a benchmark method that performed as well as the MH method in scenarios without DZ and SA studies and that was able to include the information contained in both DZ and SA studies. The binomial regression method matches these criteria but is restricted to the OR. A naive way to estimate the two other ESs using the binomial regression method would be to adapt the link function to the ES of interest (i.e. log for the RR, and identity for the RD). However, it is well known that the maximization of the binomial likelihood function using either the log or identity links is plagued by non-convergence issues and valid parameter space violation [13-15]. Therefore, we designed a new MA method based on the pseudo-likelihood (PL) concept, which consists of using working likelihood functions together with their canonical link function (e.g. the RR is estimated using a Poisson likelihood and the log link function). In our second paper, we presented this method and compared it to the MH method, using simulations covering scenarios with SA and DZ studies [16].

As emphasized in their respective Conclusions, our first two papers were restricted to the FE framework, which in practice is not always tenable [17]. In MA, the distinction between FE and random-effects (RE) frameworks is fundamental. Not only does it affect the methodology used, but it also determines the objective of the inference. Indeed, with a homogeneous treatment effect, the inference targets one single parameter (i.e. it is assumed that each primary study estimates the same ES), whereas with treatment effect heterogeneity, the objective of the MA is to summarize the distribution of a population of parameters (i.e. each primary study estimates a different ES drawn from the same

population of ESs). Compared to the MH method, which is restricted to settings with a homogeneous treatment effect, our PL approach can be adapted to account for heterogeneity in the treatment effect. This was precisely the point of our third paper in which we developed an extended PL approach and evaluated its performance under settings with rare events and treatment effect heterogeneity [18].

When the treatment effect is assumed to be heterogeneous, the investigator should focus on the whole distribution of the ES in order to provide a thorough description of the treatment effect. In this respect, a useful and intuitive measure that is underreported in current practice – despite clear recommendations in favor of its use – is the prediction interval (PI) [19-20]. Conventional RE MAs quantify the heterogeneity using an estimate of the between-study variance [21] but it has been shown that this parameter is often underestimated especially with rare events and, thus, that it does not always allow one to describe the treatment effect heterogeneity [22-23]. Therefore, in our third paper, we assessed the performance of the extended PL approach, not only according to the estimation of the mean treatment effect and the between-study variance parameters, but also according to the estimation of a PI.

Our first three papers were dedicated to the MA of count data in rare event settings. In our fourth and last paper, we tackled the issue of rare event MA of IR data [24]. The developments presented in this paper were motivated by an applied project on which we collaborated with the Lausanne University Hospital [25]. In this project, the physicians were interested to study the effect of the width of surgical margin on the rate of recurrence of phyllodes tumors of the breast. The systematic review that was conducted gathered both DZ and SA studies, making the classical weighting method unusable. Moreover, based on the physicians' contextual knowledge, the treatment

effect was assumed to be heterogeneous, which prevented us to use the MH method as well. Our first idea to meta-analyze these data was, thus, to use the random-effects Poisson (Re-Poi) model in an either univariate or bivariate approach – the latter being more flexible since involving one additional parameter [26-27]. Indeed, the Re-Poi model allows for the inclusion of both DZ and SA studies and can account for treatment effect heterogeneity. However, it is based on the equi-dispersion assumption (i.e. mean = variance), which was likely to be violated in the problem at-hand as the observational nature of the data increased the risk of having unmeasured individual characteristics differing within studies. Therefore, the use of a random-effects negative binomial (Re-NB) model seemed more appropriate since the negative binomial distribution allows for over-dispersion.

Despite several investigations in MA literature, we did not find any application of the Re-NB model for meta-analyzing IR data with a heterogeneous treatment effect, let alone in the context of rare events and SA studies. Moreover, although the Re-NB model seemed more appropriate than the Re-Poi model to analyze our data, we were not certain that it would provide good results in such an adverse setting (i.e. SA and DZ studies). Indeed, the more flexible the model is, the greater its complexity, and the harder it is to estimate it. Therefore, we decided to run simulations calibrated on our clinical dataset to assess the performance of the univariate and bivariate Re-NB models and to compare it to the performance of the univariate and bivariate Re-Poi models. Results of our simulations, along with the presentation of these four models, were provided in our fourth paper.

In what follows, we will start by providing a brief summary of the main results obtained in our four papers. We will then discuss these results, elaborate on their potential contribution to the literature of rare event MA, and make some concluding remarks. Our

four articles will form the last part of this manuscript and will be provided after the list of references for this introductory text.

Results

Paper 1

In our first paper [6], we found that the MH method without CC provided the most reliable estimates and CIs under all the scenarios considered. These findings applied to both homogeneous and heterogeneous baseline event probability and no matter the ES considered (i.e. OR, RR, and RD). The binomial regression method performed as well as the MH without CC method but was restricted to the OR. Performance of the other investigated methods was as follows. The classic inverse-variance-weighting method was found to perform poorly, even in settings with rather not so rare events. The Peto method obtained reliable estimates of the OR and provided CIs for this parameter achieving nominal coverage rates only in settings with balanced treatment arms and moderate treatment effect. The use of a CC to include the information contained in zero-event studies when using the inverse variance, Peto, or MH methods deteriorated the results obtained by these methods, especially for the two latter. As for the median unbiased estimator method, although theoretically appealing since it allows for the inclusion of zero-event studies without using a CC, it also failed to provide reliable results.

Paper 2

In our second paper [16], we developed a PL approach to account for SA and DZ studies and found that it performed very well across all simulated scenarios and no matter the ES considered (i.e. OR, RR, and RD). In order to obtain a Wald's CI with nominal coverage rates for the RD, we had to compute a calibrated CI in the spirit of the Hartung-Knapp method [28]. Nevertheless, this calibrated interval turned out to provide consistently

good results in all the simulated scenarios, as long as the proportion of SA studies remained below 75%. Moreover, by varying the underlying distribution of the baseline probability, we found that the PL approach was robust to the choice of this distribution. Finally, whereas the deletion of DZ studies had, surprisingly, no impact on the MH method's performance, our results showed that the fact of not taking into account SA studies resulted in a loss of precision for this method, as compared to the PL approach.

Paper 3

In our third paper [18], we extended the PL approach to the setting of a heterogeneous treatment effect and found that it provided good results for the estimation of the mean treatment effect and its confidence interval, even when meta-analyzing extremely rare events. These good results applied, no matter the ES considered (i.e. OR, RR, and RD). Regarding the between-study variance parameter, it was largely underestimated in rare event settings. With extremely rare events, the between-study variance was estimated to be zero in almost all simulated scenarios, even those with a genuinely high level of heterogeneity. Inversely, we found that the PI was able to better describe the treatment effect heterogeneity. Indeed, in most of the settings, and even in those where the between-study variance was estimated to be zero, the PIs were conservative. Results obtained by the PL approach were robust to the genuine underlying distribution of the ES, as long as the latter was not too asymmetrical.

Paper 4

In our fourth paper [24], we found that the Re-NB model was more performant than the Re-Poi model for MA of over-dispersed IR data with rare events and SA studies. The bivariate version of the Re-NB model gave better results than its univariate counterpart, though at the cost of more numerical issues. Finally, we found that the information available in this dataset, on which our simulations were calibrated, was not sufficient

(i.e. too few studies, too rare events, and too many SA studies) to provide reliable estimates of the between-study variance parameter, no matter the model used.

Discussion

A priori, the idea of meta-analyzing primary studies with rare events seems particularly appealing since it allows one to increase the sample size and, thus, the number of observed events. Yet, classical MA methods break down when the number of events observed in each primary study is small. Therefore, running a MA in the presence of rare events turns out to be a particularly complicated task. Our objective was to tackle the challenge of combining the lack of information contained in several studies in order to provide new guidance to meta-analysts dealing with rare events.

In this thesis, we started by assessing the performance of several FE MA methods that had been developed so far to accommodate rare events. The objective was to understand their strengths and, more importantly, their limitations. Then, we designed a MA approached based on the PL, which allows for the MA of the three most commonly used ESs with count data (i.e. OR, RR, and RD) in both treatment effect homogeneity and heterogeneity settings while including the information contained in DZ and SA studies. In addition to these developments for count data, we also proposed a Re-NB model for the MA of over-dispersed IR data with rare events and SA studies.

Based on the results obtained in our first two papers [6, 16], we are able to draw clear recommendations regarding rare event MA of count data in the context of treatment effect homogeneity. When there are no SA studies, the PL or the MH without CC methods can be used interchangeably to meta-analyze the OR, the RR, and the RD, even with very low event probability. In the presence of SA studies, the PL approach should be favored to obtain more precise estimates. These recommendations are valid for both rare event

and common event settings and, thus, apply for any meta-analysis conducted under the assumption of a homogeneous treatment effect.

We hope that these clear recommendations will help improving current practices. However, we know that there is a lot of inertia in research and that bridging the gap between theory and practice is a long-lasting process. For instance, using a CC (i.e. adding an arbitrary constant to each cell of the zero-event studies' contingency table) is still the default implementation in `Stata` [29], despite several authors arguing against its use [30-32]. We corroborated their arguments by showing that using a CC could heavily deteriorate the MH estimators' performance. Moreover, although the Peto method is restricted to the OR and provides only an approximation of this parameter [33-34], it is still recommended by Cochrane's guidelines to meta-analyze rare events [35]. Our results supported the conclusion that the Peto method should only be used with great caution since, unlike the MH method without CC and the PL approach, it did not perform well in all settings – as we showed.

When the treatment effect is assumed to be heterogeneous, things are more complicated since inference is not about one single parameter anymore but about the whole distribution of the treatment effect. Therefore, unlike what has been done in a recent simulation study where the focus was restricted to the mean treatment effect and its CI only [36], performance assessment of a MA method should include estimation of some measure of the heterogeneity. In our third paper, we used both the between-study variance and the PI to describe the heterogeneity and evaluate the estimates obtained by our PL approach for these two quantities [18].

Two key messages can be taken from the results obtained in this third paper. The first one is that a MA conducted with treatment effect heterogeneity should always report a

PI. Not only is this measure very intuitive, but, as our simulations showed, it also provides a better description of the treatment effect than the estimation of the between-study variance parameter. Indeed, whereas the between-study variance was often underestimated, especially with rare events, the computed PIs were always non-null and often conservative. This is because the formula used to compute these PIs involves both the between and within-study variances, and over-estimation of the latter tended to compensate under-estimation of the former. Underestimation of the between-study variance parameter corroborated previous findings. For instance, Engels and colleagues reanalyzed 125 published MA studies and found that 26% of them estimated a null between-study variance [37]. However, the good performance of the PI was in contradiction with what was described in previous studies [38-40]. This can be explained by the fact that these authors have used a two-stage approach, where the ESs and the variances components are first estimated and then used in further computations of other parameters. By contrast, the PL is a one-stage approach in which all parameters are estimated simultaneously [41]. Therefore, the first key message should be complemented as follows: always report the PI in RE MA and use a reliable one-stage method to conduct the analysis.

The second take-home message of our third paper is that the PL method is a reliable one-stage method to run a MA of rare events with treatment effect heterogeneity. Indeed, our results showed that it provided good estimates and CIs for the mean treatment effect parameter, even in settings with extremely rare events. As for the PIs computed with the PL approach, although (very) conservative with (very) rare events, these intervals can still serve as an interesting benchmark, allowing the exclusion of unlikely values of the treatment effect.

Beside the results that the Re-NB model was more appropriate to meta-analyze over-dispersed IR data than the Re-Poi model even in the presence of DZ and SA studies, another interesting finding of our fourth paper was the incapacity of all models to provide reliable estimates of the between-study variance parameter [24]. This corroborates one of the findings of our third paper, i.e., that estimation of this parameter fails to provide a comprehensive description of the treatment effect heterogeneity in rare event settings, and, thus, it also speaks in favor of reporting the PI. Moreover, the high rate of numerical issues of the bivariate Re-NB which was, to a lesser extent, also observed for the PL approach, indicates that, in practice, these models might have to be simplified in order to achieve convergence.

In conclusion, we showed throughout our four articles that solutions to obtain reliable results in rare event MA existed. This was particularly true in the context of a homogeneous treatment effect, where the PL and MH methods obtained good performance in all the settings covered by our simulations (the MH being slightly less precise with SA studies). When the treatment effect cannot be assumed to be homogeneous, meta-analysts should use an appropriate one-stage RE MA method, such as the PL approach for count data or the Re-NB model for IR data, and always report a PI to describe the heterogeneity. Nevertheless, it is important to keep in mind that when the events are too rare the PI can greatly overestimate the genuine heterogeneity. Moreover, in practice, one might have to simplify the estimation model because of numerical issues.

The difficulty of precisely describing the treatment effect heterogeneity and of achieving convergence of the model brings us to the limit of MA in rare event settings. Indeed, when the events are too rare, the information contained in primary studies might not be

sufficient to provide reliable estimates of all the model's parameters or even to achieve convergence. In such settings, one solution to gain information would be to resort to observational studies, which are usually based on much larger samples than RCTs and which can provide valid estimate of the treatment effect provided that the appropriate causal methodology has been used [42]. Nevertheless, further investigations are required in order to better understand the benefits of including causal observational studies when meta-analyzing rare events.

References

[1] Normand SLT. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* 1999; **18**(3): 321-359.

[2] Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal* 2018; **60**(6): 1040-1058.

[3] Kuss O, Wandrey M, Kunze M. How frequent are meta-analyses with "double-zero" studies in systematic reviews?. 2009. Available from: http://www.egms.de/static/de/meetings/gmds2009/09gmds155.shtml [Accessed on 30 November 2019].

[4] Vandermeer B, Bialy L, Hooton N, Hartling L, Klassen TP, Johnston BC, Wiebe N. Meta-analyses of safety data: a comparison of exact versus asymptotic methods. *Statistical Methods in Medical Research* 2009; **18**(4): 421-432.

[5] Salpeter SR, Greyber E, Pasternak GA, Salpeter PEE. Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. *Cochrane Database Systematic Reviews* 2010; (1): CD002967.

[6] **Piaget-Rossel R**, Taffé P. Meta-analysis of rare events under the assumption of a homogeneous treatment effect. *Biometrical Journal* 2019; **61**: 1557-1574.

[7] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**(4): 719-748.

[8] Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* 1985; **27**(5): 335-371.

[9] Plackett RL. The continuity correction in 2 x 2 tables. *Biometrika* 1964; **51**(3/4): 327-337.

[10] Parzen M, Lipsitz S, Ibrahim J, Klar N. An estimate of the odds ratio that always exists. *Journal of Computational and Graphical Statistics* 2002; **11**(2): 420-436.

[11] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; **23**(9): 1351-1375.

[12] Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine* 2007; **26**(1): 53-77.

[13] Marschner IC, Gillett AC. Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics* 2012; **13**(1): 179-192.

[14] Fitzmaurice GM, Lipsitz SR, Arriage A, Sinha D, Geenberg C, Gawande AA. Almost efficient estimation of relative risk regression. *Biostatistics* 2014; **15**(4): 745-756.

[15] Carter RE, Lipsitz SR, Tilley BC. Quasi-likelihood estimation for relative risk regression models. *Biostatistics* 2005; **6**(1): 39-44.

[16] **Piaget-Rossel R**, Taffé P. A pseudo-likelihood approach for the meta-analysis of homogeneous treatment effect: exploiting the information contained in double-zero and single-arm studies. *Journal of Statistics: Advances in Theory and Applications* 2019; **21**(2): 91-117.

[17] Kontopantelis E, Springate DA, Reeves D. A re-analysis of the Cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *Plos One* 2013; **8**(7): e69930.

[18] **Piaget-Rossel R**, Taffé P. A pseudo-likelihood approach for the meta-analysis of rare events with treatment effect heterogeneity. Submitted to *Statistical Methods in Medical Research*.

[19] Higgins JPT, Thompson SG, Spiegelhalter J. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society A* 2009; **172**(1): 137-159.

[20] IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* 2016; **6**(7): e010247.

[21] Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JPT, Langan D, Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods* 2016; **7**(1): 55-79.

[22] Bakbergenuly I, Kulinskaya E. Meta-analysis of binary outcomes via generalized linear mixed models: a simulation study. *BMC Medical Research Methodology* 2018; **18**: 70.

[23] Chung Y, Rabe-Hesketh S, Choi IH. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine* 2013; **32**(23): 4071-4089.

[24] **Piaget-Rossel R**, Taffé P. Meta-analysis of incidence rate data in the presence of zero-event and single-arm studies. *International Journal of Statistics in Medical Research* 2019; **8**: 57-66.

[25] Toussaint A, **Piaget-Rossel R**, Stormacq C, Mathevet P, Lepigeon K, Taffé P. Width of margins in phyllodes tumors of the breast: the controversy drags on? A systematic review and meta-analysis. Submitted to *Breast Cancer Research and Treatment*.

[26] Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. *Biometrical Journal* 2015; **57**(4): 633-48.

[27] Stijnen T, Hamza HT, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine* 2010; **29**(29): 3046-67.

[28] Röver C, Knapp F, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology* 2015; **15**: 99.

[29] StataCorp. Stata Statistical Software: Release 15. 2017. College Station, TX: StataCorp LP.

[30] Keus F, Wetterslev J, Gluud C, Gooszen HG, van Laarhoven CJ. Robustness assessments are needed to reduce bias in meta-analyses that include zero-event randomized trials. *The American Journal of Gastroenterology* 2009; **104**(3): 546-551.

[31] Kuss O, Gummert JF, Borgermann J. Meta-analysis with rare events should use adequate methods. *The Journal of Thoracic and Cardiovascular Surgery* 2008; **136**(1): 241.

[32] Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* 2009; **28**(5): 721-738.

[33] Brockhaus AC, Bender R, Skipka G. The Peto odds ratio viewed as a new effect measure. *Statistics in Medicine* 2014; **33**(28): 4861-4874.

[34] Brockhaus AC, Grouven U, Bender R. Performance of the Peto odds ratio compared to the usual odds ratio estimator in the case of rare events. *Biometrical Journal* 2016; **58**(6): 1428-1444.

[35] Higgins JPT, Green S. (eds). *Cochrane Handbook for Systematic Reviews of Interventions Version 6*. The Cochrane Collaboration, 2019. Available from: https://training.cochrane.org/handbook/current [Accessed on 30 November 2019].

[36] Kuss O. Statistical methods for meta-analyses including information from studies without any events – add nothing to nothing and succeed nevertheless. *Statistics in Medicine* 2015; **34**(7): 1097-1116.

[37] Engels ES, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* 2000; **19**(13): 1707-1728.

[38] Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. *Statistics in Medicine* 2016; **36**(2): 301-317.

[39] Wang CC, Lee WC. A simple method to estimate prediction intervals and predictive distributions: summarizing meta-analyses beyond means and confidence intervals. *Research Synthesis Methods* 2018; **10**(2): 255-266.

[40] Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. *Statistical Methods in Medical Research* 2019; **28**(6): 1689-1702.

[41] Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Statistics in Medicine* 2017; **36**(5): 855-75.

[42] Gray E, Marti J, Brewster DH, Wyatt JC, **Piaget-Rossel R**, Hall PS. Real-world evidence was feasible for estimating effectiveness of chemotherapy in breast cancer: a cohort study. *Journal of Clinical Epidemiology* 2019; **109**: 125-132.

Received: 16 December 2018 | Revised: 27 March 2019 | Accepted: 1 April 2019

DOI: 10.1002/bimj.201800381

Biometrical Journal

RESEARCH PAPER

Meta-analysis of rare events under the assumption of a homogeneous treatment effect

Romain Piaget-Rossel | **Patrick Taffé**

Institute of Social and Preventive Medicine, Lausanne University Hospital and Lausanne University, Lausanne, Switzerland

Correspondence
Romain Piaget-Rossel, Institute of Social and Preventive Medicine, Lausanne University Hospital and Lausanne University, Route de la Corniche 10, 1010 Lausanne, Switzerland.
Email: romain.piaget-rossel@unil.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 200021_169504/1

Abstract

We studied the performance of several meta-analysis methods in rare event settings, when the treatment effect is assumed to be homogeneous and baseline prevalences are either homogeneous or heterogeneous. We conducted extensive simulations that included the three most common effect sizes with count data: the odds ratio, the relative risk, and the risk difference. We investigated several important scenarios by varying the level of rareness, the value of the trials' arms unbalance, and the size of the treatment effect. We found that the Mantel–Haenszel method and the Binomial regression model provided the best results across all the scenarios investigated. The Peto method performed satisfactorily only when the true effect size was not too large and the degree of unbalance moderate. Inverse variance was the least reliable method. The use of a continuity correction factor slightly improved the performance of the inverse variance method but deteriorated that of the Peto and Mantel–Haenszel methods. A method based on median unbiased estimators of the probabilities provided similar results to those obtained when using the inverse variance method with a continuity correction. Therefore, when the treatment effect can be assumed to be homogeneous and for either homogeneous or heterogeneous baseline prevalences, we highly recommend using the Mantel-Haenszel method without continuity correction (for all the effect sizes) or the Binomial regression model (for the odds ratio only) to meta-analyze the data.

KEYWORDS

fixed-effect methods, homogeneous treatment effect, meta-analysis, rare events, simulation study

1 | INTRODUCTION

The objective of a meta-analysis is to combine evidence from related but independent studies in order to improve the knowledge about a specific research question and to generalize the results (Normand, 1999). When the outcome is continuous, the investigator either uses a fixed-effect (FE) or a random-effects (RE) model according to his knowledge and the expected effect of the intervention. However, with count data, the investigator further has to distinguish between the setting of homogeneous and heterogeneous baseline prevalences, yielding a total of four different frameworks (Table 1).

While a clear distinction is made between the FE and RE models in the meta-analysis' literature, the question of homogeneous versus heterogeneous baseline prevalences is seldomly discussed. Notably, in the three main published simulation studies assessing the performance of various meta-analysis methods in the context of rare events, the authors assumed a homogeneous baseline prevalence in all the simulated scenarios (Bradburn, Deeks, Berlin, & Russell Localio, 2007; Kuss, 2015; Sweeting,

TABLE 1 Frameworks for the meta-analysis of count data

	Homogeneous treatment effect	Heterogeneous treatment effects
Homogeneous baseline prevalence	Fixed effect (FE) with baseline heterogeneity	Random effects (RE) with baseline homogeneity
Heterogeneous baseline prevalences	FE with baseline heterogeneity	RE with baseline heterogeneity

Sutton, & Lambert, 2004). The goal of this paper was thus to reassess the performance of FE meta-analysis methods with rare events, considering the two settings of homogeneous versus heterogeneous baseline prevalences.

In practice, a typical setting under which the assumption of a homogeneous treatment effect is likely to hold is that of a multicenter randomized control trial, where each center has followed exactly the same research protocol. When patients between centers are comparable in terms of baseline characteristics, then the meta-analyst will likely assume a homogeneous baseline prevalence. On the contrary, when patients' baseline characteristics differed markedly, the assumption of heterogeneous baseline prevalences might be more appropriate.

The classical method to conduct a meta-analysis in the framework of a homogeneous treatment effect is that of the inverse variance (IV) (Borenstein, Hedges, Higgins, & Rothstein, 2009; Jackson & White, 2018). This method is based on the generalized least squares technique and provides an estimator corresponding to a weighted average of the primary-studies' effect sizes (ESs). This estimator has good asymptotic properties (i.e., convergence, asymptotic normality). In finite samples however, and especially with rare events, asymptotic theory breaks down and the IV method yields biased estimates and invalid confidence intervals (CI) (Lane, 2013). Specific to binary data, two other methods are commonly used in FE meta-analyses: Mantel–Haenszel (MH) (Mantel & Haenszel, 1959) and Peto (Yusuf, Peto, Lewis, Collins, & Sleight, 1985), the latter being applied for the estimation of the odds ratio (OR) only. While both methods were found to be more robust to rare events than the IV method, estimates provided by the Peto method were less reliable in settings with unbalanced trials' arms and/or large ES (Bradburn et al., 2007).

With very rare events, when some primary studies report zero event in one (single-zero studies; SZ studies) or both (double-zero studies; DZ studies) arms, the IV method may provide indefinite estimates. Under these circumstances, a straightforward way to compute the IV estimate is to exclude the problematic studies from the analysis. As for the MH and Peto methods, DZ studies are automatically excluded from the computation of the OR and the relative risk (RR; for MH), as can be checked by inspecting their mathematical formulae. However, excluding zero-event studies is suboptimal since they are likely to contain useful pieces of information, even DZ studies (e.g., no event in a sample of 50 patients is not the same as no event in a sample of 200 patients).

As discussed by Kuss, Wandrey, and Kunze (2009), meta-analyses with zero-event studies are commonly encountered in practice. In a random sample of 500 Cochrane reviews, the authors found that 34% of these reviews contained at least one meta-analysis with a DZ study. To tackle this issue, the classical solution is to use a continuity correction factor (CC) (Sweeting et al., 2004). The problem with this solution is that the resulting estimates depend on the choice of the CC (Rücker, Schwarzer, Carpenter, & Olkin, 2009). However, there is another way to include both SZ and DZ studies, without using a CC, which is based on the median unbiased estimator (MUE) method (Hirji, Tsiatis, & Mehta, 1989; Parzen, Lipsitz, Ibrahim, & Klar, 2002). This method provides estimates of the OR, RR, and risk difference (RD) that always exist (i.e., even when no event is observed in both arms). Finally, the last alternative we have investigated to estimate the OR is the Binomial regression model.

Through extensive simulations, we assessed the performance of these different methods (i.e., IV, MH, Peto, with or without CC, MUE, and Binomial regression). Except for the Peto and Binomial regression methods that are specific to the OR, all the other methods were used to estimate the OR, the RR, and the RD. Our simulations covered many important scenarios with different values of trials' arms unbalance (from strong unbalance in favor of either trials' arm to no unbalance), ES (from large reduction—or increase—in event prevalence to no effect), and baseline prevalences (from extremely rare to common events). On top of that, all these scenarios were considered with either homogeneous or heterogeneous baseline prevalences.

2 | COMBINING TRIALS UNDER THE FIXED-EFFECT FRAMEWORK

Let π_t and π_c be the probability of the event in the treated and control populations, respectively. In this paper, we focused on the three following effect sizes:

$$OR = (\pi_t * (1 - \pi_c))/(\pi_c * (1 - \pi_t)) \quad (1)$$

TABLE 2 Scale-dependency of the notion of homogeneity in treatment effect with heterogeneous baseline prevalences

π_c	π_t	RD	RR	OR
0.16	0.26	0.1	1.63	1.84
0.33	0.43	0.1	1.30	1.53
0.03	0.13	0.1	4.45	4.83
0.17	0.27	0.1	1.59	1.81
0.07	0.17	0.1	2.39	2.72

Note. The treatment is assumed to be homogeneous on the RD-scale.

$$RR = \pi_t / \pi_c \tag{2}$$

$$RD = \pi_t - \pi_c. \tag{3}$$

Under the FE model, the ES is assumed homogeneous across the primary studies and the goal of the meta-analysis is to estimate a single population parameter. Inversely, under the RE model, one assumes that each primary study seeks to estimate a different population parameter and the focus is on describing these parameters' distribution. Ideally, the selection of either framework should be grounded on contextual knowledge and not on statistical arguments. This selection will then determine the set of methods available to the meta-analyst. In the subsections below, we described six methods that can be used under the FE framework.

As already mentioned in the Introduction, when meta-analyzing count data in the FE framework, one should further distinguish between the settings of homogeneous and heterogeneous baseline prevalences (see Table 1). This subdivision has a direct implication regarding the notion of treatment effect homogeneity. With homogeneous baseline prevalences, the homogeneity of the ES holds whatever the metric adopted (OR, RR, or RD). On the contrary, with heterogeneous baseline prevalences, homogeneity of the ES depends on the scale of measurement. Indeed, assume for instance that the baseline probability π_c is heterogeneous and that the RD is homogeneous. Then, if one were to use instead the RR (or the OR), the ES estimate would turn out be heterogeneous (Table 2). Therefore, under the FE framework with baseline heterogeneity, the investigator has to specify on which scale the treatment effect is assumed homogeneous.

As this small example illustrates, the selection of the appropriate scale to measure the effect of the treatment is an important question. Actually, there is a broad debate in the literature on the advantages of absolute measures, such as the RD, versus relative measures, such as the RR and OR (Papageorgiou, Tsiranidou, Antonoglou, Deschner, & Jäger, 2015; Sinclair & Bracken, 1994).

In what follows, each primary study consists of one control group of size n_c and one treated group of size n_t. The number of events occurring in these two groups are denoted X_c and X_t, respectively.

2.1 | The inverse variance method

For $k = 1, \ldots, K$, the IV method is based on the following model:

$$\hat{\theta}_k = \theta + \epsilon_k, \quad \epsilon_k \sim N\left(0, \sigma_k^2\right), \tag{4}$$

where $\hat{\theta}_k$ is the estimator of the parameter of interest θ (i.e., RD, RR, or OR) obtained from study k and σ_k^2 its variance. The RR and OR's estimators are usually analyzed on the log scale, as their sampling distribution is more symmetrical on this scale.

The three ES estimators are computed as follows:

$$\widehat{OR}_k = \frac{\hat{\pi}_{kt} * \left(1 - \hat{\pi}_{kc}\right)}{\hat{\pi}_{kc} * \left(1 - \hat{\pi}_{kt}\right)} \tag{5}$$

$$\widehat{RR}_k = \frac{\hat{\pi}_{kt}}{\hat{\pi}_{kc}} \tag{6}$$

$$\widehat{RD}_k = \hat{\pi}_{kt} - \hat{\pi}_{kc}, \tag{7}$$

where $\hat{\pi}_{kt} = X_{kt}/n_{kt}$ and $\hat{\pi}_{kc} = X_{kc}/n_{kc}$ are the maximum likelihood estimators.

Variances of these estimators, computed using the delta method for the OR and the RR, are estimated by

$$\hat{\sigma}^2_{log\widehat{OR}_k} = \frac{1}{n_{kt}\hat{\pi}_{kt}\left(1-\hat{\pi}_{kt}\right)} + \frac{1}{n_{kc}\hat{\pi}_{kc}\left(1-\hat{\pi}_{kc}\right)} \tag{8}$$

$$\hat{\sigma}^2_{log\widehat{RR}_k} = \frac{1-\hat{\pi}_{kt}}{n_{kt}\hat{\pi}_{kt}} + \frac{1-\hat{\pi}_{kc}}{n_{kc}\hat{\pi}_{kc}} \tag{9}$$

$$\hat{\sigma}^2_{\widehat{RD}_k} = \frac{\hat{\pi}_{kt}\left(1-\hat{\pi}_{kt}\right)}{n_{kt}} + \frac{\hat{\pi}_{kc}\left(1-\hat{\pi}_{kc}\right)}{n_{kc}}. \tag{10}$$

The IV estimator is obtained by applying the generalized least squares method, assuming independence between the K primary studies and the within-study variances σ^2_k known:

$$\hat{\theta}_{IV} = \frac{\sum_{k=1}^{K} W_k\hat{\theta}_k}{\sum_{k=1}^{K} W_k}, \tag{11}$$

where $W_k = 1/\sigma^2_k$. This estimator is equivalent to the maximum likelihood estimator whenever the assumption of normality (4) holds. The variance of $\hat{\theta}_{IV}$ is given by

$$Var\left(\hat{\theta}_{IV}\right) = \frac{1}{\sum_{k=1}^{K} W_k}. \tag{12}$$

In practice, σ^2_k is not observed and $\hat{W}_k = 1/\hat{\sigma}^2_k$ is used instead.

As discussed in the Introduction section, the good performances of the IV method hold asymptotically. However, in finite samples, estimates obtained with this method systematically deviate from the true parameter value and observed coverage probabilities of the CI depart from the nominal value. These undesirable properties, which are exacerbated with rare events, can be explained by several reasons. First, primary-study estimators of the RR and OR are biased (Firth, 1993; Nemes, Jonasson, Genell, & Steineck, 2009). Therefore, even when σ^2_k is known, one has for RR and OR:

$$E\left(\hat{\theta}_{IV}|\sigma^2_k\right) = \frac{\sum_{k=1}^{K} W_k E\left(\hat{\theta}_k\right)}{\sum_{k=1}^{K} W_k} = E\left(\hat{\theta}_k\right) \neq \theta. \tag{13}$$

Second, the primary-study estimate $\hat{\theta}_k$ is correlated with its estimated variance and this correlation is not taken into account by the IV method (Berkey, Hoaglin, Mosteller, & Colditz, 1995). Third, the sampling distribution of the IV estimator is not well approximated by a normal distribution.

In very rare event settings, when some primary studies report zero event, the IV method can lead to indefinite estimates. For the RR and the OR, this happens whenever $X_{kc} = 0$ or $X_{kt} = 0$ (or both) for any k. For the RD, $\hat{\sigma}^2_k = 0$ and, thus, $\hat{W}_k = \infty$ when $X_{kc} = X_{kt} = 0$. As a result, to obtain defined IV estimates, one has to exclude DZ studies from the computation of all ES and SZ studies from the computation of OR and RR.

2.2 | The Mantel–Haenszel method

The MH method was first proposed in 1959 to estimate the risk of an exposure by means of an OR while adjusting for confounding factors (Mantel & Haenszel, 1959). This method has, then, been extended to other ES (Rothman, Greenland, & Lash, 2008) and

can be applied to conduct meta-analyses of OR, RR, and RD using the following formulae:

$$\widehat{OR}_{MH} = \frac{\sum_{k=1}^{K} X_{kt}\left(n_{kc} - X_{kc}\right)/n_k}{\sum_{k=1}^{K} X_{kc}\left(n_{kt} - X_{kt}\right)/n_k} \tag{14}$$

$$Var\left(\log\left(\widehat{OR}_{MH}\right)\right) = \frac{1}{2}\left(\frac{\sum_{k=1}^{K} R_k P_k}{R^2} + \frac{\sum_{k=1}^{K}\left(P_k S_k + Q_k R_k\right)}{RS} + \frac{\sum_{k=1}^{K} S_k Q_k}{S^2}\right) \tag{15}$$

$$\widehat{RR}_{MH} = \frac{\sum_{k=1}^{K} X_{kt} n_{kc}/n_k}{\sum_{k=1}^{K} X_{kc} n_{kt}/n_k} \tag{16}$$

$$Var\left(\log\left(\widehat{RR}_{MH}\right)\right) = \frac{\sum_{k=1}^{K}\left[\frac{\left(X_{kt} + X_{kc}\right) n_{kt} n_{kc}}{n_k^2} - \frac{X_{kt} X_{kc}}{n_k}\right]}{\left(\sum_{k=1}^{K} \frac{X_{kt} n_{kc}}{n_k}\right)\left(\sum_{k=1}^{K} \frac{X_{kc} n_{kc}}{n_k}\right)} \tag{17}$$

$$\widehat{RD}_{MH} = \frac{\sum_{k=1}^{K}\left(X_{kt} n_{kc} - X_{kc} n_{kt}\right)/n_k}{\sum_{k=1}^{K} n_{kt} n_{kc}/n_k} \tag{18}$$

$$Var\left(\widehat{RD}_{MH}\right) = \frac{\sum_{k=1}^{K}\left(\frac{n_{kt} n_{kc}}{n_k}\right)^2\left[\frac{X_{kt}\left(n_{kt} - X_{kt}\right)}{n_{kt}^2\left(n_{kt} - 1\right)} - \frac{X_{kc}\left(n_{kc} - X_{kc}\right)}{n_{kc}^2\left(n_{kc} - 1\right)}\right]}{\left(\sum_{k=1}^{K} \frac{n_{kt} n_{kc}}{n_k}\right)^2}, \tag{19}$$

where $n_k = n_{kt} + n_{kc}$, $P_k = (X_{kt} + (n_{kc} - X_{kc}))/n_k$, $Q_k = (X_{kc} + (n_{kt} - X_{kt}))/n_k$, $R_k = X_{kt}(n_{kc} - X_{kc})/n_k$, $S_k = X_{kc}(n_{kt} - X_{kt})/n_k$, $R = \sum_{k=1}^{K} R_k$, $S = \sum_{k=1}^{K} S_k$.

The MH formulae do not rely on the primary-study ES estimates (only on the counts). As a result, MH estimators are more robust to zero-event issues than IV's. For RR and OR it provides indefinite estimate only when all control groups report zero event (i.e., $X_{kc} = 0\ \forall\ k$). The variance estimates of the log (OR) and log (RR) are indefinite when either $X_{kt} = 0$ or $X_{kc} = 0\ \forall\ k$ (or both). Although quite rare, such extreme scenarios are sometimes encountered in practice. For instance, a systematic review on the occurrence of lactic acidosis with metformin use in type 2 diabetes mellitus gathered 148 studies that were all DZ studies (Salpeter, Greyber, Pasternak, & Salpeter Posthumous, 2010). From the above formulae, one can see that SZ studies contribute to the computation of all ESs (i.e., to either the numerator of the denominator for the OR and RR and to both for the RD). Contrariwise, DZ studies do not contribute to the computation of the OR and RR estimates, whereas these studies do contribute to the RD estimate (i.e., to the denominator).

Silcocks (2005) showed that the MH estimator for the OR corresponded to the maximum likelihood estimator (based on the Binomial distribution for the number of events), whenever the probability of event in the control group is homogeneous and the ratio of sample sizes is constant across the primary studies. Similarly, one can show that the MH estimators of the RR and RD correspond to their likelihood counterpart, whenever the ratio of sample sizes is constant (whatever the prevalences). For the

RD, the MH estimator is given by Equation (18). With homogeneous ratios of sample sizes (i.e., $n_{kc}/n_{kt} = R \; \forall \; k$), we have $n_{kc} = n_{kt} * R$ and $n_k = n_{kt} * (1 + R)$. Substituting these results into the above formula yields:

$$\begin{aligned}\widehat{RD}_{MH} &= \frac{\sum_{k=1}^{K} \left(X_{kt} n_{kt} R - X_{kc} n_{kt}\right) / (n_{kt} * (1 + R))}{\sum_{k=1}^{K} n_{kt}^2 R / (n_{kt} * (1 + R))} \\ &= \frac{\sum_{k=1}^{K} X_{kt} R / (1 + R) - \sum_{k=1}^{K} X_{kc} / (1 + R)}{\sum_{k=1}^{K} n_{kt} R / (1 + R)} \\ &= \frac{\sum_{k=1}^{K} X_{kt}}{\sum_{k=1}^{K} n_{kt}} - \frac{\sum_{k=1}^{K} X_{kc}}{\sum_{k=1}^{K} n_{kt} R}.\end{aligned} \tag{20}$$

Since $n_{kc} = n_{kt} * R$, the equivalence with the maximum likelihood estimator follows. Using similar arguments, one can show that $\widehat{RR}_{MH} = \widehat{RR}_{ML}$ when $n_{kc}/n_{kt} = R$.

2.3 | The Peto method

The Peto method was introduced as a user-friendly solution to estimate the OR in the setting of rare events (Yusuf et al., 1985). Peto proposed the following estimator for the log(OR):

$$\widehat{\log(OR)}_{Peto} = \frac{\sum_{k=1}^{K} \left(O_k - E_k\right)}{\sum_{k=1}^{K} V_k}, \tag{21}$$

where $O_k = X_{kt}$ is the observed number of events in the treatment group of study k, $E_k = X_k * \frac{n_{kt}}{n_k}$ is the expected number of events in the treatment group under the null hypothesis of no treatment effect, $X_k = X_{kc} + X_{kt}$ is the total number of event, and $V_k = E_k n_{kc}(n_k - X_k)/(n_k(n_k - 1))$ is the hypergeometric variance of O_i under the null. The variance of this estimator is given by:

$$Var\left(\widehat{\log(OR)}_{Peto}\right) = \frac{1}{\sum_{k=1}^{K} V_k}. \tag{22}$$

Peto estimator is obtained using exact likelihood theory (Cox, 1977) and corresponds to the estimate of the common log(OR) obtained in the first step of a Newton–Raphson procedure to maximize the conditional log-likelihood when the starting value for the log(OR) is zero (McCullagh & Nelder, 1981). Hence its other name: the "one-step estimator."

DZ studies do not contribute to the Peto log(OR) estimate (i.e., the quantities O_k, E_k and V_k are all null), whereas SZ studies do contribute. The only setting under which the Peto estimator and its variance are undefined is when all included studies are DZ studies.

2.4 | The continuity correction factor method

As already mentioned in the three previous subsections, the three classical FE methods have difficulties to deal with zero-event studies. The IV method yields indefinite estimates in the presence of either SZ or DZ studies when pooling ORs or RRs, and

TABLE 3 Strategy applied to deal with zero-event studies

	IV method		MH method		Peto method	
	Without CC	**With CC**	**Without CC**	**With CC**	**Without CC**	**With CC**
OR	Discard DZ and SZ studies	Add 0.5 to DZ and SZ studies	No action required[a]	Add 0.5 to DZ studies	No action required[a]	Add 0.5 to DZ studies
RR	Discard DZ and SZ studies	Add 0.5 to DZ and SZ studies	No action required[a]	Add 0.5 to DZ studies	—	
RD	Discard DZ studies	Add 0.5 to DZ studies	No action required[a]	No CC required[b]	—	

[a]MH and Peto methods provide estimates that are well defined even in the presence of SZ or DZ studies.
[b]The MH method provides RD estimates that include the information contained in both SZ and DZ studies.

in the presence of DZ studies when pooling RDs. As for MH and Peto, although these two methods are robust to zero-event studies, they are based on formulae that discard DZ studies when pooling ORs or RRs (for MH).

A simple remedy to the issue of SZ and DZ studies, which dates back to 1934 and has been adopted by many researchers is to use a CC. It consists in adding a constant c to each cell of each contingency table containing one or more 0 frequency (Yates, 1934). Plackett (1964) provided a detailed account of this method. In this paper, we used a CC of 0.5, which can be justified by theoretical arguments (Bhaumik et al., 2012).

In this paper, the motivation of using a CC was to allow all studies to contribute to the combined ES estimate. Therefore, we additionally evaluated IV, MH, and Peto methods with a CC. Table 3 summarizes the various strategies applied with these three methods to tackle the issue of zero-event studies:

2.5 | The median unbiased estimator method

The MUE method works in two steps. First, one computes in each primary study k the MUE of π_{kj}, $j \in \{c,t\}$ (Parzen et al., 2002):

$$\hat{\pi}_{kj} = \begin{cases} \left(1 - 0.5^{1/n_{kj}}\right)/2 & \text{if } X_{kj} = 0 \\ \left(p_{kj}^{L} + p_{kj}^{U}\right)/2 & \text{if } 0 < X_{kj} < n_{kj} \\ \left(0.5^{1/n_{kj}} + 1\right)/2 & \text{if } X_{kj} = n_{kj} \end{cases} \tag{23}$$

with $p_{kj}^{L} = F^{-1}(0.5|\alpha = X_{kj}, \beta = n_{kj} - X_{kj} + 1)$ and $p_{kj}^{U} = F^{-1}(0.5|\alpha = X_{kj} + 1, \beta = n_{kj} - X_{kj})$, where $F^{-1}(Q|\alpha,\beta)$ is the Qth quantile of the beta-distribution with parameters α and β. Note that the expression we used for p_{kj}^{U} differed from that provided in the paper of Parzen and colleagues, which, we believe, contains an error. Mathematical justifications can be found in the Appendix.

Second, from these two estimated probabilities, one can compute the primary-study ESs using Equations (5) and (7) as well as their corresponding variance using Equations (8) and (10), and combined them using a weighted average as in Equation (11). Variance of the MUE combined estimate is then given by Equation (12).

Clearly, the MUE method provides estimators that are always well defined, whatever the degree of sparseness of the events. Moreover, this method includes the information from DZ studies.

2.6 | The Binomial regression model

Another option to deal with zero-event studies is to use the Binomial logistic regression model:

$$X_{kc} \sim Binomial\left(n_{kc}, \pi_{kc}\right) \tag{24}$$

$$X_{kt} \sim Binomial\left(n_{kt}, \pi_{kt}\right) \tag{25}$$

$$logit\left(\pi_{kc}\right) = \alpha_k \tag{26}$$

$$logit\left(\pi_{kt}\right) = \alpha_k + \delta, \tag{27}$$

where α_k represents the logit of the control group probability and δ the log(OR).

Under the assumption of homogeneous baseline prevalences, $\alpha_k = \alpha, \ \forall k$, and parameters of model (24)–(27) are estimated by maximizing the following likelihood function:

$$L(\alpha, \delta) = \prod_{k=1}^{K} \binom{n_{kc}}{X_{kc}} \pi_c^{X_{kc}} \left(1 - \pi_c\right)^{n_{kc} - X_{kc}} \binom{n_{kt}}{X_{kt}} \pi_t^{X_{kt}} \left(1 - \pi_t\right)^{n_{kt} - X_{kt}}. \tag{28}$$

Note that the estimator obtained when maximizing likelihood Equation (28) corresponds to that obtained when completely pooling the data (i.e., without stratifying on k).

Under the assumption of baseline heterogeneity, α_k is treated as a nuisance parameter. One can deal with this nuisance parameter by adopting either a fixed-effects or a random-effects approach. The main advantage of a fixed-effects approach is that no distributional assumption has to be made, whereas the random-effects approach implies the choice of a distribution for α_k. However, particularly with rare events, it is advantageous to treat α_k as a random variable to limit as much as possible the number of parameters to be estimated and allow DZ studies to contribute to the estimation (DZ studies do not contribute to the likelihood with a fixed-effects model since both logits (26) and (27) are undefined).

We assumed $\alpha_k \sim N(\alpha, \sigma_\alpha^2)$ and estimated the parameters by maximizing the following marginal likelihood function:

$$L(\alpha, \delta) = \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \binom{n_{kc}}{X_{kc}} \pi_{kc}^{X_{kc}} \left(1 - \pi_{kc}\right)^{n_{kc} - X_{kc}} \binom{n_{kt}}{X_{kt}} \pi_{kt}^{X_{kt}} \left(1 - \pi_{kt}\right)^{n_{kt} - X_{kt}} f\left(\alpha_k | \alpha, \sigma_\alpha^2\right) d\alpha_k, \tag{29}$$

where $f(\cdot)$ is the Normal density.

3 | ILLUSTRATIVE EXAMPLE

3.1 | Perinatal death in post-term pregnancy

To motivate and illustrate the use of the six methods described in Section 2, we considered the systematic review conducted by Crowley (2000). In this review, the author compared the number of deaths induced by routine and selective induction of pregnancies that go beyond term. Data of the 19 randomized control trials included in Crowley's review are shown in Table 4.

All trials reported at least one arm with zero event and 11 trials were DZ studies. Most of the trials' arms included between 50 and 150 women. Except for one study where an imbalance of 2:1 in favor of the treatment group was observed, studies' arms were mostly balanced.

3.2 | Results from the methods when fitted to the illustrative dataset

Results obtained by the different methods applied to the illustrative dataset are displayed in Table 5. All methods found a decrease of the number of perinatal deaths in the treated women (i.e., those in the group of routine induction). However, the methods differed markedly in terms of the magnitude of the effect and confidence interval obtained, especially for the OR and RR. Since all the studies included in this review reported zero event in at least one arm, the IV method without CC was unable to provide finite estimates for the OR and RR. As expected in such a rare event setting, estimates obtained for these two ES were quite similar. Note that the estimate obtained using the Binomial model with homogeneous baseline prevalences corresponds to the one resulting from simply collapsing the data into a single 2 by 2 contingency table (i.e., (9/3803*4121/4122)/(3794/3803*1/4122) ≅ 0.10). Interestingly, although the Binomial model with heterogeneous baseline prevalences provided a non-zero estimate for the heterogeneity in baseline prevalences (i.e., $\hat{\sigma}_\alpha^2 = 0.37$), it provided virtually the same results as the Binomial model with homogeneous baseline prevalences.

4 | SIMULATION STUDY

4.1 | Model

We considered various scenarios, which differed according to (a) the level of rareness (extremely rare, very rare, moderately rare, common), (b) the assumption regarding baseline prevalences (homogeneous or heterogeneous), (c) the level of

TABLE 4 Illustrative dataset (perinatal death in post-term pregnancy)

	Routine induction		Selective induction	
Trial	n_c	X_c **(Deaths)**	n_t	X_t **(Deaths)**
Henry (1969)	57	2	55	0
Cole (1975)	119	0	118	0
Martin (1978)	134	1	131	0
Tylleskar (1979)	55	0	57	0
Breart (1982)	235	0	481	0
Katz (1983)	78	0	78	1
Suikkari (1983)	53	0	66	0
Sande (1983)	90	0	76	0
Cardozo (1986)	207	1	195	0
Augensen (1987)	195	0	214	0
Dyson (1987)	150	1	152	0
Witter (1987)	97	0	103	0
Bergsjo (1989)	94	1	94	0
Egarter (1989)	168	1	188	0
Martin (1989)	10	0	12	0
Heden (1991)	129	0	109	0
Hannah (1992)	1706	2	1701	0
Herabuyta (1992)	51	0	57	0
NICH (1994)	175	0	235	0
Total	**3803**	**9**	**4122**	**1**

TABLE 5 Results of meta-analysis of the illustrative dataset

	Effect size	95% Wald CI	CI's width
Odds ratio			
IV			
IV+CC	0.56	(0.25; 1.27)	1.02
MH	0.11	(0.01; 0.88)	0.87
MH+CC	0.35	(0.14; 0.88)	0.74
Peto	0.20	(0.06; 0.70)	0.64
Peto+CC	0.44	(0.18; 1.03)	0.85
MUE	0.52	(0.20; 1.38)	1.18
BinReg with homogeneous baseline prevalences*	0.10	(0.01; 0.81)	0.80
BinReg with heterogeneous baseline prevalences**	0.10	(0.01; 0.81)	0.80
Relative risk			
IV			
IV+CC	0.57	(0.25; 1.27)	1.02
MH	0.11	(0.01; 0.88)	0.87
MH+CC	0.35	(0.14; 0.89)	0.75
MUE	0.52	(0.20; 1.38)	1.18
Risk difference (values in %)			
IV	−0.15	(−0.31; 0.00)	0.31
IV+CC	−0.14	(−0.29; 0.00)	0.29
MH	−0.20	(−0.36; -0.05)	0.31
MUE	−0.11	(−0.27; 0.04)	0.31

*Likelihood Equation (28), **Likelihood Equation (29).

unbalance between the trials' arms (strong unbalance in favor of either arms, moderate unbalance in favor of either arms, no unbalance), (d) the size of the treatment effect (large reduction, moderate reduction, no effect, moderate increase, large increase). Extremely rare events had a median baseline probability of $M_{\pi_c} = 0.005$. The three other rareness levels corresponded to $M_{\pi_c} = 0.01,\ 0.05,\ 0.1$, respectively. With homogeneous baseline prevalences we set $\pi_{\mathrm{kc}} = M_{\pi_c}\ \forall k$. With heterogeneous baseline prevalences, we used $\pi_{kc} \sim Logit{-}normal(logit(M_{\pi_c}), \sigma^2_{logit})$, with $\sigma^2_{logit} = 0.5$. As regards unbalance, the five scenarios considered were obtained with mean levels $r = 0.25,\ 0.5,\ 1,\ 2,\ 4$ ($r = 0.25$ corresponding to setting with control groups four times smaller than treatment groups, in average). For the OR, and the RR, the sizes of treatment effect—measured on the log(OR) and log(RR) scales, respectively—were $-1.5,\ -0.5,\ 0,\ 0.5,\ 1.5$. For the RD, we did not consider reduction in event prevalence to avoid cases with negative treatment probability. We used the three following values: $0,\ 0.05,\ 0.1$.

Combinations of these four characteristics resulted in 200 simulation scenarios for the OR and RR, and 120 for the RD. For each scenario, 10,000 meta-analyses were generated, each of them consisting of $K = 20$ primary studies with treatment arms' sample sizes ranging from 50 to 150 (i.e., $n_{kt} \sim Uniform(50; 150)$). Sample sizes for the control arms were then obtained as $n_{kc} = n_{kt} * R$, where $R \sim Uniform(r - 0.1; r + 0.1)$ and r controls the degree of arms unbalance. Treatment probabilities π_{kt} were derived from the control probabilities and the ES considered (e.g., for RD: $\pi_{\mathrm{kt}} = \pi_{\mathrm{kc}} + RD$). Finally, the number of events in both arms were generated by two binomial draws with respective sample sizes and event probabilities.

As pointed out by a reviewer, meta-analyses of $K = 20$ studies might seem an optimistic scenario, as many published meta-analyses include a smaller number of studies. Nevertheless, for our goal to study the impact of rare events, it was better to have a large enough K to avoid fluctuation issues related to scarcity of primary studies. Therefore, we set the number of studies at a somewhat ideal level, but still realistic. For instance, Moher, Tetzlaff, Tricco, Sampson, and Altman (2007) found a median number of 23 studies out of 88 systematic reviews analyzed.

For each of the generated meta-analyses, we estimated the ES of interest, its standard error, and the 95% Wald CI. Performances of the different methods were assessed in terms of bias, and coverage rate. We additionally computed CIs' width and reported them, along with the coverage rate, in the file containing the detailed simulations results (see Supporting Information). We decided to compute median instead of mean values for the bias and CIs' width to avoid the influence of exceedingly large or small values obtained in some simulations. Both OR and RR were analyzed on the log scale.

4.2 | Results

In this section, we presented abridged results focusing on the most interesting findings for each ES in the setting of heterogeneous baseline prevalences, since results and conclusions obtained with homogeneous and heterogeneous baseline prevalences were quite similar. Furthermore, we only discussed results for $r = 1, 2, 4$ and $\log(\mathrm{ES}) = 0, -0.5, -1.5$ (ES = RR or OR), which illustrate well the issues of unbalanced sample sizes and large treatment effects. Detailed results for all the ES, simulated scenarios, and outcomes can be found in the Supporting Information.

4.2.1 | Odds ratio

In terms of bias (Figure 1), the MH and Binomial regression methods provided the best estimates whereas the IV method was the least robust across almost all settings considered. The use of a CC reduced the bias of the IV estimator but increased that of MH and Peto. The MUE method can be seen as an improved version of the IV method with performances comparable to those obtained when using the IV method with CC. The Peto method obtained similar results to those of the MH and Binomial regression methods for moderate and no treatment effects, particularly under balanced settings. However, it was clearly not reliable with large treatment effects and even failed to converge for increasing baseline probabilities (whatever the degree of unbalance). Under the scenario of no treatment effect and balanced trials, all the methods provided unbiased estimates, except the MH + CC method. Applying the Binomial model with a random intercept yielded some numerical issues, especially with very rare events (Table B11, detailed simulation results). The smallest proportion of converged runs achieved by this method was 63.4% for $\log(\mathrm{OR}) = -1.5$, $r = 0.25$ and $M_{\pi_c} = 0.005$ but in most of the scenarios considered, this proportion was above 90%.

In terms of coverage rates (Figure 2), ranking of the different meta-analysis methods was similar to what was obtained in terms of bias. The following additional observations can be made: (a) MH and Binomial regression were slightly conservative in extremely rare event settings, (b) coverage rates obtained by the IV (with or without CC) and MUE methods were below 90% in many scenarios considered, and (c) ranking between these three methods varied depending on the setting.

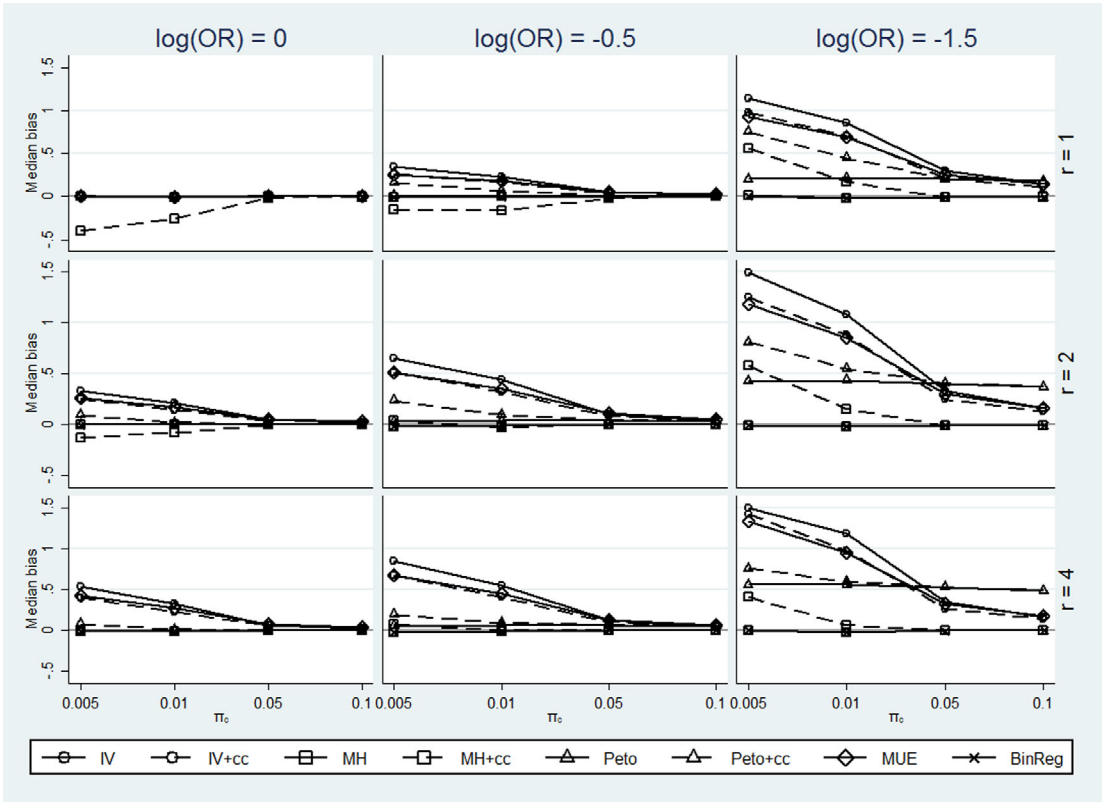

FIGURE 1 Biases obtained by the IV, IV+CC, MH, MH+CC, Peto, Peto+CC, MUE, and Binomial regression methods when estimating the log(OR). Values were right-truncated at 1.5 to allow a better visual inspection

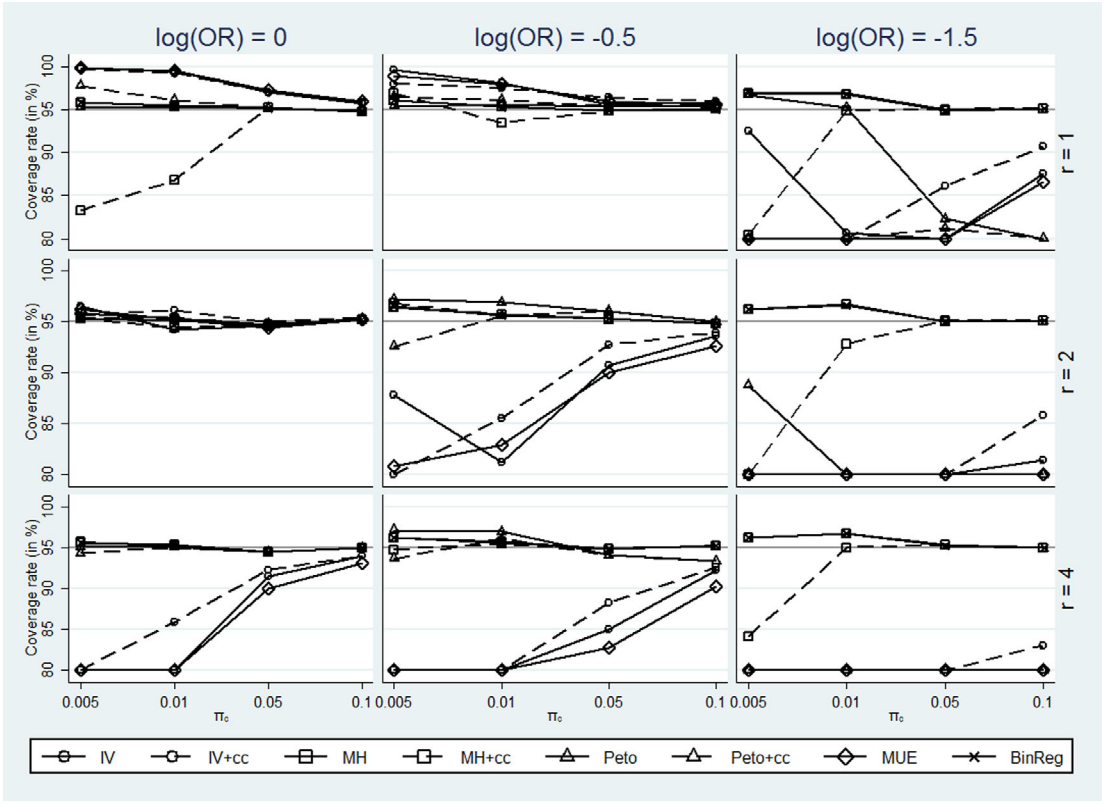

FIGURE 2 Coverage rates (in %) obtained by the IV, IV+CC, MH, MH+CC, Peto, Peto+CC, MUE, and Binomial regression methods for the log(OR). Values were left-truncated at 80 to allow a better visual inspection

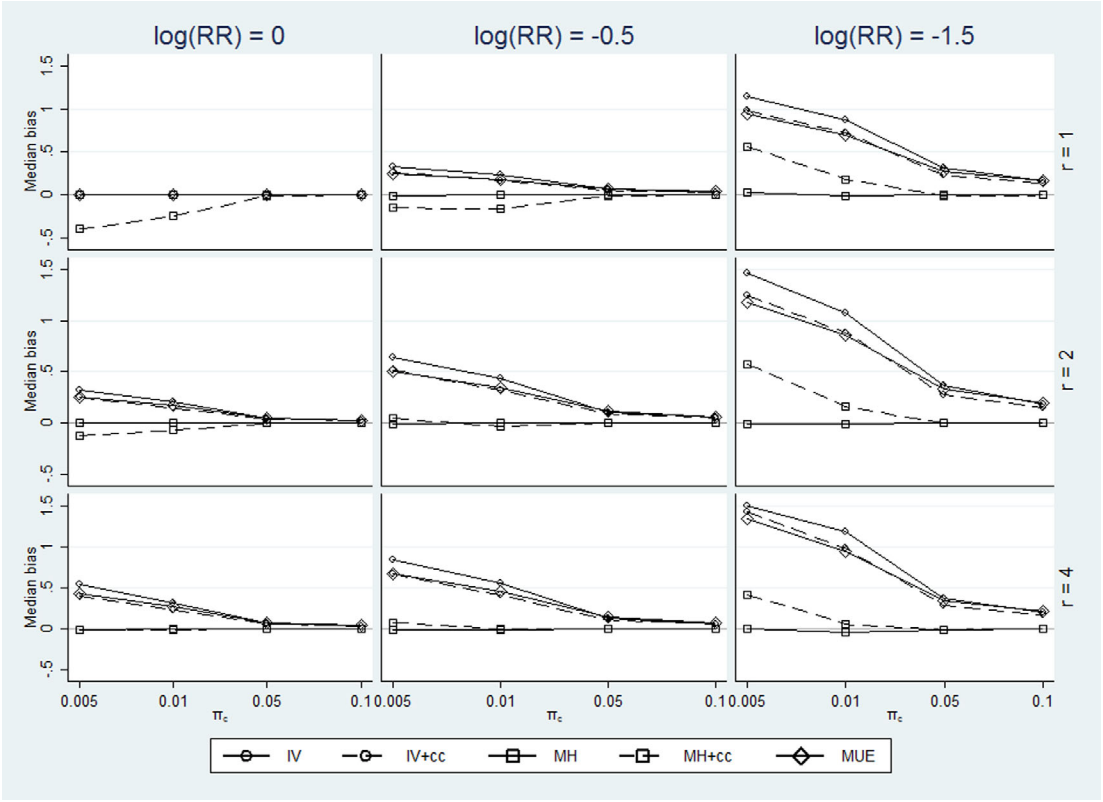

FIGURE 3 Biases obtained by the IV, IV+CC, MH, MH+CC, and MUE methods when estimating the log(RR). Values were right-truncated at 1.5 to allow a better visual inspection

4.2.2 | Relative risk

In terms of bias (Figure 3), comparisons between the different methods yielded to the same conclusions as those drawn for the OR: (a) MH estimator was the least biased, (b) the use of a CC increased the bias for MH estimator and reduced it for the IV estimator, and (c) performances of the MUE method were comparable to those of the IV method with CC.

In terms of coverage rates (Figure 4), results obtained by the six methods investigated to estimate the RR were comparable to those obtained for the OR.

4.2.3 | Risk difference

In terms of bias (Figure 5), the MH estimator of the RD was the least biased across the different settings. Ranking of IV, IV + CC, and MUE depended on the setting and none of the methods strictly dominated—or was strictly dominated by—the others. For all the methods, bias values for the RD were much smaller than those obtained for the RR and OR.

In terms of coverage rates (Figure 6), the MH method again obtained the best results. Contrarily to CIs for the OR and RR, those obtained for the RD tended to have slightly below 95% coverage rates in extremely rare event settings. Coverage rates provided by IV, IV + CC, and MUE methods did not systematically converge toward 95% as the event prevalence increased.

5 | DISCUSSION

Traditionally, in meta-analysis, a clear distinction is made between the FE and the RE models. In the FE model, one assumes that the treatment effect is homogeneous across the primary studies and the goal of the meta-analysis is to estimate a single population parameter. In the RE model, one assumes that the treatment effect is heterogeneous across the primary studies (there is a population of parameters) and focus is on characterizing the distribution of these parameters. The selection of either model must be based on contextual knowledge, as this choice has implications in terms of target of inference and appropriate statistical methods to conduct the analysis.

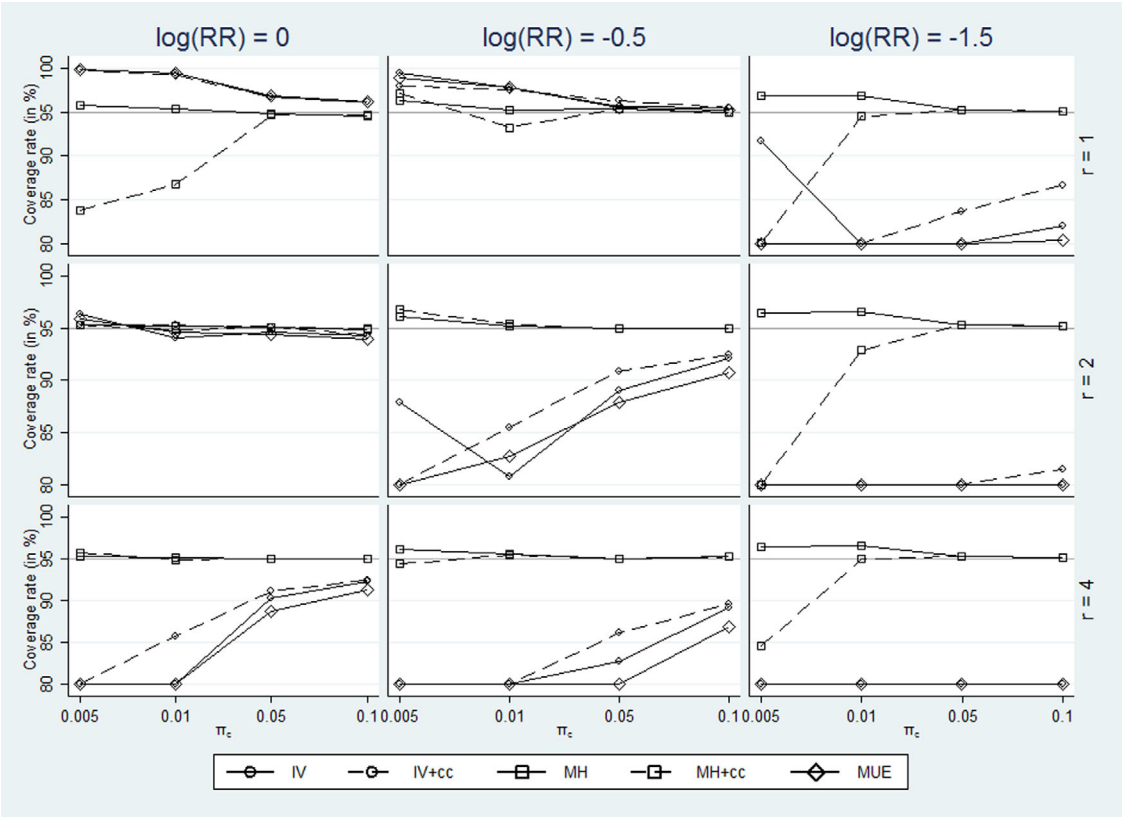

FIGURE 4 Coverage rates (in %) obtained by the IV, IV+CC, MH, MH+CC, and MUE methods for the log(RR). Values were left-truncated at 80 to allow a better visual inspection

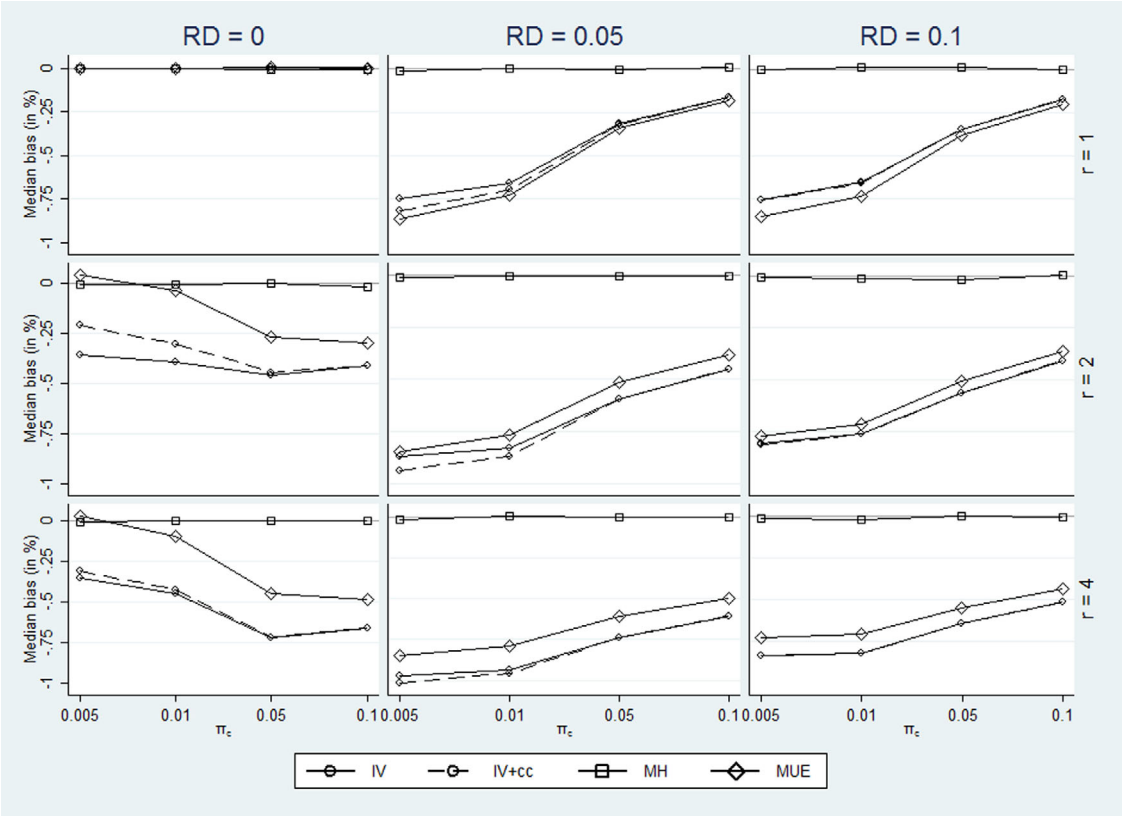

FIGURE 5 Biases (in %) obtained by the IV, IV+CC, MH, and MUE methods when estimating the RD

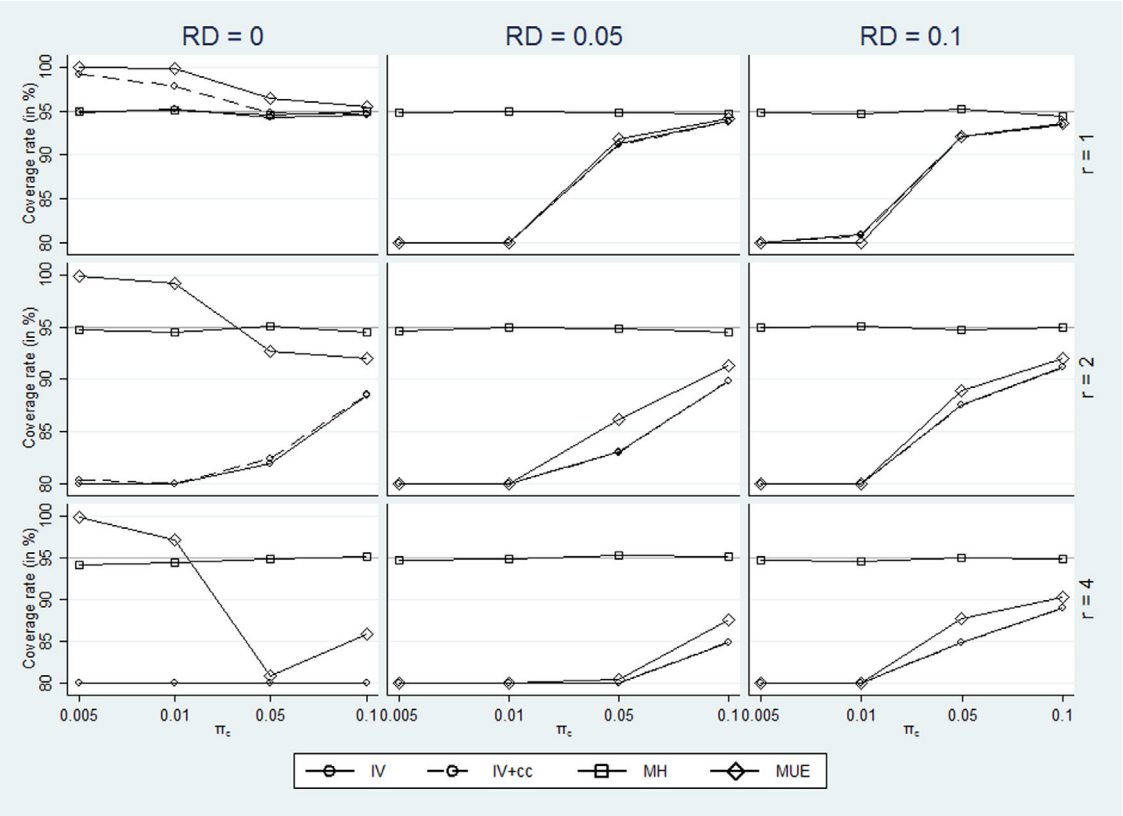

FIGURE 6 Coverage rates (in %) obtained by the IV, IV+CC, MH, and MUE methods for the RD. Values were left-truncated at 80 to allow a better visual inspection

When considering count data, one must further distinguish between settings where baseline prevalences are expected to be homogeneous or heterogeneous. Indeed, we have shown that this distinction has a direct implication regarding the notion of treatment effect homogeneity. Again, the choice between these two settings should be grounded on contextual knowledge. In the literature, the distinction between homogeneous and heterogeneous baseline prevalences is seldomly discussed. Actually, we found only one paper where simulations were carried out under the FE framework with heterogeneous baseline prevalences (Böhning & Sarol, 2000). However, the authors focused on the RD only and did not consider the issue of rare events. In the three most comprehensive simulation studies that tackled the issue of rare event meta-analysis (Bradburn et al., 2007; Kuss, 2015; Sweeting et al., 2004), the authors did not consider heterogeneous baseline prevalences. Therefore, the objective of this paper was to investigate the impact of homogeneous versus heterogeneous baseline prevalences on the performance of FE meta-analysis methods in the context of rare events and homogeneous treatment effect. Through extensive simulations, we assessed the ability of the IV, MH, Peto, with or without CC, MUE, and Binomial regression methods to estimate the three most commonly used effect sizes with count data (RD, RR, and OR) under various settings.

We found that whatever the baseline prevalences (i.e., either homogeneous or heterogeneous), under all the scenarios considered, and for all the ESs, the most reliable methods were the MH method without CC and the Binomial regression model (for estimating the OR only). Interestingly, the fact that the MH method discards DZ studies did not seem to introduce a bias. On the contrary, using a CC to include the information contained in DZ studies deteriorated dramatically this method's performance.

Under the setting of homogeneity in baseline prevalences, MH and Binomial regression estimates correspond to those of the simple pooling method. In other words, when assuming a FE model with homogeneous baseline prevalences, one does not need to apply meta-analysis techniques to compute a combined estimate. This is because, under this particular framework, the OR, RR and RD are collapsible (Greenland, Robins, & Pearl, 1999; Guo & Geng, 2005; Hernan, Clayton, & Keiding, 2011). However, when introducing heterogeneity in baseline prevalences, the OR is no more collapsible and simple pooling leads to biased estimates (this was confirmed by additional simulations; results not shown). As for the RR and RD, they remain collapsible and, thus, one can obtain unbiased estimates by simply pooling the studies, even in the presence of heterogeneous baseline prevalences. Nonetheless, our simulations have shown that simple pooling yielded CIs with coverage rates below nominal. The reason was linked to a failure of the simple pooling method to account for a larger sampling distribution of the ES.

Consistently with the statistical literature, our simulations ranked IV as the worst method in almost all rare events settings. Moreover, we found that this method failed to provide valid CIs and unbiased estimates in scenarios with common events (i.e., baseline prevalences around 10%). When estimating the RR and OR, the use of a CC improved the performance of the IV method in terms of bias. However, this was not the case for the RD and coverage rates obtained for the three ESs were not nominal. As a solution to improve the coverage rates, we additionally considered the modified Hartung–Knapp–Sidik–Jonkman (HKSJ) method (Röver, Knapp, & Friede, 2015), which involves the use of the Student's-*t* distribution and a multiplicative correction term for the variance (additional simulations; results not shown). By increasing the size of the CI, this method improved the coverage rates in scenarios where IV's CIs were below nominal. Most of the time, however, coverage rate of HKSJ's CIs remained below 95%. In addition, when coverage rates were nominal, this method provided wider CIs than those obtained using the MH method. Likewise, HKSJ's CIs were too conservative in scenarios where IV's CIs were valid.

The Peto method, which is often recommended in cases of rare events (Higgins & Green, 2008), yielded contrasted results. Whereas the amount of bias and coverage rates were similar to those obtained by the MH and Binomial regression method in scenarios with no effect or medium treatment effect, results provided by this method under settings with large treatment effect critically deteriorated, especially with large unbalance between the trial's arms. These findings seem to corroborate the conclusions made by some authors stating that the Peto method should only be applied with great caution (Brockhaus, Grouven, & Bender, 2016). Moreover, we found that Peto's estimates did not converge towards the true OR when the baseline probability increased, which gives additional credits to the argument that Peto's OR should be viewed as a different ES, and not as an OR's estimator (Brockhaus, Bender, & Skipka, 2014). Finally, the use of a CC deteriorated the performance of the Peto method.

The MUE method is an interesting—since less subjective—alternative to the use of a CC. It tackles the issue of zero-event studies elegantly and does not involve complex computations (i.e., it only requires the computation of Beta quantiles). While meta-analyses techniques based on MUE of probability have already been discussed by Li and Wang (2019), we innovated here by combining these estimates using a weighted average. The amount of bias found for the OR and RR were similar to that obtained with the IV + CC method. On the other hand, results obtained for the RD were more contrasted and the MUE method sometimes provided the most biased estimates.

Regarding the Binomial regression method, it should be noted that, under the assumption of heterogeneous baseline prevalences, the introduction of random effects yields a qualitative change from so-called "contrast-based" to "contrast-based + baseline" models (Dias & Ades, 2015). Some authors, such as Senn (2010), have argued against these models because they can yield biased results if the heterogeneity of baseline prevalences is not correctly modeled. In this paper, we circumvented this issue by assuming that the random model for α_k was the truth (i.e., in our simulation model, baseline probabilities were generated according to a Logistic-normal model). Using a fixed-effects model is a distribution-free alternative but it is not ideal with rare events since it implies more parameters to be estimated and fails to account for the information contained in DZ studies. In additional simulations, we found that the fixed-effects Binomial model yielded similar values of biases and coverages rates as its random-effects counterpart, but suffered from a loss of precision in settings of very rare events (i.e., it yielded wider confidence intervals). Finally, it should be noted that modifying the link function to obtain other ESs (i.e., log-link for the RR and identity-link for the RD) yielded numerical issues because none of these link functions insure the probabilities to be contained within the 0–1 interval. Marschner and Gillett (2012) proposed to constrain the parameter α_k during the optimization procedure but this leaded to biased estimates. An alternative would be to adapt the model to the ES considered. For instance, Böhning, Mylona, and Kimber (2015) discussed the use of a Poisson modelling to estimate risk ratios.

6 | CONCLUSIONS

Based on our findings, we make the following recommendations to applied researchers conducting meta-analyses of count data, under the framework of a homogeneous treatment effect. First, it is important to clarify the question of heterogeneous versus homogeneous baseline prevalences. When prevalences are expected to be heterogeneous, the researcher has to decide on which scale the treatment effect is assumed homogeneous (i.e., OR, RR, or RD). Second, we highly recommend using the MH method without CC in all circumstances (i.e., whatever the ES of interest, the assumption regarding the heterogeneity in baseline prevalences, and the scenario considered). To estimate the OR, the Binomial regression method is a sound alternative, which allows one to adjust for covariates. Third, the use of a CC should be definitively abandoned.

The main limitation to these recommendations is that they only apply to the framework of a homogeneous treatment effect. The reader must keep in mind that the MH method is valid only when the treatment effect is homogeneous. However, in practice, there are many situations where this assumption is likely to be violated (Kontopantelis, Springate, & Reeves, 2013). Then, although quite extensive, our simulations did not cover all the possible settings. For instance, we did not consider the case of study's

scarcity (i.e., small K), which can impact the methods' performance. Moreover, in our simulations, heterogeneity in baseline prevalences was modeled according a Logit-normal model, which matches the Binomial model described in Equation (29). Results might change depending on the distribution used to generate the baseline prevalences. Another limitation is that we did not investigate Bayesian methods. However, this choice was deliberate and motivated by the fact that these methods require the use of subjective priors, which can have—even non-informative ones—substantial effects on the estimates, especially with rare events (Lambert, Sutton, Burton, Abrams, & Jones, 2005; Senn, 2007). Future researches should focus on extending the present work to the setting of heterogeneous treatment effects, with or without baseline heterogeneity.

ACKNOWLEDGMENTS

The work of R.P.R. was supported by the Swiss National Science Foundation (grant number: 200021_169504/1). The authors would like to thank one associate editor and two anonymous reviewers for their helpful comments and suggestions.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

ORCID

Romain Piaget-Rossel https://orcid.org/0000-0002-5762-4514

REFERENCES

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine*, *14*, 395–411.

Bhaumik, D. K., Amaty, A., Normand, S. L., Greenhouse, J., Kaizar, E., Neelon, B., & Gibbons, R. D. (2012). Meta-analysis of rare binary adverse event data. *Journal of American Statistical Association*, *107*, 555–567.

Böhning, D., Mylona, K., & Kimber, A. (2015). Meta-analysis of clinical trials with rare events. *Biometrical Journal*, *57*, 633–648.

Böhning, D., & Sarol, J. (2000). Estimating risk difference in multicenter studies under baseline-risk heterogeneity. *Biometrics*, *56*, 304–308.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons, Ltd.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., & 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.

Brockhaus, A. C., Grouven, U., & Bender, R. (2016). Performance of the Peto odds ratio compared to the usual odds ratio estimator in the case of rare events. *Biometrical Journal*, *58*, 1428–1444.

Brockhaus, A. C., Bender, R., & Skipka, G. (2014). The Peto odds ratio viewed as a new effect measure. *Statistics in Medicine*, *33*, 4861–4874.

Cox, D. R. (1977). *The analysis of binary data*. New York: Chapman and Hall.

Crowley, P. (2000). Interventions for preventing or improving the outcome of delivery at or beyond term. *Cochrane Database of Systematic Reviews*, *2*, CD000170.

Daly, L. (1992). Simple SAS macros for the calculation of exact binomial and Poisson confidence intervals. *Computer in Biology and Medicine*, *22*, 351–361.

Dias, S., & Ades, A. E. (2015). Absolute or relative effects? Arm-based synthesis of trial data. *Research Synthesis Method*, *7*, 23–28.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27–38.

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*, 29–46.

Guo, J., & Geng, Z. (2005). Collapsibility of logistic regression coefficients. *Journal of the Royal Statistical Society Series B (Methodological)*, *57*, 263–267.

Hernan, M. A., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, *40*, 780–785.

Higgins, J. P. T., & Green, S. (Eds.) (2008). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. The Cochrane Collaboration. Retrieved from www.training.cochrane.org/handbook

Hirji, K. F., Tsiatis, A. A., & Mehta, C. R. (1989). Median unbiased estimation for binary data. *The American Statistician*, *43*, 7–11.

Jackson, D., & White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, *60*, 1040–1058.

Johnson, N. L., & Kotz, S. (1969). *Distribution in statistics: Discrete distributions*. New York: Wiley.

Kontopantelis, E., Springate, D. A., & Reeves, D. (2013). A re-analysis of the Cochrane library data: The dangers of unobserved heterogeneity in meta-analyses. *Plos One*, *8*, e69930.

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.

Kuss, O., Wandrey, M., & Kunze, M. (2009). *How frequent are meta-analyses with "double-zero" studies in systematic reviews?* Retrieved from http://www.egms.de/static/de/meetings/gmds2009/09gmds155.shtml

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*, 2401–2428.

Lane, P. W. (2013). Meta-analysis of incidence of rare events. *Statistical Methods in Medical Research*, *22*, 117–132.

Li, L., & Wang, X. (2019). Meta-analysis of rare binary events in treatment groups with unequal variability. *Statistical Methods in Medical Research*, *28*, 263–274.

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.

Marschner, I. C., & Gillett, A. C. (2012). Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics*, *13*, 179–192.

McCullagh, P., & Nelder, J. A. (1981). *Generalized linear models*. New York: Chapman and Hall.

Moher, D., Tetzlaff, J., Tricco, A. A., Sampson, M., & Altman, D. G. (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine*, *4*, e78.

Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, *9*, 56.

Normand, S. L. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, *18*, 321–359.

Papageorgiou, S. N., Tsiranidou, E., Antonoglou, G. N., Deschner, J., & Jäger, A. (2015). Choice of effect measure for meta-analysis of dichotomous outcomes influenced the identified heterogeneity and direction of small-study effects. *Journal of Clinical Epidemiology*, *68*, 534–541.

Parzen, M., Lipsitz, S., Ibrahim, J., & Klar, N. (2002). An estimate of the odds ratio that always exists. *Journal of Computational and Graphical Statistics*, *11*, 420–436.

Plackett, R. L. (1964). The continuity correction in 2×2 tables. *Biometrika*, *51*, 327–337.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology*. Philadelphia, PA: Lippincott Williams & Wilkins.

Röver, C., Knapp, F., & Friede, T. (2015). Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, *15*, 99.

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*, *28*, 721–738.

Salpeter, S. R., Greyber, E., Pasternak, G. A., & Salpeter Posthumous, E. E. (2010). Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. *Cochrane Database of Systematic Reviews*, *4*, CD002967.

Senn, S. (2007). Trying to be precise about vagueness. *Statistics in Medicine*, *26*, 1417–1430.

Senn, S. (2010). Hans van Houwelingen and the art of summing up. *Biometrical Journal*, *52*, 85–94.

Silcocks, P. (2005). An easy approach to the Robins-Breslow-Greenland variance estimator. *Epidemiologic Perspectives & Innovations*, *2*, 9.

Sinclair, J. C., & Bracken, M. B. (1994). Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology*, *47*, 881–889.

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.

Yates, F. (1934). Contingency tables involving small numbers and the chi2 test. *Supplement of the Journal of the Royal Statistical Society*, *1*, 217–235.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases*, *27*, 335–371.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Piaget-Rossel R, Taffé P. Meta-analysis of rare events under the assumption of a homogeneous treatment effect. *Biometrical Journal*. 2019;61:1557–1574. https://doi.org/10.1002/bimj.201800381

APPENDIX: DERIVATION OF p_{kj}^U FOR THE MEDIAN UNBIASED ESTIMATOR METHOD

Parzen et al. (2002) showed that p_{kj}^U could be obtained by solving

$$0.5 = \sum_{i=0}^{X_{kj}} \left(p_{kj}^U\right)^i \left(1 - p_{kj}^U\right)^{n_{kj}-i} \tag{A.1}$$

Then, one can use the following relationship between the cumulative Beta distribution function and the cumulative Binomial distribution function (Daly, 1992; Johnson & Kotz, 1969):

$$F\left(p|\alpha,\beta\right)=\sum_{i=\alpha}^{n}\binom{n}{i}p^{j}(1-p)^{j} \tag{A.2}$$

where $F(\cdot|\alpha,\beta)$ is the cumulative Beta distribution with integer parameters α and $\beta=n-\alpha+1$. Plugging Equation (A.2) into Equation (A.1) yields

$$0.5=1-F\left(p_{kj}^{U}|\alpha=X_{kj}+1,\beta=n_{kj}-\alpha+1\right) \tag{A.3}$$

where $\beta=n_{kj}-(X_{kj}+1)+1=n_{kj}-X_{kj}$, and not $n_{kj}-X_{kj}+2$ as Parzen and colleagues wrote in their paper on page 425.

Finally, an expression for p_{kj}^{U} is given by

$$p_{kj}^{U}=F^{-1}\left(0.5|\alpha=X_{kj}+1,\beta=n_{kj}-X_{kj}\right) \tag{A.4}$$

which corresponds to the expression we used in this paper.

Journal of Statistics: Advances in Theory and Applications
Volume 21, Number 2, 2019, Pages 91-117
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100122046

A PSEUDO-LIKELIHOOD APPROACH FOR THE META-ANALYSIS OF HOMOGENEOUS TREATMENT EFFECTS: EXPLOITING THE INFORMATION CONTAINED IN SINGLE-ARM AND DOUBLE-ZERO STUDIES

ROMAIN PIAGET-ROSSEL and PATRICK TAFFÉ

Unisanté-Centre Universitaire de Médecine Générale
et Santé Publique
Route de la Corniche 10
1010 Lausanne
Switzerland
e-mail: romain.piaget-rossel@unil.ch
patrick.taffe@chuv.ch

Abstract

Mantel-Haenszel is a fixed-effect meta-analysis method, which performs quite well under the assumption of a homogeneous treatment effect, even in the presence of very rare events. However, this method fails to account for the information contained in single-arm and double zero studies. In this paper, we developed a pseudo-likelihood approach, which allows the inclusion of both single-arm and double-zero studies in the combined effect size estimate. Using Monte-Carlo simulations, we evaluated the behaviour of these two methods when subject to an increasing proportion of single-arm and double-zero studies. We found that the exclusion of double-zero studies did not impact the performance of the Mantel-Haenszel method, whereas the exclusion of single-

2010 Mathematics Subject Classification: Primary 92B15; Secondary 62P10.
Keywords and phrases: double-zero studies, Mantel-Haenszel, meta-analysis, Monte-Carlo simulations, pseudo-likelihood approach, single-arm studies.
Received April 12, 2019; Revised May 3, 2019

arm studies reduced its efficiency compared to the pseudo-likelihood approach. We thus recommend using the pseudo-likelihood approach when the meta-analysis includes single-arm studies. With only double-zero studies, the Mantel-Haenszel can safely be used.

1. Introduction

Meta-analysis is concerned with the synthesis of information contained in independent but related studies, called "primary studies" (Normand [12]). With homogeneous treatment effects (i.e., under the "fixed-effect model"), each primary study seeks to estimate the same population parameter, commonly referred to as "effect size" (ES), and the objective of a meta analysis is to combine all the available evidence into one single and more precise estimate. Under this framework, Piaget-Rossel and Taffé [13] have shown that the Mantel-Haenszel (MH) method without continuity correction (CC) performs very well, even in the presence of very rare events. However, this method excludes double-zero (i.e. studies reporting zero event in both control and treatment arms; DZ) and single-arm (i.e., studies that report results for only one arm; SA) studies from the computation of the combined ES.

DZ studies are typically encountered in rare events settings, where the probability of the event of interest can be so small that it might be unfeasible to design a study with proper sample size, (i.e., so that at least a couple of events are observed). Using a CC allows the inclusion of DZ studies (Sweeting et al. [15]). However, this method is usually not recommended (Efthimiou [4]). Indeed, not only do the resulting estimates depend on the choice of correction used (Kuss et al. [7]; Keus et al. [6]), but this also introduces a bias in the estimates (Piaget-Rossel and Taffé [13]).

Although almost never discussed in the literature, the issue of SA studies is an important one, especially when considering the meta-analysis of observational studies. For example, a systematic review on the surgical management of phyllodes tumors of the breast found that 3 out of 11 studies only included patients with a resection margin above or equal to 10mm, (i.e., these studies did not have any patients with a margin below 10mm) (Toussaint [16]).

Intuitively, SA and DZ studies do carry some information regarding the probability of the event. Consequently, it is useful to devise statistical methods, which allows the inclusion of this information in the computation of the ES estimate. In this paper, to tackle this issue, we have adopted a pseudo-likelihood (PL) approach, which allows the inclusion of SA and DZ studies, without the use of a CC. The basic idea was to adopt a working model for the counts in each arm and treat the (heterogeneous) baseline prevalences as random nuisance parameters. The distribution of the nuisance parameter was not assumed to be known, hence the denomination "pseudo-likelihood".

In this paper, we investigated the performance of this PL approach, which explicitly includes SA and DZ studies, and compared it with the MH method. By means of Monte-Carlo simulations, we evaluated the behaviour of these two methods when subject to an increasing proportion of DZ and SA studies. ES of interest were the odds ratio (OR), the relative risk (RR), and the risk difference (RD), which are the three most commonly-used ES in meta-analyses of binary data. We focused on the specific framework of a homogeneous treatment effect, as the MH method can be seen as the gold standard when there are no SA and DZ studies (Piaget-Rossel and Taffé [13]). Baseline prevalences were assumed to be heterogeneous, as the setting of homogeneous baseline prevalences is more restrictive. The remainder of this paper is structured as follows. In Section 2, we present the PL approach. Section 3 describes our simulation model and presents the results obtained. Section 4 outlines the main findings and makes some recommendations regarding the best method to use when conducting a FE meta-analysis in the presence of SA or DZ studies.

2. The Pseudo-likelihood Approach

Under the assumption of heterogeneous baseline prevalences, the binomial likelihood writes:

$$L = \prod_{k=1}^{K} \prod_{j \in \{t,c\}} \binom{n_{kj}}{x_{kj}} \pi_{kj}(\alpha_k)^{x_{kj}} (1 - \pi_{kj}(\alpha_k))^{n_{kj} - x_{kj}},$$

where n_{kj} is the sample size of arm j ($j = c$ for control and t for treatment) in study k, x_{kj} is the number of events occurring in arm j of study k, $\pi_{kj}(\alpha_k)$ is the inverse of the link function (i.e., the probability of the event) in arm j of study k, and α_k is a nuisance parameter. To deal with the nuisance parameters, one can either treat them as fixed or random quantities. Particularly with rare events, it is advantageous to treat these quantities as random parameters to limit as much as possible the number of parameters to be estimated and allow SA and DZ studies to contribute to the estimation:

$$L = \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \left[\prod_{j \in \{t,c\}} \binom{n_{kj}}{x_{kj}} \pi_{kj}(\alpha_k)^{x_{kj}} (1 - \pi_{kj}(\alpha_k))^{n_{kj} - x_{kj}} \right] f(\alpha_k) d\alpha_k, \qquad (1)$$

where $\pi_{kj}(\alpha_k)$ is the probability of event in study k and treatment arm j defined as a function of α_k, and $f(\alpha_k)$ is the density function of the random variable α_k. Usually, the density function $f(\alpha_k)$ is unknown and to cope with it we have adopted a pseudo-likelihood approach.

For estimating the OR, the following pseudo-likelihood may be used:

$$L^p = \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \left[\prod_{j \in \{t,c\}} \binom{n_{kj}}{x_{kj}} \pi_{kj}(\alpha_k)^{x_{kj}} (1 - \pi_{kj}(\alpha_k))^{n_{kj} - x_{kj}} \right] \phi(\alpha_k | \alpha, \sigma^2) d\alpha_k, \quad (2)$$

with link function $\text{logit}(\pi_{kj}(\alpha_k)) = \alpha_k + \beta T_{kj}$, where T_{kj} is an indicator for treatment arm (i.e., $T_{kj} = 1$ if $j = t$ and 0 if $j = c$), and $\phi(\cdot | \alpha, \sigma^2)$ is the normal density with mean α and variance σ^2. Note that β corresponds to the log(OR). This is a pseudo-likelihood since in Equation (1), $\phi(\cdot | \alpha, \sigma^2)$ is not assumed to be the true density function, it is only a "working" density function.

To estimate the RR, one may adopt the log link function $\log(\pi_{kj}(\alpha_k))$ $= \alpha_k + \beta T_{kj}$. However, to insure probabilities contained within 0-1, one has to constrain the α_k parameter during the optimization procedure (Marschner and Gillett [10]). The main drawback of this approach is that the imposition of the parameter constraint may lead to biased estimates, particularly when the risk level is either low or high. Therefore, to cope with this issue, we proposed to approximate the binomial distribution by a Poisson distribution with parameter $\lambda_{kj} = \pi_{kj} * n_{kj}$ and used the pseudo-likelihood function

$$L^p = \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \left[\prod_{j \in \{t,c\}} \frac{\lambda_{kj}(\alpha_k)^{x_{kj}} \exp(-\lambda_{kj}(\alpha_k))}{x_{kj}!} \right] \phi(\alpha_k | \alpha, \sigma^2) d\alpha_k. \quad (3)$$

The numerical advantages of this approach are obvious as the log link is canonical for the Poisson likelihood. In this model, β corresponds to the log(RR).

Similarly, for the RD, we used the canonical identity link function and approximated the binomial distribution of the counts by a normal distribution for the proportions:

$$L^p = \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \left[\prod_{j \in \{C,T\}} \frac{1}{\sqrt{2\pi * \tau_{kj}^2}} \exp\left(\frac{p_{kj} - \pi_{kj}(\alpha_k)}{2\tau_{kj}^2} \right) \right] \phi(\alpha_k | \alpha, \sigma^2) d\alpha_k, \quad (4)$$

where $p_{kj} = x_{kj} / n_{kj}$, $\pi_{kj} \equiv E(p_{kj}) = \alpha_k + \beta T_{kj}$, and $\tau_{kj}^2 \equiv \mathrm{Var}(p_{kj}) = \pi_{kj} * (1 - \pi_{kj}) / n_{kj}$. Given that the variances τ_{kj}^2 are unknown and difficult to estimate (the above model is highly nonlinear and computation of the integral with good precision is challenging, particularly with rare events), we decided to use instead the robust sandwich estimate of the variance-covariance matrix (White [17]). Moreover, to achieve appropriate empirical coverage rates, we selected the 98.5th quantile of the standard normal distribution to compute the

Wald CI for the RD. This yielded a "calibrated" CI in the spirit of that obtained by using the modified Hartung-Knapp-Sidik-Jonkman method (Röver et al. [14]).

3. Simulation Study

3.1. Model

Given our objective (study the impact of SA and DZ studies), we decided to consider large meta-analyses to avoid fluctuations issues related to scarcity of primary studies. Therefore, we set the number of primary studies at $K = 20$. Although this scenario might seem optimistic, as many published meta-analyses, notably in the Cochrane Library, included less primary studies, it is nevertheless a realistic one (see, for instance, Moher et al. [11], which found a median number of 23 studies out of 88 systematic reviews analyzed).

In each primary-study, treatment arms' sample sizes ranged from 50 to 150 (i.e., $n_{kt} \sim$ discrete-uniform {50; 150}). Control arms' sample sizes were generated as $n_{kc} = n_{kt} + r$, with $r \sim$ discrete-uniform {– 15; 15}. Baseline prevalences π_{kc} were obtained as random draws from a continuous uniform distribution with range $[p - p/5; p + p/5]$, p being the mean of the distribution. This distribution provided a realistic level of heterogeneity in baseline prevalences under the assumption of a homogeneous treatment effect. Probabilities in the treated group π_{kt} were derived from the control probabilities and ES considered (i.e., $\pi_{kt} = \pi_{kc} * OR/(\pi_{kc} * OR + 1 - \pi_{kc})$; $\pi_{kt} = \pi_{kc} * RR$; $\pi_{kt} = \pi_{kc} + RD$). Finally, the number of events in both arms were generated by two binomial draws with respective sample sizes and event probabilities. Notice that our simulations models are different from the models described in (2), (3), or (4), hence the terminology "pseudo-likelihood". We investigated the impact of both DZ and SA studies on the performance of the proposed approaches. To study the impact of DZ studies, we considered four

scenarios with different mean values for the baseline prevalences $p = \{0.1, 0.007, 0.0035, 0.0015\}$. Under the null hypothesis of no treatment effect, these four probabilities yielded approximately 0, 25, 50, and 75% of DZ studies per meta-analysis, respectively. When studying the impact of SA studies, we set $p = 0.1$, which ensured almost all meta-analyses to be free of DZ studies. Each primary study was defined as SA or not-SA study by means of a Bernoulli draw with probability m. In the subset of SA studies, the arm to be removed was then designated by a Bernoulli draw with probability 0.5. We considered four different scenarios with respective m values $\{0, 0.25, 0.5, 0.75\}$.

The impact of SA and DZ studies was assessed for various values of the ES, which are reported in Table 1. Since the RD is an absolute measure, we had to derive it from the log(RR) and mean baseline prevalence p, to avoid generating probabilities in the treatment group below 0. For instance, for $p = 0.15\%$ and log(RR) = – 1.5, one obtained RD as $\exp(-1.5)^{*}0.0015 - 0.0015 \cong -0.0012$.

For each scenario considered, 10,000 meta-analyses were generated. For each of the generated meta-analyses, we estimated the ES of interest, its standard error and the 95% Wald confidence interval (CI). For the RD, we computed a calibrated CI using the 98.5^{th} quantile; for the OR and RR, the usual 97.5^{th} quantile was used. Performance of the MH and PL methods were assessed in terms of bias, coverage rate and width of the CI. We decided to compute median instead of mean values for the bias and CI's width to avoid the influence of exceedingly large or small values. We also reported the proportion of converged runs. For the MH method, a run was reported as non-converged when either of the estimates obtained (i.e., the ES or its variance) where undefined. For the OR and the RR, this happened when either $x_{kt} = 0$ or $x_{kc} = 0, \forall k$ (or both); for the RD this happened only when both $x_{kt} = 0$ and $x_{kc} = 0, \forall k$. Non-converged runs were also reported by the MH method when the K primary studies

were SA studies. Both OR and RR estimates were analyzed on the log scale because the sampling distribution was more symmetrical on this scale.

Table 1. Size of the treatment effect considered in the simulations

p	log(OR)*	log(RR)**	RD (in %)	Description of the effect
0.0015	− 1.5 (0.22)	− 1.5 (0.22)	− 0.12	Large negative ES
0.0035	− 1.5 (0.22)	− 1.5 (0.22)	− 0.27	Large negative ES
0.007	− 1.5 (0.22)	− 1.5 (0.22)	− 0.54	Large negative ES
0.1	− 1.5 (0.22)	− 1.5 (0.22)	− 7.77	Large negative ES
0.0015	− 0.5 (0.61)	− 0.5 (0.61)	− 0.06	Moderate negative ES
0.0035	− 0.5 (0.61)	− 0.5 (0.61)	− 0.14	Moderate negative ES
0.007	− 0.5 (0.61)	− 0.5 (0.61)	− 0.28	Moderate negative ES
0.1	− 0.5 (0.61)	− 0.5 (0.61)	− 3.93	Moderate negative ES
0.0015	0 (1)	0 (1)	0	Null ES
0.0035	0 (1)	0 (1)	0	Null ES
0.007	0 (1)	0 (1)	0	Null ES
0.1	0 (1)	0 (1)	0	Null ES
0.0015	0.5 (1.65)	0.5 (1.65)	0.10	Moderate positive ES
0.0035	0.5 (1.65)	0.5 (1.65)	0.23	Moderate positive ES
0.007	0.5 (1.65)	0.5 (1.65)	0.45	Moderate positive ES
0.1	0.5 (1.65)	0.5 (1.65)	6.49	Moderate positive ES
0.0015	1.5 (4.48)	1.5 (4.48)	0.52	Large positive ES
0.0035	1.5 (4.48)	1.5 (4.48)	1.22	Large positive ES
0.007	1.5 (4.48)	1.5 (4.48)	2.44	Large positive ES
0.1	1.5 (4.48)	1.5 (4.48)	34.82	Large positive ES

Note. p = mean baseline prevalence;
*ORs reported between brackets;
**RRs reported between brackets.

3.2. Results

3.2.1. Double-zero studies

Reducing the mean baseline prevalence increased the proportion of DZ studies and, thus, the proportion of excluded studies by the MH method when computing the OR (Table 2). On the other hand, with the PL approach DZ studies are allowed and no study is excluded. We found that even when the proportion of discarded studies was very high (75%), the MH method still performed very well, and provided unbiased estimates and CIs with proper coverage rates. Likewise, the PL approach performed very well and provided results quite similar to those produced by the MH method. Except in the large negative ES scenario with extremely small baseline prevalences (i.e., $p = 0.0015$), where both methods reported large biases (median value > 0.75 for a log(OR) of – 1.5), estimates obtained by the two methods were good across all scenarios investigated. Both methods provided conservative CIs in case of extremely rare events. Finally, the proportion of converged runs indicated that PL was more computationally involved and could run into numerical issues, especially with very rare events.

Regarding the RR, results obtained by the MH and PL methods were comparable to those obtained for the OR (Table 3). Both methods provided unbiased estimates and valid CIs in almost all settings investigated. Again, the fact that MH discarded DZ studies did not alter the performance of this method.

For the RD, both methods included DZ studies and, thus, no study was excluded (Table 4). Similar to the OR and RR cases, biases obtained were small and coverage rates nominal. However, CIs provided by the PL approach were wider than those obtained with the MH method.

Table 2. Impact of double-zero studies for the estimation of the OR

	Excluded studies (in %)		Median bias		Coverage rate (in %)		Median CI's width		Converged runs (in %)	
p **(in %)**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Large negative ES										
10	0.02	0	0.001	0.001	95.20	95.18	0.64	0.64	100.00	98.97
0.70	43.69	0	− 0.015	− 0.022	97.09	97.26	2.50	2.50	95.69	80.77
0.35	64.26	0	0.145	0.125	96.33	96.40	3.28	3.28	79.19	73.67
0.15	79.83	0	0.747	0.782	96.81	96.78	4.45	4.40	46.35	44.97
Moderate negative ES										
10	0.00	0	− 0.001	− 0.000	95.24	95.20	0.47	0.47	100.00	99.21
0.70	34.42	0	− 0.005	− 0.016	95.91	95.92	1.75	1.76	99.96	80.68
0.35	57.62	0	− 0.018	− 0.024	97.59	97.74	2.53	2.54	98.50	87.21
0.15	76.71	0	0.090	0.090	99.50	99.63	3.92	3.92	79.27	76.05
Null ES										
10	0.00	0	0.002	0.001	95.04	95.04	0.41	0.41	100.00	99.29
0.70	27.05	0	− 0.005	− 0.005	96.01	96.20	1.52	1.52	100.00	80.80
0.35	50.78	0	0.001	− 0.000	96.91	96.87	2.19	2.19	99.76	84.93
0.15	73.02	0	0.001	0.000	99.71	99.68	3.40	3.40	89.73	85.30

Table 2. (Continued)

	Excluded studies (in %)		Median bias		Coverage rate (in %)		Median CI's width		Converged runs (in %)	
p **(in %)**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Moderate positive ES										
10	0.00	0	− 0.000	− 0.000	95.39	95.41	0.38	0.38	100.00	99.29
0.70	18.35	0	0.008	0.008	95.75	95.59	1.36	1.36	100.00	82.56
0.35	41.17	0	0.016	0.019	96.17	96.05	1.94	1.96	99.89	81.78
0.15	67.06	0	− 0.015	− 0.001	98.46	98.77	3.05	3.10	94.00	87.52
Large positive ES										
10	0.00	0	0.000	0.000	95.27	95.13	0.35	0.35	100.00	99.12
0.70	4.04	0	0.016	0.015	95.79	95.87	1.18	1.18	100.00	94.99
0.35	17.32	0	0.021	0.022	96.42	96.34	1.67	1.66	99.86	82.46
0.15	45.04	0	0.010	0.020	97.28	97.44	2.53	2.53	94.76	81.18

Note: p = mean baseline prevalence, ES = effect size, MH = Mantel-Haenszel method, PL = pseudo-likelihood method, CI = confidence interval. See Table 1 for the ES' values in the different scenarios.

Table 3. Impact of double-zero studies for the estimation of the RR

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width		Converged runs (in %)	
p (in %)	MH	PL	MH	PL	MH	PL	MH	PL	MH	PL
Large negative ES										
10	0.02	0	− 0.003	− 0.003	95.17	95.60	0.64	0.65	100.00	99.27
0.70	43.42	0	− 0.015	− 0.014	96.93	96.95	2.49	2.49	95.80	94.83
0.35	64.28	0	0.149	0.128	96.55	96.52	3.28	3.28	78.73	78.68
0.15	79.77	0	0.756	0.789	97.32	97.36	4.44	4.38	46.31	46.62
Moderate negative ES										
10	0.00	0	− 0.001	− 0.001	94.69	95.52	0.44	0.45	100.00	99.15
0.70	34.32	0	− 0.006	− 0.007	95.97	96.02	1.74	1.74	99.98	99.02
0.35	57.81	0	− 0.013	− 0.014	97.96	97.92	2.54	2.53	98.34	97.68
0.15	76.64	0	0.104	0.099	99.53	99.53	3.88	3.92	78.51	78.52
Null ES										
10	0.00	0	− 0.001	− 0.000	95.50	96.35	0.37	0.39	100.00	99.03
0.70	27.01	0	0.001	0.001	95.75	95.72	1.51	1.52	100.00	99.20
0.35	50.96	0	0.000	− 0.001	96.90	96.93	2.19	2.18	99.75	98.81
0.15	73.15	0	0.001	0.000	99.68	99.69	3.39	3.39	90.16	89.89

Table 3. (Continued)

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width		Converged runs (in %)	
p **(in %)**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Moderate positive ES										
10	0.00	0	0.002	0.002	94.93	96.37	0.33	0.35	100.00	98.65
0.70	18.27	0	0.012	0.011	95.17	95.20	1.34	1.35	100.00	99.62
0.35	41.30	0	0.002	0.001	96.63	96.57	1.93	1.93	99.87	99.06
0.15	67.08	0	− 0.014	− 0.004	98.44	98.54	3.06	3.06	94.12	93.77
Large positive ES										
10	0.00	0	0.000	− 0.000	95.15	96.92	0.28	0.31	100.00	98.64
0.70	4.02	0	0.008	0.008	95.49	95.63	1.16	1.17	100.00	99.57
0.35	17.37	0	0.014	0.016	96.18	96.15	1.65	1.65	99.92	99.67
0.15	44.98	0	− 0.002	− 0.004	97.08	97.04	2.51	2.51	94.49	94.06

Note: p = mean baseline prevalence, ES = effect size, MH = Mantel-Haenszel method, PL = pseudo-likelihood method, CI = confidence interval. See Table 1 for the ES' values in the different scenarios.

Table 4. Impact of double-zero studies for the estimation of RD

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width (in %)		Converged runs (in %)	
p (in %)	MH	PL	MH	PL	MH	PL	MH	PL	MH	PL
Large negative ES										
10	0	0	− 0.010	− 0.004	94.66	95.44	2.917	3.344	100.00	99.99
0.70	0	0	0.005	0.007	94.21	95.05	0.802	0.900	100.00	100.00
0.35	0	0	0.005	0.006	94.16	94.14	0.567	0.632	99.97	99.95
0.15	0	0	0.011	0.006	95.75	94.52	0.369	0.411	97.28	97.22
Moderate negative ES										
10	0	0	− 0.013	− 0.007	94.66	95.48	3.344	3.819	100.00	99.98
0.70	0	0	0.001	− 0.002	94.70	95.52	0.919	1.037	100.00	100.00
0.35	0	0	− 0.003	0.002	94.73	95.63	0.648	0.726	100.00	99.99
0.15	0	0	0.007	0.004	93.79	95.13	0.423	0.473	98.90	98.86
Null ES										
10	0	0	− 0.021	− 0.029	94.73	95.35	3.700	4.225	100.00	99.99
0.70	0	0	− 0.002	− 0.003	94.90	95.67	1.026	1.160	100.00	99.99
0.35	0	0	− 0.001	0.001	94.92	96.02	0.722	0.815	100.00	100.00
0.15	0	0	− 0.000	0.000	94.84	97.59	0.469	0.528	99.79	99.77

Table 4. (Continued)

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width (in %)		Converged runs (in %)	
***p* (in %)**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Moderate positive ES										
10	0	0	− 0.004	− 0.006	94.80	95.83	4.162	4.767	100.00	99.96
0.70	0	0	− 0.012	− 0.011	94.90	95.90	1.178	1.334	100.00	99.98
0.35	0	0	− 0.001	0.001	95.10	96.01	0.833	0.939	100.00	100.00
0.15	0	0	0.001	− 0.001	94.92	95.81	0.542	0.608	99.99	99.99
Large positive ES										
10	0	0	0.017	0.006	94.41	95.67	5.071	5.776	100.00	99.96
0.70	0	0	− 0.000	− 0.002	94.93	95.98	1.683	1.914	100.00	99.95
0.35	0	0	− 0.002	− 0.005	94.89	95.55	1.199	1.356	100.00	99.98
0.15	0	0	− 0.009	− 0.008	94.60	94.92	0.784	0.885	100.00	100.00

Note: *p* = mean baseline prevalence, ES = effect size, MH = Mantel-Haenszel method, PL = pseudo-likelihood method, CI = confidence interval. See Table 1 for the ES' values in the different scenarios.

3.2.2. Single-arm studies

Whatever the ES considered, the MH method excluded SA studies, whereas this was not the case with the PL approach.

Regarding the OR, both methods provided unbiased estimates and nominal coverage rates in all scenarios considered (Table 5). However, as the proportion of SA studies increased, the MH method became less and less efficient and CIs provided by this method were wider than those obtained using the PL approach, especially for large negative ES. Finally, SA studies created less numerical issues than DZ studies (the proportion of converged runs was always above 98%).

Results obtained for the RR were quite comparable to those obtained when estimating the OR (Table 6). In terms of bias, both methods provided virtually identical results and all values observed were below 2% (in absolute terms). Coverage rates were always nominal. Again, the MH method was less and less efficient as the proportion of SA studies increased and provided CIs much wider than the PL method.

Table 7 reports the results obtained by the MH and PL methods when estimating the RD. Median values of the bias were all below 0.05% in absolute terms. Regarding coverage rates, both methods provided valid CIs across all the scenarios, except when the proportion of SA studies was 75%, in which case the coverage rate of the PL's CI was below nominal. Regarding precision of the estimates, similar to what was observed for the OR and RR, the PL approach outperformed the MH method for increasing proportions of SA studies.

Table 5. Impact of single-arm studies for the estimation of OR

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width		Converged runs (in %)	
SA (in %)	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Large negative ES										
0	0.03	0	− 0.004	− 0.004	95.14	95.15	0.64	0.64	100.00	98.79
25	24.98	0	− 0.007	− 0.006	94.99	95.06	0.74	0.69	100.00	98.79
50	49.98	0	− 0.001	− 0.002	95.30	95.09	0.92	0.75	100.00	98.63
75	74.94	0	− 0.015	− 0.007	95.85	94.97	1.32	0.84	99.26	98.39
Moderate negative ES										
0	0.00	0	0.002	0.002	95.03	95.05	0.47	0.46	100.00	99.22
25	25.03	0	− 0.002	− 0.003	94.98	95.00	0.54	0.50	100.00	99.26
50	49.97	0	− 0.003	0.000	95.06	95.13	0.66	0.55	100.00	99.19
75	75.06	0	− 0.007	− 0.002	95.25	95.00	0.95	0.62	99.64	98.81
Null ES										
0	0.00	0	0.002	0.002	95.27	95.19	0.41	0.41	100.00	99.26
25	25.01	0	0.001	− 0.001	94.96	95.06	0.48	0.45	100.00	99.32
50	50.09	0	0.000	0.001	94.93	94.63	0.59	0.49	100.00	99.29
75	74.82	0	0.002	− 0.001	95.39	94.95	0.84	0.55	99.76	99.02

Table 5. (Continued)

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width		Converged runs (in %)	
SA (in %)	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Moderate positive ES										
0	0.00	0	0.001	0.001	94.85	94.94	0.38	0.38	100.00	99.29
25	24.76	0	0.003	0.001	94.94	94.57	0.44	0.41	100.00	99.24
50	50.11	0	0.001	0.002	94.96	95.03	0.54	0.45	100.00	99.42
75	74.77	0	− 0.000	− 0.000	95.41	95.41	0.77	0.51	99.65	99.15
Large positive ES										
0	0.00	0	0.003	0.003	95.01	94.99	0.35	0.35	100.00	99.29
25	24.89	0	− 0.001	− 0.001	95.18	94.95	0.40	0.38	100.00	99.32
50	49.86	0	− 0.001	− 0.001	95.37	94.89	0.49	0.41	100.00	99.31
75	74.94	0	0.004	0.002	94.55	94.47	0.70	0.47	99.76	99.25

Note: SA = proportion of single-arm studies, ES = effect size, MH = Mantel-Haenszel method, PL = pseudo-likelihood method, CI = confidence interval. See Table 1 for the ES' values in the different scenarios.

Table 6. Impact of single-arm studies for the estimation of RR

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width		Converged runs (in %)	
SA (in %)	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Large negative ES										
0	0.03	0	− 0.001	− 0.001	95.01	95.36	0.64	0.65	100.00	99.21
25	25.02	0	− 0.005	− 0.004	95.63	95.66	0.74	0.70	100.00	99.42
50	50.07	0	− 0.003	− 0.003	94.73	95.32	0.91	0.76	100.00	99.52
75	74.87	0	− 0.015	− 0.007	95.61	95.72	1.32	0.84	99.32	99.53
Moderate negative ES										
0	0.00	0	− 0.000	0.000	94.62	95.53	0.43	0.45	100.00	99.12
25	25.14	0	− 0.001	− 0.001	95.32	95.97	0.50	0.49	100.00	99.55
50	50.13	0	− 0.001	0.000	95.12	95.91	0.62	0.53	100.00	99.54
75	74.79	0	0.001	0.001	95.37	95.96	0.88	0.59	99.69	99.56
Null ES										
0	0.00	0	− 0.001	− 0.001	94.98	96.06	0.37	0.39	100.00	99.11
25	25.03	0	− 0.002	− 0.002	95.35	96.21	0.43	0.42	100.00	99.45
50	50.01	0	0.002	0.001	95.29	95.85	0.53	0.46	100.00	99.41
75	74.95	0	− 0.005	− 0.001	95.19	95.43	0.76	0.52	99.61	99.62

Table 6. (Continued)

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width		Converged runs (in %)	
SA (in %)	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Moderate positive ES										
0	0.00	0	0.000	0.001	95.50	96.89	0.33	0.35	100.00	99.07
25	24.95	0	0.001	0.001	94.66	96.01	0.38	0.38	100.00	99.18
50	49.94	0	0.000	0.001	94.96	96.27	0.47	0.42	100.00	99.35
75	74.97	0	0.001	− 0.001	95.29	95.83	0.67	0.46	99.64	99.43
Large positive ES										
0	0.00	0	0.001	0.001	94.66	96.48	0.28	0.31	100.00	98.53
25	25.03	0	− 0.001	− 0.000	95.25	96.92	0.32	0.33	100.00	98.99
50	49.99	0	0.002	0.001	95.31	96.52	0.40	0.36	100.00	98.71
75	74.84	0	0.004	0.003	94.91	95.75	0.57	0.40	99.64	98.80

Note: SA = proportion of single-arm studies, ES = effect size, MH = antel-Haenszel method, PL = pseudo-likelihood method, CI = confidence interval. See Table 1 for the ES' values in the different scenarios.

Table 7. Impact of single-arm studies for the estimation of RD

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width (in %)		Converged runs (in %)	
SA (in %)	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Large negative ES										
0	0.00	0	0.004	0.009	94.83	95.66	0.029	0.033	100.00	99.98
25	25.03	0	0.017	0.007	94.69	94.86	0.034	0.036	100.00	99.96
50	50.08	0	0.006	0.013	95.11	94.05	0.041	0.039	100.00	99.89
75	74.86	0	0.008	0.035	94.88	90.92	0.059	0.043	99.66	98.89
Moderate negative ES										
0	0.00	0	0.001	− 0.011	95.34	95.66	0.033	0.038	100.00	99.99
25	25.11	0	− 0.001	− 0.011	94.81	94.90	0.039	0.041	100.00	99.96
50	50.11	0	− 0.033	− 0.014	94.92	94.07	0.047	0.044	100.00	99.95
75	74.87	0	0.040	0.029	94.71	91.56	0.067	0.049	99.65	98.94
Null ES										
0	0.00	0	− 0.012	− 0.006	95.01	95.53	0.037	0.042	100.00	99.97
25	24.93	0	− 0.003	− 0.017	95.22	95.60	0.043	0.045	100.00	99.96
50	49.90	0	− 0.006	0.009	94.84	94.95	0.052	0.049	100.00	99.95
75	74.93	0	− 0.005	0.001	94.62	90.65	0.075	0.053	99.65	99.07

Table 7. (Continued)

	Excluded studies (in %)		Median Bias		Coverage rate (in %)		Median CI's width (in %)		Converged runs (in %)	
SA (in %)	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**	**MH**	**PL**
Moderate positive ES										
0	0.00	0	− 0.004	− 0.023	95.01	95.70	0.042	0.048	100.00	100.00
25	24.90	0	− 0.011	− 0.031	94.63	95.38	0.048	0.051	100.00	99.98
50	50.21	0	− 0.020	− 0.007	94.83	94.47	0.059	0.055	100.00	99.88
75	74.81	0	0.038	0.011	95.13	91.17	0.084	0.060	99.76	98.77
Large positive ES										
0	0.00	0	0.006	0.016	94.87	95.64	0.051	0.058	100.00	99.95
25	24.99	0	0.017	0.026	94.84	95.63	0.058	0.062	100.00	99.87
50	50.11	0	− 0.008	− 0.021	94.62	94.23	0.072	0.066	100.00	99.91
75	74.89	0	0.007	0.006	94.60	91.35	0.102	0.072	99.62	98.92

Note: SA = proportion of single-arm studies, ES = effect size, MH = antel-Haenszel method, PL = pseudo-likelihood method, CI = confidence interval. See Table 1 for the ES' values in the different scenarios.

4. Discussion

The MH method has been shown to perform extremely well under the assumption of a homogeneous treatment effect (Bradburn et al. [2]; Piaget-Rossel and Taffé [13]). It involves simple computations, can be applied to compute the three classical ESs, (i.e., OR, RR, and RD) and is readily available in most of the statistical packages. However, DZ and SA studies are excluded from the computation of the combined ES estimate. In this paper, we have developed a novel approach based on the formulation of a PL, which allows one to include both SA and DZ studies into the meta-analysis. Using simulations, we compared the performance of this PL approach to that of the MH method, in settings with increasing proportion of SA and DZ studies.

Our proposed PL method performed very well for all three ESs. For the RD, we found that the calibrated Wald's CI computed using the 98.5^{th} quantile of the standard normal distribution provided nominal coverage rates, except in the settings with 75% of SA studies. This shows that using the normal distribution as a working distribution for the baseline prevalences does not impact the performance of this method. In additional simulations, we found that using an asymmetrical beta distribution to generate the baseline prevalences – instead of the uniform distribution described in Subsection 3.1 – did not alter this conclusion (results not shown). This finding challenges Dias and Ades's statement [3] that "unless the baseline model is correctly specified, the relative effect estimates will be biased".

We found that both the MH and PL methods provided reliable results, whatever the proportion of DZ studies. Biases, coverage rates, and CIs' width provided by these two methods were quite similar. The only noticeable difference was that the PL's calibrated CIs for the RD were wider than the CIs obtained with the MH method. These results suggest that under the assumption of a homogeneous treatment effect, DZ studies do not contain relevant information for the meta-analysis.

This is quite unexpected given that, for instance, a DZ study of size 150 does seem to convey more information regarding the low probability of an event compared to a DZ study of size 50. Moreover, Friedrich et al. [5] argued that deleting DZ studies in balanced trials might bias the treatment effect away from the null. Nevertheless, from our results, we conclude that MH is a valid FE meta-analysis method, even in the presence of DZ studies. The PL method is a good alternative, but it has the disadvantage of being more computationally-involved and may run into numerical issues, especially when the proportion of DZ studies is high. When the meta-analysis includes only DZ studies, none of the methods work.

Regarding SA studies, results suggested that they contained more relevant information than DZ studies. Indeed, whereas bias and coverage rates obtained by the PL and MH methods were found to be similar, the latter provided CIs wider as the proportion of SA studies increased, suggesting a loss of precision related to the non-inclusion of the information contained in SA studies. Based on these results, PL should be favored in the presence of SA studies. With 100% SA studies, the MH method breaks down, whereas we found that the PL approach still performed very well (additional simulations; results not shown).

In additional simulations, we compared our PL approach to the beta-binomial model discussed by Kuss [8], which also allows including SA and DZ studies. We found that both methods performed similarly, whatever the proportion of SA and DZ studies and the ES considered (results not shown). However, the Beta-Binomial model encountered more numerical issues (e.g., the number of converged runs when estimating the OR was systematically below 90%).

To sum up, in settings with DZ studies, we recommend using the MH method, although this method exclude the information contained in these studies. In settings with SA studies, we recommend using the PL approach, which was shown to be more efficient. However, when the ES

of interest is the RD and the proportion of SA studies is very high (75% or more), the meta-analyst should be aware that the CIs computed using the PL approach may have coverage rates slightly below nominal.

The main limitation to these recommendations is that they only apply to the framework of homogeneous treatment effects. In practice, there are many situations where this assumption is likely to be violated (Kontopantelis et al. [9]) and the reader must keep in mind that the MH method is not valid when treatment effects are heterogeneous (Kuss [8]). As for the PL approach, it can easily be adapted to account for heterogeneity in treatment effects by including regressors (meta-regression). Moreover, our simulations did not cover all the possible settings. For instance, we did not consider the case of study's scarcity (i.e., small K), which can impact the methods' performance.

In a future research we will focus on adapting the PL approach to the framework of heterogeneous treatment effects. It would also be worth seeking alternatives to our calibrated CIs for the RD, such as the use of the profile likelihood method (Böhning et al. [1]).

Acknowledgement

The work of RPR was supported by the Swiss National Science Foundation (Grant Number: 200021_169504/1).

References

[1] D. Böhning, R. Kuhnert and S. Rattanasiri, Meta-Analysis of Binary Data Using Profile Likelihood, CRC Press, 2008.

[2] M. J. Bradburn, J. J. Deeks, J. A. Berlin and A. Russell Localio, Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events, Statistics in Medicine 26(1) (2007), 53-77.

DOI: https://doi.org/10.1002/sim.2528

[3] S. Dias and A. E. Ades, Absolute or relative effects? Arm-based synthesis of trial data, Research Synthesis Methods 7(1) (2015), 23-28.

DOI: https://doi.org/10.1002/jrsm.1184

[4] O. Efthimiou, Practical guide to the meta-analysis of rare events, Evidence-Based Mental Health 21(2) (2018), 72-76.

DOI: http://dx.doi.org/10.1136/eb-2018-102911

[5] J. O. Friedrich, N. K. Adhikari and J. Beyene, Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data, BMC Medical Research Methodology 7 (2007); Article 5.

DOI: https://doi.org/10.1186/1471-2288-7-5

[6] F. Keus, J. Wetterslev, C. Gluud, H. G. Gooszen and C. J. van Laarhoven, Robustness assessments are needed to reduce bias in meta-analyses that include zero-event randomized trials, American Journal of Gastroenterology 104(3) (2009), 546-551.

[7] O. Kuss, J. F. Gummert and J. Borgermann, Meta-analysis with rare events should use adequate methods, The Journal of Thoracic and Cardiovascular Surgery 136(1) (2008), 241.

[8] O. Kuss, Statistical methods for meta-analyses including information from studies without any events – add nothing to nothing and succeed nevertheless, Statistics in Medicine 34(7) (2015), 1097-1116.

DOI: https://doi.org/10.1002/sim.6383

[9] E. Kontopantelis, D. A. Springate and D. Reeves, A re-analysis of the Cochrane library data: The dangers of unobserved heterogeneity in meta-analyses, Plos One 8(7) (2013), e69930.

DOI: https://doi.org/10.1371/journal.pone.0069930

[10] I. C. Marschner and A. C. Gillett, Relative risk regression: Reliable and flexible methods for log-binomial models, Biostatistics 13(1) (2012), 179-192.

DOI: https://doi.org/10.1093/biostatistics/kxr030

[11] D. Moher, J. Tetzlaff, A. A. Tricco, M. Sampson and D. G. Altman, Epidemiology and reporting characteristics of systematic reviews, PLoS Medicine 4(3) (2007), e78.

DOI: https://doi.org/10.1371/journal.pmed.0040078

[12] S. L. T. Normand, Meta-analysis: Formulating, evaluating, combining, and reporting, Statistics in Medicine 18(3) (1999), 321-359.

DOI: https://doi.org/10.1002/(SICI)1097-0258(19990215)18:3<321::AID-SIM28>3.0.CO;2-P

[13] R. Piaget-Rossel and P. Taffé, Meta-analysis of rare events under the assumption of a homogeneous treatment effect, Accepted for publication in Biometrical Journal (2019).

[14] C. Röver, F. Knapp and T. Friede, Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies, BMC Medical Research Methodology 15 (2015); Article 99.

DOI: https://doi.org/10.1186/s12874-015-0091-1

[15] M. J. Sweeting, A. J. Sutton and P. C. Lambert, What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data, Statistics in Medicine 23(9) (2004), 1351-1375.

DOI: https://doi.org/10.1002/sim.1761

[16] A. Toussaint, Width of Margins in Phyllodes Tumors of the Breast: The Controversy Drags on: A Systematic Review, Diplôme inter-Universitaire de Sénologie et Pathologie Mammaire, François-Rabelais University and Rennes University, Tours and Rennes, 2017.

[17] H. White, Maximum likelihood estimation of misspecified models, Econometrica 50(1) (1982), 1-26.

■

A pseudo-likelihood approach for the meta-analysis of rare events with treatment effect heterogeneity

Journal:	*Statistical Methods in Medical Research*
Manuscript ID	SMM-19-0549
Manuscript Type:	Original Article
Keywords:	Meta-analysis, Prediction interval, Pseudo-likelihood, Rare events, Treatment effect heterogeneity
Abstract:	In the presence of rare events and treatment effect heterogeneity, most of the alternative meta-analysis methods developed so far have focused on the odds ratio. However, using this effect size raises some challenges as it is a non-collapsible measure. In this paper, we have extended the pseudo-likelihood approach, previously proposed for homogeneous treatment effect, to the setting of treatment effect heterogeneity. This approach allows one not only to estimate the odds ratio, but also the relative risk and the risk difference, both collapsible and more intuitive effect sizes. Using simulations, we assessed the performance of the extended pseudo-likelihood approach in settings with rare events and treatment effect heterogeneity. Unlike conventional random-effects meta-analyses, where the focus is usually on the mean treatment effect parameter, its confidence interval, and, less frequently, on the heterogeneity parameter, we also considered the performance of the prediction interval. Our results showed that the extension of the pseudo-likelihood approach provided good estimates of the mean treatment effect. Moreover, we found that reporting the prediction interval was advantageous since even when the heterogeneity parameter was estimated as zero, the width of the interval was always non-null and, in most of the scenarios considered, even conservative.

Title:
A pseudo-likelihood approach for the meta-analysis of rare events with treatment effect heterogeneity

Authors:
Romain PIAGET-ROSSEL[1]: romain.piaget-rossel@unil.ch;
Patrick TAFFÉ[*]: patrick.taffe@unisante.ch

Mailing address:
University of Lausanne, Center for Primary Care and Public Health
Route de la Corniche 10, 1010 Lausanne, Switzerland

Abstract

In the presence of rare events and treatment effect heterogeneity, most of the alternative meta-analysis methods developed so far have focused on the odds ratio. However, using this effect size raises some challenges as it is a non-collapsible measure. In this paper, we have extended the pseudo-likelihood approach, previously proposed for homogeneous treatment effect, to the setting of treatment effect heterogeneity. This approach allows one not only to estimate the odds ratio, but also the relative risk and the risk difference, both collapsible and more intuitive effect sizes. Using simulations, we assessed the performance of the extended pseudo-likelihood approach in settings with rare events and treatment effect heterogeneity. Unlike conventional random-effects meta-analyses, where the focus is usually on the mean treatment effect parameter, its confidence interval, and, less frequently, on the heterogeneity parameter, we also considered the performance of the prediction interval. Our results showed that the extension of the pseudo-likelihood approach provided good estimates of the mean treatment effect. Moreover, we found that reporting the prediction interval was advantageous since even when the heterogeneity parameter was estimated as zero, the width of the interval was always non-null and, in most of the scenarios considered, even conservative.

Key words: Meta-analysis, Prediction interval, Pseudo-likelihood, Rare events, Treatment effect heterogeneity

[*] Corresponding author.

1. Introduction

Statistically assessing the effect of a treatment on the incidence of a rare event is a challenging task. Indeed, because of monetary and time constraints, it is often not feasible to design a study of proper size (i.e. one that allows observing a sufficiently large number of events to permit reliable inference). Therefore, in rare event settings, it seems particularly relevant to conduct meta-analyses (MAs) to increase the sample size from n_k to $\sum_{k=1}^{K} n_k$, n_k being the sample size of study *k*. However, as many researches have now highlighted, the classical MA method that consists in weighting each primary study's effect size (ES) by the inverse of its variance plus the between-study heterogeneity (if any) breaks down with rare events[1,2]. This is mostly due to the fact that this method implies assumptions that only hold asymptotically[3,4]. Primary-studies plagued by event scarcity being commonly encountered in practice (e.g. in a random sample of 500 Cochrane reviews, Vandermeer et al. found that 30% of them contained at least one study with no event in either arms[5]), alternative methods that better accommodate rare events have received considerable attention in medical research[6].

When the treatment effect can be assumed homogeneous across the primary studies, it has been shown that the Mantel-Haenszel (MH) method without continuity correction (CC) yielded unbiased estimates of the odds ratio (OR), relative risk (RR) and risk difference (RD), even in situations of extremely rare events[7]. However, this method excludes double-zero (DZ) and single-arm (SA) studies. To cope with this, the binomial likelihood has been recommended and the link function adapted to get the OR (logit), the RR (log), and the RD (identity)[8]. However, it is well known that the maximization of the binomial likelihood function using either the log or identity links is plagued by non-convergence issues and valid parameter space violation[9,10,11]. Therefore, many alternative methods have been developed,

particularly in the setting of primary studies[12]. However, in the context of MA, little is known regarding the performance of these alternative methods. Very recently, one study investigated the performance of a pseudo-likelihood (PL) approach in the context of a homogeneous treatment effect[13]. It was found that this approach performed as well as the MH method when both study's arms were present and was more efficient when there were many SA studies.

When the treatment effect is heterogeneous, although some authors have proposed ways to estimate the RR or the RD[14,15,16], the main focus with binary data has been on the estimation of the OR[17,18,19,20,21]. However, there are fundamental difficulties with the OR – not to mention the problem of interpretation – as it is (generally) a non-collapsible measure[7,22]. For instance, when pooling studies having adopted different sampling design such as pairing or matching, one has to distinguish between conditional and marginal ORs. These issues do not arise when computing the RR or the RD. Therefore, the goal of this study was to extend the pseudo-likelihood approach of Piaget-Rossel and Taffé[13] to the setting of a heterogeneous treatment effect and assess the performance of this methodology to estimate the OR, RR, and RD.

Curiously, despite clear recommendations for reporting the prediction interval (PI) in the presence of treatment effect heterogeneity, current practice is still to report the mean treatment effect parameter μ along with its CI, as well as some measure of between-study heterogeneity, such as the between-study variance τ^2 [23]. In a recent and often cited simulation study, the author assessed the performance of several MA methods based only on the estimation of μ along with its CI, without even considering τ^2 [24]. Focusing on the mean treatment effect in settings with between-study heterogeneity is clearly not sufficient since one fails to describe the whole range of values that can be taken by the ES. Additionaly reporting an estimate of the between-study variance parameter does not always allow one to comprehensively describe the treatment effect heterogeneity. Indeed, estimation of this

parameter is difficult (especially with rare events) and estimates obtained can be zero despite the presence of genuine heterogeneity[25]. The use of the conventional PI may solve this conundrum and offer a better description of the treatment effect heterogeneity. Indeed, its computation involving not only the between-study variance, but also the within-study variance[26], non-null PIs can be obtained even when the estimation of the between-study variance parameter is zero.

Therefore, in this study, the focus was not only on the mean and between-study variance parameters, but also on the performance of the conventional prediction interval. Recently, many authors have investigated the performance of several alternative PIs and concluded that the conventional approach did not always perform well[27,28,29]. However, they conducted their analyses using the two-stage model, where the ESs and their variance are estimated first and then used in further computations of other parameters[30]. In contrast, a one-stage approach was used in the present paper.

The remaining of this paper is organized as follows. In Section 2, we proposed an extension of the PL approach that accounts for between-study heterogeneity in the treatment effect. We described the models underlying this approach for the OR, the RR, and the RD. To illustrate the use of the PL approach in rare event settings, we then compared it to the DerSimonian and Laird (DL) method using data from a systematic review on the effect of anti-infective-treated central venous catheters on catheter-related bloodstream infection in the acute care setting[31]. Section 4 was dedicated to Monte-Carlo simulations (model's description and results), which were conducted to assess the performance of the PL method under various rare-event scenarios. Finally, we discussed the results obtained and provided recommendations in Section 5.

2 Meta-analysis with between-study heterogeneity in the treatment effect

In MA of binary data, it is common to assume that the number of events X_{kj} in study k and treatment arm j follows a binomial distribution

$$X_{kj} : \text{bin}(n_{kj}, \pi_{kj}), \quad k \in \{1,...,K\}, \quad j \in \{C,T\}$$

where n_{kj} denotes the sample size and π_{kj} the probability of an event in study k and treatment arm j. To measure the treatment effect, π_{kC} and π_{kT} are contrasted using an ES (e.g. RR = π_{kT}/π_{kC}). When this ES is assumed to vary across the studies, the meta-analyst should use a so-called "random-effects" approach, which simply consists in using a MA method that allows for between-study heterogeneity in the treatment effect. Since each study is estimating a different ES, inference should target the whole distribution of the ES (and not simply its mean) to provide a thorough overview of the treatment effect.

2.1 The pseudo-likelihood approach

The PL approach for the OR is based on the following model:

$$X_{kj} : \text{Bin}(n_{kj}, \pi_{kj}) \quad k \in \{1,...,K\} \quad j \in \{C,T\}$$

$$\text{logit}(\pi_{kj}) = \psi_k + \mu_k T_{kj}$$

$$\begin{pmatrix} \psi_k \\ \mu_k \end{pmatrix} : \text{Normal}\left(\begin{pmatrix} \psi \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2_\psi & \cdot \\ \sigma_{\psi\mu} & \sigma^2_\mu \end{pmatrix} \right) \tag{1}$$

where X_{kj} is the number of events occurring in arm j of study k, n_{kj} is the sample size, π_{kj} is the probability of an event, and T_{kj} is the treatment dummy, taking value 1 for the treated group and 0 otherwise. Note that since model (1) allows for a correlation between the two arms, it does not impose any restriction on how variability in the two groups compares. The two

parameters of interest are the mean μ and the between-study variance $\sigma_\mu^2 \equiv \tau^2$, both measured on the log(OR)'s scale. Note that this is a "pseudo-likelihood" approach because we are not assuming that the bivariate distribution of ψ_k and μ_k is the correct one.

Estimates of μ and τ^2 are obtained by maximizing the following marginal pseudo-likelihood function:

$$L = \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[\prod_{j \in \{C,T\}} \binom{n_{kj}}{X_{kj}} \pi_{kj}(\psi_k, \mu_k)^{X_{kj}} (1 - \pi_{kj}(\psi_k, \mu_k))^{n_{kj} - X_{kj}} \right] f(\psi_k, \mu_k) d\psi_k d\mu_k$$

where $\pi_{kj}(\psi_k, \mu_k) = \text{invlogit}\left(\psi_k + \mu_k T_{kj}\right)$ and $f(\psi_k, \mu_k)$ is the bivariate Normal density from model (1).

For the RR, the Binomial distribution is approximated by means of a Poisson distribution with parameter $\lambda_{kj} = n_{kj} * \pi_{kj}$ and the following model is used:

$$X_{kj} : \text{Poisson}(\lambda_{kj}) \quad k \in \{1, ..., K\} \quad j \in \{C, T\}$$

$$\log(\pi_{kj}) = \psi_k + \mu_k T_{kj}$$

$$\begin{pmatrix} \psi_k \\ \mu_k \end{pmatrix} : \text{Normal}\left(\begin{pmatrix} \psi \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_\psi^2 & \cdot \\ \sigma_{\psi\mu} & \sigma_\mu^2 \end{pmatrix} \right) \tag{2}$$

Again, the two parameters of interest are the mean treatment effect μ and the between-study variance $\sigma_\mu^2 \equiv \tau^2$, now measured on the log(RR)'s scale.

Estimates of μ and τ^2 are obtained by maximizing the following marginal pseudo-likelihood function:

$$L = \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[\prod_{j \in \{C,T\}} \frac{\lambda_{kj}(\psi_k, \mu_k)^{X_{kj}} \exp\left[-\lambda_{kj}(\psi_k, \mu_k)\right]}{X_{kj}!} \right] f(\psi_k, \mu_k) d\psi_k d\mu_k$$

where $\lambda_{kj}\left(\psi_k,\mu_k\right)=n_{kj}\exp\left(\psi_k+\mu_k T_{kj}\right)$ and $f(\psi_k,\mu_k)$ is the bivariate Normal distribution from model (2).

To obtain an estimate for the RD, the PL approach involves the approximation of the Binomial distribution by means of a Normal distribution for $p_{kj}=X_{kj}/n_{kj}$:

$$p_{kj}:\ \text{Normal}(\pi_{kj},\upsilon_{kj}^2)\quad k\in\{1,...,K\}\quad j\in\{C,T\}$$

$$\pi_{kj}=\psi_k+\mu_k T_{kj}$$

$$\begin{pmatrix}\psi_k\\ \mu_k\end{pmatrix}:\ \text{Normal}\left(\begin{pmatrix}\psi\\ \mu\end{pmatrix},\begin{pmatrix}\sigma_\psi^2 & \cdot\\ \sigma_{\psi\mu} & \sigma_\mu^2\end{pmatrix}\right) \tag{3}$$

With this modeling, the mean treatment effect μ and the heterogeneity $\sigma_\mu^2\equiv\tau^2$ are measured on the RD's scale.

The two parameters of interest μ and τ^2 are estimated by maximizing the following marginal pseudo-likelihood function:

$$L=\prod_{k=1}^{K}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left[\prod_{j\in\{C,T\}}\frac{1}{\sqrt{2\pi*\upsilon_{kj}^2}}\exp\left(\frac{p_{kj}-\pi_{kj}\left(\psi_k,\mu_k\right)}{2\upsilon_{kj}^2}\right)\right]f(\psi_k,\mu_k)d\psi_k d\mu_k$$

where $\pi_{kj}\left(\psi_k,\mu_k\right)\equiv E(p_{kj})=\psi_k+\mu_k T_{kj}$, $\upsilon_{kj}^2\equiv Var(p_{kj})=\pi_{kj}*(1-\pi_{kj})/n_{kj}$ and $f(\psi_k,\mu_k)$ is the bivariate Normal distribution from model (3). Given that the variances υ_{kj}^2 are unknown and difficult to estimate (the above model is highly non-linear and computation of the integral with good precision is challenging, particularly with rare events), we decided to use instead the robust sandwich estimate of the variance-covariance matrix[32]. Moreover, to achieve appropriate empirical coverage rates, we computed calibrated CIs for the RD using the 98.5th quantile of the Standard Normal distribution[13].

2.2 Illustrative example

To illustrate the method described above, we used the data of a MA conducted by Niël-Weise et al. (2007) on the effect of anti-infective-treated central venous catheters on catheter-related bloodstream infection in the acute care setting (Table 1)[31]. This MA was based on 18 studies, in which catheter-related bloodstream infection was a rare event. There were 5 single-zero and 1 double-zero studies, and the average proportion of observed event was about 4.38% and 1.26% in the control and treated arms, respectively.

Table 1. Dataset from the study on the efficacy of anti-infective catheters

	Standard catheter (control)		Anti-infective catheter (treatment)	
Study (k)	Nb of infections (x_{kC})	Nb of patients (n_{kC})	Nb of infections (x_{kT})	Nb of patients (n_{kT})
1	3	117	0	116
2	3	35	1	44
3	9	195	2	208
4	7	136	0	130
5	6	157	5	151
6	4	139	1	98
7	3	177	1	174
8	2	39	1	74
9	19	103	1	97
10	2	122	1	113
11	7	64	0	66
12	1	58	0	70
13	5	175	3	188
14	11	180	6	187
15	0	105	0	118
16	1	262	0	252
17	3	362	1	345
18	1	69	4	64

Results obtained by applying the DL and PL methods on the illustrative dataset are displayed in Table 2. One first thing worth mentioning is that the DL approach cannot handle SZ and DZ studies when computing the RR or the OR without adding a CC. However, since the use

of a CC has been shown to lead to biased estimates in rare event settings[7], we computed combined estimates by excluding SZ and DZ studies when using the DL method.

Both PL and DL methods found that, in average, a patient with an anti-infective catheter will have less infection than a patient with a standard one. However, estimates obtained by these two methods differed widely. The mean OR and RR computed using the DL method were around 0.45, whereas those computed using the PL method were below 0.3. When measured on the RD scale, the estimates of the mean treatment effect was -1.78% for the DL method and -3.11% for the PL approach. Notice that for the PL approach, we had to simplify model (3) by assuming $\sigma_{\psi\mu} = 0$ to achieve convergence.

Another noticeable difference between the PL and DL approaches was that the latter did not detect any heterogeneity when computing the OR and the RR, whereas the former provided estimates of 0.71 and 0.62, respectively. This translated into large differences in terms of PIs' width, such that a DL user would conclude that the treatment is beneficial to almost all (i.e. at least 95%) the patients, whereas a PL user would conclude that the treatment is not effective for all patients (i.e. PI obtained contained the value 1). On the RD scale, inversely, the DL method obtained a larger estimate of τ^2 and, thus, a larger PI, which contained the value 0. No CI was reported for the heterogeneity when using the DL method because this method does not include an estimation of the standard error for this parameter.

Table 2. Meta-analysis of the illustrative dataset

Effect size	Method	μ (CI)	τ^2 (CI)	Prediction interval
OR	DL	0.44 (0.27; 0.72)	0 (.)	[0.26; 0.75]
	PL	0.28 (0.13; 0.58)	0.71 (0.12; 4.04)	[0.04; 1.97]
RR	DL	0.46 (0.28; 0.74)	0 (.)	[0.27; 0.76]
	PL	0.29 (0.14; 0.59)	0.62 (0.10; 3.94)	[0.05; 1.81]
RD*	DL	-1.78 (-2.84; -0.71)	0.02 (.)	[-5.25; 1.70]
	PL	-3.11 (-5.35; -0.88)	0.00 (0.00; 0.02)	[-5.64; -0.59]

Note:
*Results provided in %.

3 Simulation study

3.1 Model

We generated meta-analyses of $K = 20$ primary studies with treatment arms' sample sizes ranging from 50 to 150 (i.e. n_{kT} : uniform$\{50;150\}$). Control arms' sample sizes were generated as $n_{kC} = n_{kT} + r$, with r : uniform$\{-15;15\}$. The event probability in the control group π_{kC} was assumed to be slightly heterogeneous and, thus, drawn from a uniform distribution with range $\left[\pi_{kC}^{L};\pi_{kC}^{U}\right]$, $\pi_{kC}^{L} = M_{\pi_c} - M_{\pi_c}/5$ and $\pi_{kC}^{U} = M_{\pi_c} + M_{\pi_c}/5$, where M_{π_c} is the mean event probability in the control group. The treatment group probability π_{kT} was derived from π_{kC} and the ES considered (i.e. $\pi_{kT} = \pi_{kC} * OR_k / \left(\pi_{kC} * OR_k + 1 - \pi_{kC}\right)$, $\pi_{kT} = \pi_{kC} * RR_k$ or $\pi_{kT} = \pi_{kC} + RD_k$). ESs were generated according to a uniform distribution with mean μ_{ES} and variance τ_{ES}^2 :

$$ES : \text{ uniform}\left[ES^{L};ES^{U}\right]$$

where both bounds were derived using the uniform distribution's properties (i.e. $ES^{L} = \mu_{ES} - \sqrt{3\tau_{ES}^2}$, $ES^{U} = \mu_{ES} + \sqrt{3\tau_{ES}^2}$) and ES = log(OR), log(RR) or RD. Finally, the number of events in both arms were generated by two binomial draws with respective sample sizes and event probabilities.

To study the issue of rare events, we considered three different values for the mean probability of an event in the control group M_{π_c} : 10%, 1%, and 0.5%. We also considered different values for the treatment effect and the heterogeneity (measured on the log(OR)'s scale), setting $\mu_{\log(OR)} \in \{-1.5;-0.5;0;0.5;1.5\}$ and $\tau_{\log(OR)}^2 \in \{0;0.1;0.5\}$. Values of $\mu_{\log(OR)}$ represent situations of large (negative or positive), moderate (negative or positive) and null

treatment effect, respectively. As for $\tau^2_{\log(OR)}$, we chose these three values to illustrate the impact of null, moderate and large heterogeneity in the treatment effect, respectively. To illustrate the impact of the choice of these three parameters' values on the scarcity level, we reported in the Appendix (Table A1) the average proportion of control and treated arms with zero event in the 45 scenarios considered.

For the log(OR), values of $\log(OR)^L$ and $\log(OR)^U$ were directly computed as $\mu_{\log(OR)} - \sqrt{3\tau^2_{\log(OR)}}$ and $\mu_{\log(OR)} + \sqrt{3\tau^2_{\log(OR)}}$. For the two other ESs, to insure generating probabilities within the range $[0,1]$, the lower and upper bounds were derived from $\log(OR)^L$ and π^L_{kC}, and $\log(OR)^U$ and π^L_{kC}, respectively. More precisely, for $b \in \{L,U\}$, we computed

$$\log(RR)^b = \log\left(\frac{\pi^L_{kC} * \exp\left(\log(OR)^b\right)}{\pi^L_{kC} * \exp\left(\log(OR)^b\right) + 1 - \pi^L_{kC}}\right) - \log\left(\pi^L_{kC}\right)$$

$$RD^b = \frac{\pi^L_{kC} * \exp\left(\log(OR)^b\right)}{\pi^L_{kC} * \exp\left(\log(OR)^b\right) + 1 - \pi^L_{kC}} - \pi^L_{kC}$$

Finally, for $ES \in \{\log(RR), RD\}$, μ_{ES} and τ^2_{ES} were retrieved from the lower and upper bounds values of the respective ES as $\mu_{ES} = \left(ES^L + ES^U\right)/2$ and $\tau^2_{ES} = \left(ES^U - ES^L\right)^2/12$.

For each scenario, we made 10 000 iterations. For each of the generated meta-analyses, we fitted model (1), (2), or (3), depending on the ES of interest. In order to achieve a sufficiently large number of converged runs, we simplified these three models by assuming $\sigma_{\psi\mu} = 0$.

We estimated the mean and the between-study variance of the ES of interest (i.e. $\hat{\mu}_{ES}$ and $\hat{\tau}^2_{ES}$) plus their respective standard error, from which we derived the 95% Wald CI for μ_{ES} and τ^2_{ES}, as well as the lower and upper bounds of the 95% PI[26]:

$$PI = \hat{\mu}_{ES} \pm t_{K-2,1-\alpha/2}\sqrt{\hat{\tau}^2_{ES} + \hat{V}\left(\hat{\mu}_{ES}\right)} \quad (4)$$

where $t_{K-2,1-\alpha/2}$ is the (1-α/2)th quantile of the t-student distribution with *K*-2 degrees of freedom and $\hat{V}\left(\hat{\mu}_{ES}\right)$ is the estimated variance of the mean ES's estimator.

Performance of the PL approach was assessed in terms of the difference between the median value of the estimates and the true value of the parameters μ_{ES} and τ^2_{ES}, coverage rate of the Wald CIs obtained for these two parameters, and difference between the median value of the estimated lower and upper bounds of the 95% PI and the 0.025th and 0.975th quantiles of the uniform distribution. We decided to report median instead of mean values of the estimates to avoid the influence of exceedingly large or small values obtained in some simulations.

3.2 Results

3.2.1 Odds ratio

Results obtained by the PL model when estimating the OR are provided in Table 3. Although we simplified model (1) by assuming $\sigma_{\psi\mu} = 0$, this model still encountered numerical issues, even in the scenarios with common events. Overall, the proportion of converged runs varied between 45 and 75%.

In the 15 scenarios with common events (i.e. $M_{\pi_c} = 0.1$), the PL approach obtained good results. Estimates of the mean treatment effect were virtually unbiased and coverage rates for this parameter were nominal in most of the scenarios and never felt below 91%. Regarding

the heterogeneity, the model slightly underestimated $\hat{\tau}^2_{\log(OR)}$, especially in scenarios with large negative mean treatment effect, and CIs obtained for this parameter were too conservative (i.e. most of the time above 98%). As for the lower and upper bounds of the PIs, they were close to the 0.025th and 97.5th theoretical quantiles and all computed PIs contained these two quantiles within their two bounds (i.e. they were conservative).

With rare events (i.e. $M_{\pi_c} = 0.01$), things remained very good regarding the estimation of the mean treatment effect, except in the scenarios with large heterogeneity where estimates tended to have a small positive bias and coverage rates were sligthly below nominal. In all scenarios, estimates of the heterogeneity were biased towards zero and CIs computed for this parameter were not reliable. However, performance of the PIs was satisfactory, since they contained the two theoretical quantiles, except in some scenarios with very large heterogeneity.

With extremely rare events (i.e. $M_{\pi_c} = 0.005$), estimates of the mean treatment effect was unbiased, except in scenarios with large heterogeneity where a small positive bias was observed. Coverage rates of this parameter's CIs were nominal in the 15 scenarios considered with extremely rare events. Estimates provided for the heterogeneity parameter was 0, except in the scenario with large heterogeneity and large positive treatment effect, and CIs were not reliable. As for the PIs, they were larger than under the scenarios with more common events but, again, they always contained both theoretical quantiles when the heterogeneity was moderate. With large heterogeneity, the lower theoretical quantile was not always contained within the PI but the estimated lower bound was rather close to the theoretical one.

Table 3. Pseudo-likelihood performance for the meta-analysis of the odds ratio

M_{π_c}	$\tau^2_{\log(OR)}$	$\mu_{\log(OR)}$	$[q_{0.025}; q_{0.975}]$	Conv. runs*	$\hat{\mu}_{\log(OR)}$	Co. rate $\mu_{\log(OR)}$*	$\hat{\tau}^2_{\log(OR)}$	Co. rate $\tau^2_{\log(OR)}$*	$\text{PI}_{\log(OR)}$
0.1	0	-1.5	[.]	67.31	-1.52	95.80	-	-	-
0.1	0	-0.5	[.]	69.77	-0.51	95.40	-	-	-

0.1	0	0	[.]	70.75	-0.01	94.73	-	-	-
0.1	0	0.5	[.]	72.10	0.50	95.56	-	-	-
0.1	0	1.5	[.]	76.56	1.50	95.04	-	-	-
0.1	0.1	-1.5	[-2.02; -0.98]	63.42	-1.50	93.83	0.02	95.48	[-2.03; -0.95]
0.1	0.1	-0.5	[-1.02; 0.02]	63.92	-0.49	93.73	0.06	98.27	[-1.03; 0.08]
0.1	0.1	0	[-0.52; 0.52]	64.45	0.01	93.05	0.07	98.99	[-0.56; 0.59]
0.1	0.1	0.5	[-0.02; 1.02]	66.03	0.51	92.75	0.07	98.91	[-0.09; 1.10]
0.1	0.1	1.5	[0.98; 2.02]	66.80	1.51	93.29	0.08	99.52	[0.90; 2.11]
0.1	0.5	-1.5	[-2.66; -0.34]	57.97	-1.46	91.19	0.38	99.43	[-2.78; -0.12]
0.1	0.5	-0.5	[-1.66; 0.66]	58.99	-0.48	91.80	0.43	99.54	[-1.87; 0.92]
0.1	0.5	0	[-1.16; 1.16]	58.22	0.02	92.63	0.45	98.88	[-1.40; 1.44]
0.1	0.5	0.5	[-0.66; 1.66]	58.94	0.51	93.18	0.46	98.44	[-0.92; 1.94]
0.1	0.5	1.5	[0.34; 2.66]	59.37	1.50	92.82	0.46	98.33	[0.08; 2.93]
0.01	0	-1.5	[.]	52.33	-1.58	97.47	-	-	-
0.01	0	-0.5	[.]	53.58	-0.54	96.45	-	-	-
0.01	0	0	[.]	54.33	-0.01	96.30	-	-	-
0.01	0	0.5	[.]	54.11	0.48	96.05	-	-	-
0.01	0	1.5	[.]	55.39	1.50	95.56	-	-	-
0.01	0.1	-1.5	[-2.02; -0.98]	51.21	-1.56	97.10	0.00	85.55	[-2.77; -0.33]
0.01	0.1	-0.5	[-1.02; 0.02]	52.36	-0.50	95.95	0.00	89.85	[-1.33; 0.41]
0.01	0.1	0	[-0.52; 0.52]	52.31	0.01	95.55	0.00	94.74	[-0.71; 0.79]
0.01	0.1	0.5	[-0.02; 1.02]	51.49	0.53	95.47	0.00	97.23	[-0.13; 1.23]
0.01	0.1	1.5	[0.98; 2.02]	52.38	1.54	95.23	0.00	99.28	[0.94; 2.17]
0.01	0.5	-1.5	[-2.66; -0.34]	51.28	-1.42	93.82	0.00	98.77	[-2.68; -0.13]
0.01	0.5	-0.5	[-1.66; 0.66]	47.98	-0.38	92.48	0.00	99.69	[-1.37; 0.74]
0.01	0.5	0	[-1.16; 1.16]	46.49	0.11	92.30	0.12	99.82	[-0.91; 1.24]
0.01	0.5	0.5	[-0.66; 1.66]	44.66	0.60	91.85	0.21	99.88	[-0.49; 1.79]
0.01	0.5	1.5	[0.34; 2.66]	46.27	1.58	92.09	0.29	99.90	[0.34; 2.85]
0.005	0	-1.5	[.]	47.25	-1.61	97.60	-	-	-
0.005	0	-0.5	[.]	51.05	-0.55	96.90	-	-	-
0.005	0	0	[.]	51.38	-0.01	96.71	-	-	-
0.005	0	0.5	[.]	53.09	0.49	96.53	-	-	-
0.005	0	1.5	[.]	52.70	1.51	96.02	-	-	-
0.005	0.1	-1.5	[-2.02; -0.98]	48.35	-1.54	96.93	0.00	80.41	[-3.03; 0.17]
0.005	0.1	-0.5	[-1.02; 0.02]	50.21	-0.52	96.87	0.00	89.34	[-1.68; 0.73]
0.005	0.1	0	[-0.52; 0.52]	50.40	-0.00	95.91	0.00	91.38	[-1.00; 1.08]
0.005	0.1	0.5	[-0.02; 1.02]	50.35	0.52	96.15	0.00	95.09	[-0.35; 1.47]
0.005	0.1	1.5	[0.98; 2.02]	50.26	1.53	96.18	0.00	98.50	[0.77; 2.34]
0.005	0.5	-1.5	[-2.66; -0.34]	47.29	-1.40	94.46	0.00	99.43	[-2.90; 0.24]
0.005	0.5	-0.5	[-1.66; 0.66]	47.66	-0.38	94.06	0.00	99.60	[-1.59; 0.97]
0.005	0.5	0	[-1.16; 1.16]	47.21	0.15	93.69	0.00	99.83	[-0.96; 1.39]
0.005	0.5	0.5	[-0.66; 1.66]	47.31	0.67	94.23	0.00	99.67	[-0.36; 1.83]
0.005	0.5	1.5	[0.34; 2.66]	46.04	1.64	94.07	0.15	99.81	[0.53; 2.84]

Note: $q_{0.025}$ = 0.025th quantile of the uniform distribution. $q_{0.975}$ = 0.975th quantile of the uniform distribution. Conv. runs = proportion of converged runs. Co. rate θ = Coverage rate of the Wald CI interval for the parameter θ.

*Values reported in %

3.2.2 Relative risk

Results obtained by the PL approach when meta-analyzing the RR are displayed in Table 4. Overall, they were similar to those obtained when estimating the OR. Estimates of the mean treatment effect were close to the true parameter's values, even with extremely rare events, and the largest biases were observed in scenarios with large negative mean treatment effect. Coverage rates of the CIs computed for this parameter were close to nominal, with most of the values contained between 93 and 97%. For the heterogeneity parameter, estimates were biased towards zero, especially in rare-event scenarios, and CIs were either invalid (i.e. coverage rates largely below 95%) or too conservative (i.e. coverage rates close to 100%). Nevertheless, performance of the PIs were much better with bounds close to the theoretical quantiles in many scenarios. When PIs were too wide, which was especially the case in scenarios with extremely rare events and large negative treatment effect, they at least contained the two theoretical quantiles.

Table 4. Pseudo-likelihood performance for the meta-analysis of the relative risk

M_{π_c}	$\tau^2_{\log(RR)}$	$\mu_{\log(RR)}$	$[q_{0.025}; q_{0.975}]$	Conv. runs*	$\hat{\mu}_{\log(RR)}$	Co. rate $\mu_{\log(RR)}$*	$\hat{\tau}^2_{\log(RR)}$	Co. rate $\tau^2_{\log(RR)}$*	$\text{PI}_{\log(RR)}$
0.1	0	-1.44	[.]	54.92	-1.45	95.94	-	-	-
0.1	0	-0.47	[.]	61.37	-0.47	96.09	-	-	-
0.1	0	0.00	[.]	65.74	-0.00	96.32	-	-	-
0.1	0	0.45	[.]	72.11	0.45	96.64	-	-	-
0.1	0	1.25	[.]	81.42	1.26	96.68	-	-	-
0.1	0.01	-1.44	[-1.60; -1.27]	54.14	-1.45	96.16	0.00	72.29	[-1.84; -1.04]
0.1	0.01	-0.47	[-0.62; -0.31]	58.96	-0.47	95.62	0.00	85.27	[-0.74; -0.19]
0.1	0.01	-0.00	[-0.15; 0.15]	64.36	0.00	95.70	0.00	90.28	[-0.23; 0.24]
0.1	0.01	0.45	[0.30; 0.59]	70.08	0.45	96.39	0.00	94.64	[0.24; 0.66]
0.1	0.01	1.25	[1.13; 1.37]	80.29	1.25	96.38	0.00	97.87	[1.08; 1.43]
0.1	0.24	-1.44	[-2.25; -0.64]	48.71	-1.41	92.24	0.14	99.28	[-2.24; -0.56]
0.1	0.22	-0.49	[-1.26; 0.29]	50.80	-0.47	92.70	0.16	99.73	[-1.33; 0.42]
0.1	0.21	-0.03	[-0.78; 0.72]	51.62	-0.01	92.64	0.16	99.80	[-0.86; 0.85]
0.1	0.19	0.41	[-0.30; 1.12]	53.90	0.43	92.56	0.15	99.78	[-0.38; 1.25]
0.1	0.13	1.18	[0.60; 1.76]	58.19	1.19	93.44	0.09	99.90	[0.54; 1.85]
0.01	0	-1.49	[.]	41.31	-1.63	97.75	-	-	-
0.01	0	-0.50	[.]	39.32	-0.55	96.72	-	-	-
0.01	0	0.00	[.]	39.25	-0.02	96.56	-	-	-
0.01	0	0.49	[.]	39.73	0.47	96.07	-	-	-

0.01	0	1.47	[.]	45.20	1.48	95.82	-	-	-
0.01	0.01	-1.49	[-1.66; -1.33]	41.07	-1.60	97.59	0.00	53.61	[-2.89; -0.31]
0.01	0.01	-0.50	[-0.66; -0.33]	39.75	-0.54	97.08	0.00	62.67	[-1.37; 0.37]
0.01	0.01	-0.00	[-0.16; 0.16]	38.20	-0.02	96.47	0.00	71.48	[-0.73; 0.74]
0.01	0.01	0.49	[0.33; 0.66]	40.87	0.49	95.86	0.00	80.72	[-0.14; 1.14]
0.01	0.01	1.47	[1.31; 1.63]	46.36	1.48	95.41	0.00	92.24	[0.96; 2.02]
0.01	0.25	-1.49	[-2.32; -0.67]	41.42	-1.54	96.45	0.00	94.70	[-2.85; -0.22]
0.01	0.25	-0.50	[-1.32; 0.32]	40.35	-0.47	95.19	0.00	97.82	[-1.40; 0.56]
0.01	0.25	-0.00	[-0.82; 0.81]	40.56	0.06	94.72	0.00	99.05	[-0.75; 0.96]
0.01	0.24	0.49	[-0.32; 1.30]	43.39	0.55	94.22	0.00	99.44	[-0.21; 1.39]
0.01	0.23	1.46	[0.67; 2.25]	47.70	1.53	93.06	0.07	99.78	[0.78; 2.35]
0.005	0	-1.50	[.]	39.49	-1.62	98.02	-	-	-
0.005	0	-0.50	[.]	37.99	-0.60	97.34	-	-	-
0.005	0	0.00	[.]	37.67	-0.04	96.84	-	-	-
0.005	0	0.50	[.]	36.84	0.47	96.55	-	-	-
0.005	0	1.49	[.]	40.98	1.48	96.56	-	-	-
0.005	0.01	-1.50	[-1.66; -1.33]	38.50	-1.64	97.87	0.00	48.89	[-3.34; 0.19]
0.005	0.01	-0.50	[-0.66; -0.33]	38.99	-0.57	97.00	0.00	61.60	[-1.76; 0.72]
0.005	0.01	-0.00	[-0.16; 0.16]	38.07	-0.03	97.03	0.00	67.49	[-1.05; 1.06]
0.005	0.01	0.50	[0.33; 0.66]	37.67	0.48	95.78	0.00	73.56	[-0.41; 1.44]
0.005	0.01	1.49	[1.32; 1.65]	40.17	1.50	96.51	0.00	86.68	[0.75; 2.25]
0.005	0.25	-1.50	[-2.32; -0.68]	39.44	-1.59	96.50	0.00	90.79	[-3.24; 0.24]
0.005	0.25	-0.50	[-1.32; 0.32]	38.22	-0.46	96.08	0.00	96.91	[-1.68; 0.88]
0.005	0.25	-0.00	[-0.82; 0.82]	38.56	0.05	95.59	0.00	98.45	[-1.00; 1.24]
0.005	0.25	0.49	[-0.32; 1.31]	39.47	0.58	95.92	0.00	98.78	[-0.36; 1.63]
0.005	0.24	1.48	[0.67; 2.29]	44.93	1.59	95.08	0.00	99.75	[0.74; 2.50]

Note: $q_{0.025}$ = 0.025th quantile of the uniform distribution. $q_{0.975}$ = 0.975th quantile of the uniform distribution. Conv. runs = proportion of converged runs. Co. rate θ = Coverage rate of the Wald CI interval for the parameter θ.

*Values reported in %

3.2.3 Risk difference

Results regarding the estimation of the RD using the PL approach are provided in Table 5. As for the OR and RR, estimation of model (3) was difficult and the proportion of converged runs was never above 80%. More numerical issues were observed in scenarios with large heterogeneity and extremely rare events, especially with large positive mean treatment effect.

Estimates of the mean treatment effect parameter were satisfactory in all scenarios considered, with virtually unbiased estimates and nominal coverage rates of the CIs, even in scenarios with extremely rare events. Conversely, estimates of the heterogeneity parameter were most of the times biased and CIs invalid, with coverage rates largely below 95%. Notice

however that the true parameter's value was quite low in most of the scenarios simulated (i.e. results are provided in ‰), meaning that in absolute terms, biases were not so large. Moreover, the performance of the PIs, which are also used to quantify the heterogeneity in the treatment effect, was much better. Although the median PIs tended to be too conservative (i.e. larger than the theoretical inter-quantile range), they contained $q_{0.025}$ and $q_{0.975}$ within their lower and upper bounds in the 15 scenarios simulated.

Table 5. Pseudo-likelihood performance for the meta-analysis of the risk difference

M_{π_c}	τ^2_{RD} **	μ_{RD} *	$[q_{0.025}; q_{0.975}]$	Conv. runs*	$\hat{\mu}_{RD}$ *	Co. rate μ_{RD} *	$\hat{\tau}^2_{RD}$ **	Co. rate τ^2_{RD} *	PI_{RD}
0.1	0	-6.10	[.]	78.81	-6.11	95.11	-	-	-
0.1	0	-2.99	[.]	74.52	-2.98	94.57	-	-	-
0.1	0	0.00	[.]	70.96	0.01	94.74	-	-	-
0.1	0	4.54	[.]	67.83	4.54	95.00	-	-	-
0.1	0	20.04	[.]	60.89	20.03	95.20	-	-	-
0.1	0.00	-6.07	[-6.38; -5.76]	78.92	-6.07	94.84	0.00	7.04	[-7.75; -4.42]
0.1	0.02	-2.93	[-3.71; -2.14]	73.81	-2.94	95.23	0.00	20.27	[-5.13; -0.88]
0.1	0.05	0.09	[-1.12; 1.31]	70.05	0.10	94.62	0.03	32.20	[-2.87; 2.89]
0.1	0.12	4.66	[2.85; 6.47]	66.06	4.67	94.46	0.44	40.56	[0.07; 9.31]
0.1	0.41	20.18	[16.86; 23.49]	58.33	20.14	94.55	1.47	39.12	[12.03; 28.28]
0.1	0.11	-5.39	[-7.10; -3.68]	74.97	-5.41	94.66	0.00	17.84	[-7.31; -3.59]
0.1	0.67	-1.34	[-5.61; 2.92]	66.53	-1.37	94.77	0.47	75.54	[-6.12; 3.43]
0.1	1.54	2.33	[-4.13; 8.79]	57.36	2.29	94.67	1.51	93.87	[-5.93; 10.48]
0.1	3.24	7.55	[-1.82; 16.93]	40.88	7.46	93.69	3.06	97.58	[-4.20; 19.10]
0.1	9.64	23.09	[6.93; 39.24]	5.01	22.57	90.42	6.95	86.43	[5.08; 40.04]
0.01	0	-0.62	[.]	53.58	-0.62	93.75	-	-	-
0.01	0	-0.31	[.]	52.82	-0.32	94.62	-	-	-
0.01	0	0.00	[.]	52.00	-0.00	94.54	-	-	-
0.01	0	0.51	[.]	50.39	0.49	94.42	-	-	-
0.01	0	2.69	[.]	36.11	2.58	93.22	-	-	-
0.01	0.00	-0.62	[-0.65; -0.59]	53.05	-0.61	94.14	0.00	6.28	[-1.12; -0.11]
0.01	0.00	-0.31	[-0.39; -0.23]	53.06	-0.31	94.80	0.00	12.17	[-0.99; 0.26]
0.01	0.00	0.01	[-0.12; 0.14]	52.36	0.01	94.16	0.00	17.60	[-0.92; 0.79]
0.01	0.00	0.53	[0.32; 0.75]	49.39	0.51	94.35	0.05	18.38	[-0.98; 2.03]
0.01	0.01	2.74	[2.18; 3.29]	34.93	2.63	93.82	0.22	7.17	[-0.45; 5.73]
0.01	0.00	-0.55	[-0.72; -0.38]	53.08	-0.55	94.89	0.00	9.50	[-1.09; -0.04]
0.01	0.01	-0.12	[-0.57; 0.33]	51.93	-0.14	93.90	0.00	31.46	[-1.01; 0.58]
0.01	0.02	0.31	[-0.42; 1.04]	49.74	0.29	93.91	0.04	45.67	[-1.13; 1.71]
0.01	0.05	1.01	[-0.18; 2.20]	44.22	0.95	93.92	0.12	54.86	[-1.35; 3.30]
0.01	0.34	3.91	[0.86; 6.95]	11.01	3.55	88.83	0.43	91.01	[-0.83; 7.91]
0.005	0	-0.31	[.]	48.83	-0.31	94.29	-	-	-
0.005	0	-0.16	[.]	48.51	-0.15	94.50	-	-	-

0.005	0	0.00	[.]	48.84	-0.01	94.49	-	-	-
0.005	0	0.26	[.]	45.94	0.23	94.65	-	-	-
0.005	0	1.37	[.]	33.47	1.25	93.82	-	-	-
0.005	0.00	-0.31	[-0.32; -0.29]	48.89	-0.31	93.58	0.00	7.05	[-0.68; 0.03]
0.005	0.00	-0.15	[-0.19; -0.11]	49.02	-0.16	93.82	0.00	10.93	[-0.65; 0.23]
0.005	0.00	0.01	[-0.06; 0.07]	48.64	0.00	95.05	0.00	14.73	[-0.67; 0.54]
0.005	0.00	0.27	[0.16; 0.38]	46.98	0.25	94.74	0.02	14.91	[-0.78; 1.29]
0.005	0.00	1.39	[1.11; 1.68]	31.07	1.28	92.82	0.10	5.85	[-0.86; 3.45]
0.005	0.00	-0.27	[-0.36; -0.19]	48.71	-0.28	93.74	0.00	10.96	[-0.67; 0.07]
0.005	0.00	-0.06	[-0.29; 0.16]	48.09	-0.06	94.86	0.00	24.82	[-0.67; 0.40]
0.005	0.01	0.16	[-0.21; 0.53]	47.70	0.14	94.40	0.01	34.34	[-0.77; 1.01]
0.005	0.01	0.51	[-0.09; 1.12]	42.06	0.47	93.94	0.05	38.62	[-1.02; 2.00]
0.005	0.09	2.03	[0.44; 3.62]	14.66	1.77	88.20	0.18	62.14	[-1.05; 4.60]

Note: $q_{0.025}$ = 0.025th quantile of the uniform distribution. $q_{0.975}$ = 0.975th quantile of the uniform distribution. Conv. runs = proportion of converged runs. Co. rate θ = Coverage rate of the Wald CI interval for the parameter θ.
*Values reported in %.
**Values reported in ‰.

4 Discussion

In this paper, we proposed an extension of the PL approach to the frameworks of treatment effect heterogeneity. This approach has been shown to perform very well for meta-analyzing a homogeneous treatment effect, providing unbiased estimates and nominal coverage rates, even in scenarios with extremely rare events[13]. However, the assumption of a homogeneous treatment effect is often too restrictive[33]. Therefore, the objective of this paper was to assess the performance of the extended PL approach in settings with treatment effect heterogeneity and rare events.

Overall, we found that the extended PL approach provided good results (i.e. low biases and nominal coverage rates) for the estimation of the mean OR, RR, and RD, even when meta-analyzing extremely rare events. Moreover, the use of a PI to quantify the degree of between-study heterogeneity was found to be more reliable than estimation of τ^2.

Estimates of the mean treatment effect were virtually unbiased in most of the scenarios considered and whatever the scale chosen. Moreover, the Wald CIs obtained for this

parameter provided coverage rates that were close to nominal. The only settings where the performance of the PL method somewhat deteriorated for this parameter were scenarios with extremely rare events and large heterogeneity. For the RD, this was particularly the case with large positive mean treatment effect.

Although providing a good estimate of the mean treatment effect is a desirable feature, this is not sufficient in the presence of treatment effect heterogeneity and the between-study variance should also be estimated. However, our simulations corroborated previous findings that it is difficult to obtain a good estimate of this parameter (it is very often underestimated), especially with rare events[19,34]. Therefore, assessing the degree of heterogeneity based only on this parameter's estimate can lead to dubious conclusions.

Fortunately, the computation of a PI allows one to alleviate – if not solve – this issue. Indeed, in most of the settings studied, our simulations have shown that when the heterogeneity was not too asymmetric, the computed PIs were conservative and contained the 2.5th and 97.5th quantiles of the ES's distribution. This is in contrast with previous simulations studies that found suboptimal coverage under certain circumstances[27,29]. However, this can be explained by the two-stage approach adopted by these authors. In this study, we used a one-stage approach (i.e. all model's parameters are estimated simultaneously), which, as shown, performed much better. Notice that formula (4) includes both the between and the within-study variances. With rare events, the within-study component is likely to dominate the between-study component and large values of the former to compensate underestimation of the latter.

The specificity of the PL approach is that it is based on working distributions; we are neither making the assumption that the outcome model is the correct one, nor that the baseline prevalence and the treatment effect are genuinely normally distributed. Actually, in models (2) and (3), the Poisson distribution in the former and the normal distribution in the latter

served as an approximation for the true binomial distribution, whereas the bivariate normal distribution serves as an approximation for the uniform distribution used to simulate the data. Therefore, it was important to assess the robustness of this model. In additional simulations, we considered various ES's distributions (normal and gamma; results not shown). When using the normal distribution, results were virtually identical to those obtained with the uniform distribution, whereas results obtained with the gamma distribution (with scale parameter 1 and shape parameter 0.1, 0.25, and 0.5) slightly deteriorated the PI's performance. The more asymmetrical the underlying distribution of the ES, the more the 0.975^{th} quantile was underestimated. These findings lead us to the conclusion that the PL approach is robust to the true ES's distribution, as long as this distribution is not too asymmetric.

In order to obtain a sufficient number of converged runs, we had to simplify models (1) – (3) by assuming a zero correlation between ψ_k and μ_k. As such this amounts to imposing a non-smaller variability of the event probability in the treatment group than in the control group, which is a defensible assumption. For instance, Bhaumik and colleagues adopted assumption[20]. Another way to proceed is to recode the treatment dummy (i.e. T_{kj}) 1 as 1/2 and 0 as -1/2, which implies an equal variability in the two groups[35]. If one is not willing to make any assumption regarding how the variability in the two treatment arms compare, then one should use models (1) – (3) without constraining the correlation to zero. However, the price is an increased risk of running into a numerical issue.

Based on all these considerations, when the treatment effect is expected to be heterogeneous but not too asymmetrically distributed, we recommend the use of the PL approach for the meta-analysis of rare events, whatever the ES considered (i.e. OR, RR, or RD). Moreover, we urge researchers to always report the PI (and not simply the mean and between-study

variance) when conducting a MA under the assumption of a heterogeneous treatment effect. Indeed, although the range covered by the PI was often too wide with rare events, this can serve as an interesting benchmark (i.e. if a value does not fall within the computed PI, it is very unlikely that this value will be observed for a new study).

Funding

The work of RpR was supported by the Swiss National Science Foundation (grant number: 200021_169504/1).

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

References

[1] Sweeting MJ., Sutton AJ., Lambert, PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*. 2004; **23**: 1351-1375.

[2] Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*. 2007; **26**: 53-77.

[3] Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions ? *Biometrical Journal*. 2018; **60**: 1040-1058.

[4] Shuster, JJ. Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine*. 2009; **29**: 1259-1265.

[5] Vabdermeer B, Bialy L, Hooton N, Hartling L, Klassen TP, Johnston BC, Wiebe N. Meta-analyses of safety data: a comparison of exact versus asymptotic methods. *Statistical Methods in Medical Research*. 2009; **18(4)**: 421-432.

[6] Nietert PJ, Wahlquist AE, Herbert TL. Characteristics of recent biostatistical methods adopted by researchers publishing in general/internal medicine journals. *Statistics in Medicine*. 2013; **32**(1): 1-10.

[7] Piaget-Rossel R, Taffé P. Meta-analysis of rare events under the assumption of a homogeneous treatment effect. *Biometrical Journal*. 2019; 1-18.

[8] Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*. 2000; **19(24)**: 3417-3432.

[9] Marschner IC, Gillett AC. Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics*. 2012; **13(1)**: 179-192.

[10] Fitzmaurice GM, Lipsitz SR, Arriage A, Sinha D, Geenberg C, Gawande AA. Almost efficient estimation of relative risk regression. *Biostatistics*. 2014; **15(4)**: 745-756.

[11] Carter RE, Lipsitz SR, Tilley BC. Quasi-likelihood estimation for relative risk regression models. *Biostatistics*. 2005; **6(1)**: 39-44.

[12] Dwivedi AK, Indika M, Lee S, Tarwater P. Methods for estimating relative risk in studies of common binary outcomes. *Journal of Applied Statistics*. 2014; **3**: 484-500.

[13] Piaget-Rossel R, Taffé P. A pseudo-likelihood approach for the meta-analysis of homogeneous treatment effect: exploiting the information contained in double-zero and single-arm studies. *Journal of Statistics: Advances in Theory and Applications*. 2019; **21(2)**: 91-117.

[14] Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework oft he generalized linera mied model with applications in sparse data. *Statistics in Medicine.* 2010; **29**: 3046-3067.

[15] Cai T, Parast L, Ryan L. Meta-analysis for rare events. *Statistics in Medicine*. 2010; **29**: 2078-2089.

[16] Lane PW. Meta-analysis of incidence of rare events. *Statistical Methods in Medical Research*. 2011; **22(2)**: 117-132.

[17] Amatya A, Bhaumik DK, Normand SL, Geenhouse J, Kaizar E, Neelon B, Gibbons RD. Likelihood-based random-effect meta-analysis of binary events. *Journal of Biopharmaceutical Statistics*. 2015; **25**: 984-1004.

[18] Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine*. 2018; **37**: 1059-1085.

[19] Bakbergenuly I, Kulinskaya E. Meta-analysis of binary outcomes via generalized linear mixed models: a simulation study. *BMC Medical Research Methodology*. 2018; **18**: 70.

[20] Bhaumik DK, Amatya A, Normand SLT, Greenhouse J, Kaizar E, Neelon B, Gibbons RD. Meta-analysis of rare binary adverse event data. *Journal of American Statistical Association*. 2012; **107**: 555-567.

[21] Tian L, Cai T, Pfeffer MA, Pinakov N, Cremieux PY, Wei LJ. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2x2 tables with all available data but without artificial continuity correction. *Biostatistics*. 2009; **10(2)**: 275-281.

[22] Richardson SR, Robins JM, Wang L. On modeling and estimation for the relative risk and risk difference. *Journal of the American statistical association*. 2017; **112(519)**: 1121-1130.

[23] IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016; **6**: e010247.

[24] Kuss O. Statistical methods for meta-analyses including information from studies without any events – add nothing to nothing and succeed nevertheless. *Statistics in Medicine*. 2015; **34**: 1097-1116.

[25] Guolo A, Varin C. Random-effects meta-analysis: the number of studies matters. *Statistical Methods in Medical Research*. 2017; **26(3)**: 1500-1518.

[26] Higgins JPT, Thompson SG, Spiegelhalter J. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society A*. 2009;**172**:137-159.

[27] Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. *Statistics in Medicine* 2016; **36**: 301-317.

[28] Wang CC, Lee WC. A simple method to estimate prediction intervals and predictive distributions: summarizing meta-analyses beyond means and confidence intervals. *Research Synthesis Methods*. 2018; **10**: 255-266.

[29] Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. *Statistical Methods in Medical Research* 2019; **28(6)**: 1689-1702.

[30] Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Statistics in Medicine*. 2017; **36(5)**: 855-75.

[31] Niël-Weise BS, Stijnen T, van den Broek PJ. Anti-infective-treated central venous catheters: a systematic review of randomized controlled trials. *Intensive Care Medicine*. 2007; **33(12)**: 2058-2068.

[32] White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982; **50**: 1-26.

[33] Kontopantelis E, Springate DA, Reeves D. A re-analysis of the Cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *Plos One*. 2013;**8**:e69930.

[34] Piaget-Rossel R, Taffé P. Meta-analysis of incidence rate data in the presence of zero-event and single-arm studies. *International Journal of Statistics in Medical Research*. 2019; **8**: 57-66.

[35] Turner RM, Omar RZ, Yang M, et al. A multilevel model framework for meta-analysis of clinical trials with binary outomces. *Statistics in Medicine*. 2000; **19(24)**: 3417-3432.

Appendix

Table A1. Proportions of treatment arms reporting zero event in the 45 scenarios considered

	$M_{\pi_c} = 0.1$		$M_{\pi_c} = 0.01$		$M_{\pi_c} = 0.005$	
	CA = 0	TA = 0	CA = 0	TA = 0	CA = 0	TA = 0
$\tau^2_{\log(OR)} = 0$						
$\mu_{\log(OR)} = -1.5$	0.07	11.75	38.89	80.05	61.73	89.80
$\mu_{\log(OR)} = -0.5$	0.12	0.66	39.09	55.66	61.29	75.03
$\mu_{\log(OR)} = 0$	0.14	0.09	39.10	38.60	61.08	60.82
$\mu_{\log(OR)} = 0.5$	0.11	0.00	38.08	22.05	61.77	45.20
$\mu_{\log(OR)} = 1.5$	0.12	0.00	39.23	2.58	61.33	13.33
$\tau^2_{\log(OR)} = 0.1$						
$\mu_{\log(OR)} = -1.5$	0.13	12.79	39.29	79.91	61.06	89.28
$\mu_{\log(OR)} = -0.5$	0.08	0.95	38.16	54.86	61.59	73.04
$\mu_{\log(OR)} = 0$	0.10	0.17	38.34	38.99	61.46	60.98
$\mu_{\log(OR)} = 0.5$	0.08	0.02	38.89	22.80	61.86	45.54
$\mu_{\log(OR)} = 1.5$	0.05	0.00	38.67	3.78	60.91	15.53
$\tau^2_{\log(OR)} = 0.5$						
$\mu_{\log(OR)} = -1.5$	0.08	16.88	38.10	76.83	61.90	87.30
$\mu_{\log(OR)} = -0.5$	0.04	3.29	38.91	53.10	61.51	70.13
$\mu_{\log(OR)} = 0$	0.09	0.74	39.07	38.20	62.11	58.14
$\mu_{\log(OR)} = 0.5$	0.09	0.20	38.18	25.39	61.37	44.11
$\mu_{\log(OR)} = 1.5$	0.09	0.00	38.03	6.81	61.08	18.59

Note: "CA = 0" = average proportion of control arms with zero event; "TA = 0" = average proportion of treatment arms reporting zero event. Results based on 10,000 MA generated, with 20 primary studies each.

Meta-Analysis of Incidence Rate Data in the Presence of Zero-Event and Single-Arm Studies

Romain Piaget-Rossel and Patrick Taffé*

University of Lausanne, Center for Primary Care and Public Health (Unisanté), Division of Biostatistics, Route de la Corniche 10, 1010 Lausanne, Switzerland

Abstract: Unlike the classical two-stage DerSimonian and Laird meta-analysis method, the one-stage random-effects Poisson and Negative-binomial models have the great advantage of including the information contained in studies reporting zero event in one or both arms and in studies with one missing arm. Since the Negative-binomial distribution relaxes the assumption of equi-dispersion made by the Poisson, it should perform better when data exhibit over-dispersion. However, the superiority of the Negative-binomial model with rare events and single-arm studies is unclear and needs to be investigated. Moreover, to the best of our knowledge, this model has never been investigated in the context of a meta-analysis of incidence rate data with heterogeneous intervention effect. Therefore, we assessed the performance of the univariate and bivariate random-effects Poison and Negative-binomial models using simulations calibrated on a real dataset from a study on the surgical management of phyllodes tumors. Results suggested that the bivariate random-effects Negative-binomial model should be favored for the meta-analysis of incidence rate data exhibiting over-dispersion, even in the presence of zero-event and single-arm studies.

Keywords: Incidence rate, Meta-analysis, Negative-binomial model, Poisson model, Rare events, Random effects.

1. INTRODUCTION

Meta-analysis is considered as the gold standard of evidence-based medicine [1]. By combining the results of related but independent studies, it allows to evaluate the effect of a treatment (or an intervention; here-after we will use this latter terminology to emphasize the fact that we are in an observational framework) in situations where primary studies taken separately would not have sufficient power to detect a statistically significant effect [2]. Meta-analyses are thus particularly useful when studying rare events. For example, Nissen and Wolski studied the impact of a diabetes drug on the incidence of myocardial infarctions and cardiovascular deaths [3], whereas Niël-Weise, Stijnen, and van den Broek conducted a meta-analysis on the effect of anti-infective-treated central venous catheters on the incidence of catheter-related bloodstream infections [4].

When the data at hand are counts of events over time, the effect size (ES) of interest is often the incidence rate (IR) and different intervention arms can be contrasted using the incidence rate ratio or the incidence rate difference. In this setting, the most commonly-used approach, which can be applied in both fixed-effect (FE) and random-effects (RE) frameworks, consists in computing a weighted average of the primary study ESs with weights proportional to the inverse of each ES's variance [5] (the so-called "two-stage" approach [6]). Although this approach is very popular and enjoys good asymptotic properties, its use is problematic in small/finite samples, especially with rare events, when some studies report no event in one or both arms, and when some studies' arms are missing. Indeed, the ES and/or the weight computed in a single-zero (SZ), double-zero (DZ), or single-arm (SA) study are indefinite. To deal with SZ and DZ studies, researchers sometimes use a continuity correction factor [7]. However, this method is flawed and suffers from several criticisms [8-10]. In addition, SA studies are still excluded from the meta-analysis.

Under the assumption of a homogeneous intervention effect, Mantel-Haenszel (MH) is an alternative to the classical inverse variance method and has been shown to be very performant, even with very rare events [11]. Unlike the inverse variance method, the MH method can cope with SZ studies. However, DZ studies are simply discarded with that method and, therefore, do not contribute to the ES estimate. Similarly, the MH method fails to include the information contained in SA studies. Piaget-Rossel and Taffé have shown that only the exclusion of SA studies impacted the performance of this method (i.e. a loss of precision was observed in settings with a large proportion of SA studies) [12]. Another limitation of the MH method is that it is only valid under the assumption of a homogeneous intervention effect [13].

To improve these simple methods, one-stage or exact methods based on the likelihood principle have been developed. Such methods use the information contained in all the studies (i.e. including SZ, DZ, and SA studies) and allow for the inclusion of covariates. A natural way to model IR data is to use a Poisson likelihood [14]. This model can be adapted to the setting of a heterogeneous intervention effect by introducing random effects [15-16] and can be used to model the IR using either a univariate or a bivariate modelling approach [17]. One important limitation of the

*Address correspondence to this author at the University of Lausanne, Center for Primary Care and Public Health (Unisanté), Division of Biostatistics, Route de la Corniche 10, 1010 Lausanne, Switzerland; Tel: +41 (0)21 314 72 56; E-mail: patrick.taffe@unisante.ch

E-ISSN: 1929-6029/19

Poisson model is its reliance on the equi-dispersion assumption (i.e. the mean of the distribution is equal to its variance), which rarely holds with count data (because of unmeasured individual characteristics differing within studies for instance). A way to relax the equi-dispersion assumption is to replace the Poisson distribution by the Negative-binomial [18]. Although we found some applications of the Negative-binomial model for the meta-analysis of individual patient data in a two-stage approach or in the context of a homogeneous intervention effect [19-20], we are not aware of the use of the random-effects Negative-binomial (Re-NB) model for the meta-analysis of incidence rate data within a framework of a heterogeneous intervention effect, especially with rare events and SA studies.

Therefore, the goal of this paper was to assess the appropriateness of the Re-NB model for the meta-analysis of IR data in the presence of SZ, DZ, and even SA studies. Using simulations calibrated on a real clinical dataset, we compared this model with the random-effects Poisson (Re-Poi) model. We considered both univariate and bivariate versions of these two models. The data we used came from a recent systematic review on the impact of the width of the resection's margin on the rate of local recurrences in phyllodes tumors [21]. The use of a model allowing for over-dispersion seemed particularly adapted to this example where the exposure (i.e. the width of the resection's margin) had not been randomized and accounting for patient-level covariates affecting the incidence rate of recurrences at the analysis stage was difficult. We restricted our analyses to the framework of a heterogeneous intervention effect because the assumption of a homogeneous intervention effect was not plausible for the example considered.

In the remaining of this paper, we start by describing the illustrative example. Then, Section 3 presents the different models under investigation. In Section 4, we illustrate these models using data from the illustrative example and present results from a simulation study. Finally, Section 5 contains the discussion and some concluding remarks.

2. ILLUSTRATIVE EXAMPLE

The dataset used in this paper came from a systematic review on the surgical management of phyllodes tumors [21]. The author conducted a

Table 1: Data Extract from the Study on Surgical Management of Phyllodes Tumors

Study	Tumor's type	Control arm (i.e. margin < 10mm)			Intervention arm (i.e. margin ≥ 10mm)		
		n	t	Y	n	t	Y
1	Benign	14	1020.6	3	8	583.2	0
1	Borderline	4	291.6	2	11	801.9	4
1	Malignant	4	291.6	3	5	364.5	2
2	Benign	-	-	-	7	522.2	0
2	Malignant	-	-	-	3	98.1	0
3	Benign	104	10712	4	30	3090	1
3	Borderline	34	2856	2	23	1932	0
4	Benign	56	3976	7	44	3124	6
4	Borderline	1	71	0	3	213	0
4	Malignant	1	71	0	3	213	0
5	Malignant	10	1399	6	14	1958.6	4
6	Benign	126	9450	5	14	1050	0
6	Borderline	19	1121	4	13	767	1
6	Malignant	1	15	0	9	135	5
7	Benign	16	665.6	4	18	748.8	0
7	Borderline	1	57	0	2	114	0
7	Malignant	1	45	1	2	90	0
8	Malignant	6	726	0	30	3630	0
9	Benign	53	3074	0	5	290	0
9	Borderline	5	290	1	2	116	0
10	Benign	-	-	-	179	6748.3	12
10	Borderline	-	-	-	43	1406.1	3
10	Malignant	-	-	-	32	979.2	1

Note: n = sample size; t = person-months (number of patients × mean follow-up); Y = number of recurrences.

systematic review to assess the impact of the width of the resection's margin on the rate of tumor's recurrences (Table **1**).

The dataset entails 10 primary studies on tumors patients who underwent a surgical intervention to remove their tumor. Each tumor was classified as either benign, borderline, or malignant. Two arms were defined according to the margin of resection used during the surgery: intervention arm included patients with a margin above or equal to 10mm, and patients whose resection's margin was below 10mm belonged to the control arm. Study 2 and 10 corresponded to SA studies as they reported results only for margins above ten millimeters. One third of the control arms and more than half of the intervention arms reported zero event. Sample sizes and person-months varied widely across the studies.

3. MODELS TO COMBINE INCIDENCE RATES

3.1. The Random-Effects Poisson Model

3.1.1. Univariate Modelling

Let Y_{ijk} be the number of events occurring in study *i* ($i = 1, \ldots, 10$), type of tumors *j* ($j \in$ {benign, borderline, malignant}) and arm *k* (*k* = *C* for control and *I* for intervention). Assume that the number of events is conditionally distributed as a Poisson variable with mean $\lambda_{ijk} = \mu_{ijk} * t_{ijk}$, where μ_{ijk} denotes the incidence rate and t_{ijk} the person-time. Consider the following univariate Re-Poi model:

$$Y_{ijk}|\lambda_{ijk} \sim Poisson(\lambda_{ijk})$$

$$\lambda_{ijk} = \mu_{ijk} * t_{ijk}$$

$$\log(\mu_{ijk}) = \beta_{0i} + \beta_{1i} * I_{ijk} + \gamma' * X_{ijk} + \theta' * Z_{ijk} * I_{ijk}$$

$$\begin{pmatrix}\beta_{0i}\\ \beta_{1i}\end{pmatrix} \sim N[\beta, \Omega], \beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \Omega = \begin{pmatrix}\sigma^2_{\beta_0} & 0\\ 0 & \sigma^2_{\beta_1}\end{pmatrix}$$

where I_{ijk} is an indicator variable taking the value *0.5* if *k* = *I* and *-0.5* if *k* = *C*, X_{ijk} is a vector of covariate affecting the baseline incidence rate, Z_{ijk} is a vector of covariates affecting the intervention effect. For the sake of clarity, we have separated the covariates affecting the baseline incidence rate from those affecting the intervention effect. Notice, however, that X_{ijk} and Z_{ijk} may contain the same covariates. In this model, one makes the assumption that the residual variance of the log(IR) is the same in the control and intervention groups [22]. Observe that $\beta_{0i} - \frac{1}{2}\beta_{1i}$ represents the residual log(IR) in the control group and $\beta_{0i} + \frac{1}{2}\beta_{1i}$ the residual log(IR) in the intervention group in study *i*. Therefore, β_{1i} is the residual log(IRR) in study *i* and β_1 the mean residual log(IRR) across the 10 studies. The vector of parameters γ measures the change in baseline log(IR) associated with a one-unit change of X_{ijk}, whereas θ allows one to account for a differential effect of the intervention according to the covariates contained in Z_{ijk}. Finally, $\sigma^2_{\beta_0}$ captures the residual baseline log(IR) heterogeneity, whereas $\sigma^2_{\beta_1}$ measures the residual heterogeneity of the intervention effect.

The likelihood function writes:

$$\prod_{i=1}^{10} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[\left(\prod_{j \in D} \prod_{k \in \{C,I\}} P(Y_{ijk}|\lambda_{ijk}, \beta_{0i}, \beta_{1i}) \right) \phi(\beta_{0i}, \beta_{1i}|\, \beta, \Omega) \right] d\beta_{0i} d\beta_{1i}$$

where *D* = {benign, borderline, malignant}, $P(Y_{ijk}|\lambda_{ijk}, \beta_{0i}, \beta_{1i})$ is the Poisson density with mean $\lambda_{ijk} = \mu_{ijk} * t_{ijk}$, and $\phi(\beta_{0i}, \beta_{1i}|\, \beta, \Omega)$ is the bivariate normal density with mean β and variance-covariance matrix Ω.

3.1.2. Bivariate Modelling

In the bivariate approach, the log incidence rates of events are modelled separately for the intervention and control arms and the variances of the residuals log(IR) for the control and intervention groups are allowed to be different. Therefore, the bivariate Re-Poi model is given by:

$$Y_{ijC}|\lambda_{ijC} \sim Poisson(\lambda_{ijC})$$

$$\lambda_{ijC} = \mu_{ijC} * t_{ijC}$$

$$\log(\mu_{ijC}) = \alpha_{iC} + \gamma' * X_{ijC} + \theta'_C * Z_{ijC}$$

$$Y_{ijI}|\lambda_{ijI} \sim Poisson(\lambda_{ijI})$$

$$\lambda_{ijI} = \mu_{ijI} * t_{ijI}$$

$$\log(\mu_{ijI}) = \alpha_{iI} + \gamma' * X_{ijI} + \theta'_I * Z_{ijI}$$

$$\begin{pmatrix}\alpha_{iC}\\ \alpha_{iI}\end{pmatrix} \sim N[\alpha, \Omega], \alpha = \begin{pmatrix}\alpha_C\\ \alpha_I\end{pmatrix}, \Omega = \begin{pmatrix}\sigma^2_C & \sigma_{CI}\\ \sigma_{CI} & \sigma^2_I\end{pmatrix}$$

where X_{ijI} and Z_{ijI} are defined as in the univariate Re-Poi model. Note that α_{iC} represents the residual log(IR) in the control group of study *i* and α_{iI} the residual log(IR) in the intervention group of study *i*. Therefore, $\alpha_{iI} - \alpha_{iC}$ is the residual log(IRR) in study *i* and $\alpha_I - \alpha_C$ the mean residual log(IRR) across the studies. Again, X_{ijk} contains the covariates affecting the baseline incidence, whereas Z_{ijk} contains those affecting the intervention effect. This bivariate model is more flexible than the univariate since it allows estimating two distinct variance parameters (σ^2_C, σ^2_I). Moreover, the covariance (σ_{CI}) links the two processes.

The likelihood is given by

$$\prod_{i=1}^{10} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[\left(\prod_{j \in D} \prod_{k \in \{C,I\}} P(Y_{ijk}|\lambda_{ijk}, \alpha_{iC}, \alpha_{iI}) \right) \phi(\alpha_{iC}, \alpha_{iI}|\, \alpha, \Omega) \right] d\alpha_{iC} d\alpha_{iI}$$

where $P(Y_{ijk}|\lambda_{ijk}, \alpha_{iC}, \alpha_{iI})$ is the Poisson density, $\phi(\alpha_{iC}, \alpha_{iI}|\, \alpha, \Omega)$ denotes the bivariate normal density with mean α and variance-covariance matrix Ω.

3.2. The Random-Effects Negative-Binomial Model

3.2.1. Univariate Modelling

To obtain a Negative-binomial model [18], one can introduce over-dispersion into the Re-Poi model by including a random multiplicative coefficient ν_i in the expression of the mean/variance parameter: $\lambda_{ijk} = \mu_{ijk} * t_{ijk} * \nu_i$, where $\nu_i \sim \text{Gamma}\left(\frac{1}{\eta}, \eta\right)$ (i.e. a one parameter Gamma distribution with unit mean and variance η, $\eta > 0$). With $Y_{ijk}|\lambda_{ijk} \sim Poisson(\lambda_{ijk})$, it can be shown that the marginal expectation $E(Y_{ijk}) = \lambda_{ijk}$ and marginal variance $Var(Y_{ijk}) = \lambda_{ijk}(1 + \eta\lambda_{ijk})$, thereby allowing the variance to differ from the mean. Consequently, the univariate Re-NB model is given by:

$$Y_{ijk}|\lambda_{ijk} \sim Poisson(\lambda_{ijk})$$

$$\lambda_{ijk} = \mu_{ijk} * t_{ijk} * \nu_i$$

$$\log(\mu_{ijk}) = \beta_{0i} + \beta_{1i} * I_{ijk} + \gamma' * X_{ijk} + \theta' * Z_{ijk} * I_{ijk}$$

$$\begin{pmatrix}\beta_{0i}\\ \beta_{1i}\end{pmatrix} \sim N[\beta, \Omega], \beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \Omega = \begin{pmatrix}\sigma^2_{\beta_0} & 0\\ 0 & \sigma^2_{\beta_1}\end{pmatrix}, \nu_i \sim \Gamma\left(\frac{1}{\eta}, \eta\right)$$

and the likelihood writes:

$$\prod_{i=1}^{10} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left\{ \left[\prod_{j \in D} \prod_{k \in \{C,I\}} \int_{-\infty}^{+\infty} \left(P(Y_{ijk}|\lambda_{ijk}, \beta_{0i}, \beta_{1i}, \nu_i) \right. \right. \right.$$
$$\left. \left. \left. \Gamma(\nu_i|1/\eta, \eta) \right) d\nu_i \right] \phi(\beta_{0i}, \beta_{1i}|\beta, \Omega) \right\} d\beta_{0i} d\beta_{1i}$$

where $P(Y_{ijk}|\lambda_{ijk}, \beta_{0i}, \beta_{1i}, \nu_i)$ is the Poisson density, $\Gamma(\nu_i|1/\eta, \eta)$ is the Gamma density with mean 1 and variance η, $\phi(\beta_{0i}, \beta_{1i}|\beta, \Omega)$ is the bivariate Normal density with mean β and variance-covariance matrix Ω.

3.2.2. Bivariate Modelling

The bivariate Re-NB model is similar in structure to the bivariate Re-Poi model, except for the addition of the over-dispersion terms ν_{iC} and ν_{iI}:

$$Y_{ijC}|\lambda_{ijC} \sim Poisson(\lambda_{ijC})$$

$$\lambda_{ijC} = \mu_{ijC} * t_{ijC} * \nu_{iC}$$

$$\log(\mu_{ijC}) = \alpha_{iC} + \gamma' * X_{ijC} + \theta'_C * Z_{ijC}$$

$$Y_{ijI}|\lambda_{ijI} \sim Poisson(\lambda_{ijI})$$

$$\lambda_{ijT} = \mu_{ijI} * t_{ijI} * \nu_{iI}$$

$$\log(\mu_{ijI}) = \alpha_{iI} + \gamma' * X_{ijI} + \theta'_I * Z_{ijI}$$

$$\begin{pmatrix}\alpha_{iC}\\ \alpha_{iI}\end{pmatrix} \sim N[\alpha, \Omega], \alpha = \begin{pmatrix}\alpha_C\\ \alpha_I\end{pmatrix}, \Omega = \begin{pmatrix}\sigma^2_C & \sigma_{CI}\\ \sigma_{CI} & \sigma^2_I\end{pmatrix}$$

$$\nu_{iC} \sim \Gamma\left(\frac{1}{\eta_C}, \eta_C\right), \nu_{iI} \sim \Gamma\left(\frac{1}{\eta_I}, \eta_I\right)$$

The likelihood is given by:

$$\prod_{i=1}^{10} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left\{ \left[\prod_{j \in D} \prod_{k \in \{C,I\}} \int_{-\infty}^{+\infty} \left(P(Y_{ijk}|\lambda_{ijk}, \alpha_{iC}, \alpha_{iT}, \nu_{ik}) \right. \right. \right.$$
$$\left. \left. \left. \Gamma(\nu_{ik}|1/\eta_k, \eta_k) \right) d\nu_{ik} \right] \phi(\alpha_{iC}, \alpha_{iI}|\alpha, \Omega) \right\} d\alpha_{iC} d\alpha_{iI}$$

where $P(Y_{ijk}|\lambda_{ijk}, \alpha_{iC}, \alpha_{iT}, \nu_{ik})$ is the Poisson density, $\Gamma(\nu_{ik}|1/\eta_k, \eta_k)$ is the Gamma density with mean 1 and variance η_k ($k = I, C$), $\phi(\alpha_{iC}, \alpha_{iI}|\alpha, \Omega)$ is the bivariate Normal density with mean α and variance-covariance matrix Ω. Close inspection of the likelihood function reveals that the bivariate Re-NB model allows not only two distinct variances to be estimated (i.e. one for each arm), but also two distinct over-dispersion parameters (η_C, η_I), which makes this model more flexible than the more commonly-used bivariate Re-Poi model.

4. NUMERICAL ANALYSES

All the numerical analyses were conducted using Stata/IC 15.1 [23]. We used the command mepoisson to fit both Re-Poi models, menbreg to fit the univariate Re-NB model, and gsem to fit the bivariate Re-NB model. To integrate the likelihood, we used the mean-variance adaptive Gauss-Hermite quadrature method with seven integration points (the default implementation in Stata). We set the maximum number of iterations at 1001.

4.1. Specifications of the Log Incidence Rate

For the univariate Re-Poi and the univariate Re-NB models, we considered the following log(IR) specification:

$$\log(\mu_{ijk}) = \beta_{0i} + \beta_{1i} * I_{ijk} + \beta_2 * M_{ijk} + \beta_3 * I_{ijk} * M_{ijk}$$

where M_{ijk} is an indicator for malignant tumors taking the value 1 for malignant and 0 otherwise and I_{ijk} is defined as in Section 3. The focus was on estimating the mean intervention effect for non-malignant tumors ($E(\beta_{1i}) \equiv \log IRR$), the residual heterogeneity of the intervention effect ($Var(\beta_{1i}) \equiv \sigma^2_{\log IRR}$), and the difference in the mean intervention effects between malignant and non-malignant tumors ($\beta_3 \equiv \Delta_{\log IRR}$).

For the bivariate models, we considered the following specifications for the log(IR):

$$\log(\mu_{ijC}) = \alpha_{iC} + \gamma_C * M_{ijC}$$

$$\log(\mu_{ijT}) = \alpha_{iT} + \gamma_T * M_{ijT}$$

Again, parameters of interest were the mean intervention effect for non-malignant tumors ($E(\alpha_{it} -$

$\alpha_{ic}) \equiv \log IRR$), the residual heterogeneity of the intervention effect ($Var(\alpha_{it} - \alpha_{ic}) \equiv \sigma^2_{\log IRR}$) and the difference in the mean intervention effect between malignant and non-malignant tumors ($\gamma_T - \gamma_C \equiv \Delta_{\log IRR}$).

4.2. Application to the Surgical Management of Phyllodes Tumors

Table **2** reports the results obtained by fitting the four models described in Section 3 to the data presented in Table **1**. Changing from the univariate to the bivariate framework had a smaller effect on the estimates obtained by the Re-Poi model than on those obtained by the Re-NB model.

For each model, the estimated effect of a margin of resection above 10mm was a reduction of the rate of local recurrence for non-malignant tumors (i.e. all estimates of $\log IRR$ were below zero). This reduction was larger for the Re-NB models, with the largest reduction estimated by the univariate model. Estimations of the residual heterogeneity of the intervention effect by both Re-Poi models were around 0.4, whereas the Re-NB models provided values close to zero for this parameter (the univariate Re-NB model even found an absence of residual heterogeneity).

With the two Re-Poi models, estimates of the impact of the margin was greater for malignant tumors (i.e. $\widehat{\Delta}_{\log IRR} < 0$), whereas with the Re-NB models the opposite result was obtained. While with the univariate Re-NB model it was still found that the margin above 10mm reduced the rate of local recurrences of malignant tumors, this was not the case anymore with the bivariate model (i.e. the estimated coefficients of $\log IRR$ and $\Delta_{\log IRR}$ cancelled each other out).

4.3. Simulations

To identify the best fitting model from Section 3, we conducted Monte-Carlo simulations that were calibrated in order to mimic the data from the studies selected in the systematic review on phyllodes tumors (Table **1**). The number of events were generated according to the bivariate Re-NB model described in Subsection 3.2.2. We set the parameters of this model in order to investigate four different scenarios (see Table **3** below). The first scenario used values closed to the estimates obtained when fitting the model on the example dataset. This was our baseline scenario, from which we derived the three others. Scenario 2 corresponded to a Poisson framework with no over-dispersion, scenario 3 was devised to study the impact of having a large amount of residual heterogeneity of the intervention effect, whereas the last scenario investigated the situation with no mean intervention effect.

For each scenario, we simulated $N = 1000$ datasets. Performance of each model was assessed by the median relative bias (i.e. relative difference between the median estimate and the true parameter's value), and coverage rate and median width of the 95% Wald confidence intervals (CIs) obtained for the three

Table 2: Estimation of the Three Parameters of Interest in the Illustrative Example Dataset

Model	$\log IRR$	$\sigma^2_{\log IRR}$	$\Delta_{\log IRR}$
Univariate Re-Poi	-0.27 (-1.15; 0.61)	0.40 (0.02 ; 7.35)	-0.33 (-1.58 ; 0.91)
Bivariate Re-Poi	-0.33 (-1.36 ; 0.70)	0.45 (0.02 ; 8.25)	-0.32 (-1.57 ; 0.92)
Univariate Re-NB	-0.58 (-1.66 ; 0.50)	0 (.)	0.24 (-1.70 ; 2.18)
Bivariate Re-NB	-0.48 (-1.61 ; 0.64)	0.02 (0.01 ; 0.03)	0.48 (-1.52 ; 2.48)

Note: 95% Wald confidence intervals are provided between parentheses. $\log IRR$ = mean intervention effect for the non-malignant tumors (i.e. impact of margin ≥ 10mm vs margin < 10mm); $\sigma^2_{\log IRR}$ = residual heterogeneity of the intervention effect; $\Delta_{\log IRR}$ = difference in the mean intervention effect between malignant and non-malignant tumors.

Table 3: Value of the Different Parameters under the Four Simulated Scenarios

Scenarios	η_C	η_I	α_C	α_I	γ_C	γ_I	σ^2_C	σ^2_I	σ_{CI}
1) Baseline	0.45	1.65	-6.5	-7	0.9	1.4	0.40	0.25	0.30
2) No over-dispersion	**0**	**0**	-6.5	-7	0.9	1.4	0.40	0.25	0.30
3) Large residual heterogeneity	0.45	1.65	-6.5	-7	0.9	1.4	**1.60**	**1.00**	0.30
4) No mean intervention effect	0.45	1.65	-6.5	**-6.5**	0.9	**0.9**	0.40	0.25	0.30

Note: η_k, α_k, γ_k, σ^2_k and σ_{CI}, for k = C, I, correspond to the parameters of the bivariate Re-NB model described in Subsection 3.2.2. Values in bold represent changes compared to the baseline scenario.

parameters of interest (i.e. the mean intervention effect for non-malignant tumors $\log IRR$, the residual heterogeneity of the intervention effect $\sigma^2_{\log IRR}$, and the difference in mean intervention effect between malignant vs non-malignant tumors $\Delta_{\log IRR}$). We decided to compute the median instead of the mean for the bias and CI's width to avoid the influence of exceedingly large or small values obtained in some simulations. Since the numerical algorithm used to estimate the different models sometimes failed to converge, we also reported the proportion of converged runs achieved by each model.

4.3.1. Scenario 1: Baseline

The baseline scenario corresponded to the situation with moderate mean intervention effect for non-malignant tumors ($\log IRR = -0.5$), moderate difference in mean intervention effect between malignant and non-malignant tumors ($\Delta_{\log IRR} = 0.5$) and small residual heterogeneity of the intervention effect ($\sigma^2_{\log IRR} = 0.05$). Results for this scenario are displayed in Table **4**. Overall, the univariate and bivariate versions of the Re-Poi model provided more similar results than the two versions of the Re-NB models. Moreover, these latter ran into more convergence issues.

The bivariate Re-NB models provided the best estimates for the mean intervention effect for non-malignant tumors; biases were lower and coverage rates closer to nominal (i.e. 95%). Regarding the estimation of the mean intervention effect between malignant and non-malignant tumors, all the models obtained small relative bias but only the bivariate Re-NB model's CIs provided acceptable coverage rates (91.14%). The model that performed the best for the estimation of the intervention's residual heterogeneity parameter was again the bivariate Re-NB model (although the relative bias was almost 300%). Coverage rates for this parameter were much too low whatever the model considered.

4.3.2. Scenario 2: No Over-dispersion

In the scenario without over-dispersion, all the models tended to encounter more numerical issues than under the scenario 1, especially the bivariate Re-NB model whose proportion of converged runs was below 10% (Table **5**). Again the bivariate Re-NB model provided the best estimate for the mean intervention effect for non-malignant tumors (relative bias < 5%). However, the CIs provided by this model for this parameter were too conservative. For the residual heterogeneity parameter, the bivariate Re-Poi model was the only one to obtain unbiased estimates but its CIs displayed the lowest coverage rates (15%). Compared to scenario 1, coverage rates obtained for parameter $\Delta_{\log IRR}$ was now satisfactory for all models.

4.3.3. Scenario 3: Large Residual Heterogeneity

With large residual heterogeneity ($\sigma^2_{\log IRR} = 2$; Table **6**), biases in the mean intervention effect for non-malignant tumors and difference in mean intervention effect estimates were more or less similar to those obtained in the baseline scenario (i.e. Scenario 1; Table **4**). However, coverage rates tended to be lower and confidence intervals wider. Regarding the estimate of the residual heterogeneity parameter, both Re-NB models underestimated this parameter (median relative bias = -63.77% for the univariate model and -30.90% for the bivariate one), whereas the

Table 4: Models Performances under Scenario 1

True para-meter value	Model	Estimate	Relative bias (in %)	Coverage rate (in %)	CI's width	Converged runs (in %)
$\log IRR = -0.5$	Uni Re-Poi	-0.68	-35.15	87.14	2.21	77.0
	Bi Re-Poi	-0.78	-55.24	84.21	2.24	85.4
	Uni Re-NB	-0.64	-27.90	91.91	2.20	59.2
	Bi Re-NB	-0.57	-13.77	94.76	2.49	55.3
$\sigma^2_{\log IRR} = 0.05$	Uni Re-Poi	1.02	1936.65	23.12	6.07	77.0
	Bi Re-Poi	1.02	1933.33	9.94	5.03	85.4
	Uni Re-NB	0.41	722.40	47.05	3.43	59.2
	Bi Re-NB	0.20	295.24	4.34	0.12*	55.3
$\Delta_{\log IRR} = 0.5$	Uni Re-Poi	0.46	-8.63	68.05	2.28	77.0
	Bi Re-Poi	0.49	-2.81	66.67	2.22	85.4
	Uni Re-NB	0.48	-3.46	89.54	3.22	59.2
	Bi Re-NB	0.48	-4.98	91.14	3.43	55.3

Note: Median values are provided for the estimate and CI's width. Confidence intervals were computed using the Wald method.
*Percentage value.

Table 5: Models Performances under Scenario 2

True para-meter value	Model	Estimate	Relative bias (in %)	Coverage rate (in %)	CI's width	Converged runs (in %)
$\log IRR = -0.5$	Uni Re-Poi	-0.54	-7.36	97.10	1.53	20.7
	Bi Re-Poi	-0.54	-7.12	95.40	1.36	54.4
	Uni Re-NB	-0.56	-12.06	94.68	1.29	79.0
	Bi Re-NB	-0.48	4.94	100	1.61	8.8
$\sigma^2_{\log IRR} = 0.05$	Uni Re-Poi	0.17	239.88	85.99	3.84	20.7
	Bi Re-Poi	0.05	6.04	15.07	0.04*	54.4
	Uni Re-NB	0.24	372.99	46.71	1.05	79.0
	Bi Re-NB	0.17	245.45	37.50	0.63	8.8
$\Delta_{\log IRR} = 0.5$	Uni Re-Poi	0.50	0.44	96.62	1.86	20.7
	Bi Re-Poi	0.50	-0.02	95.59	1.72	54.4
	Uni Re-NB	0.51	1.14	95.32	1.71	79.0
	Bi Re-NB	0.41	-18.26	97.73	1.98	8.8

Note: Median values are provided for the estimate and CI's width. Confidence intervals were computed using the Wald method.
*Percentage value.

Re-Poi models overestimated it (median relative bias around 30%). All CIs obtained for this parameter were wider than those obtained in the baseline scenario, which improved all models' coverage rates, especially for the univariate Re-Poi and Re-NB models, which obtained values close to nominal.

4.3.4. Scenario 4: No Mean Intervention Effect

The last scenario investigated was that of an intervention with a null average effect for both malignant and non-malignant tumors (i.e. $\log IRR = 0$ and $\Delta_{\log IRR} = 0$). Results obtained under this scenario are provided in Table **7**. They were virtually identical to those obtained under the scenario of efficient intervention for non-malignant tumors and non-efficient for malignant tumors (i.e. the baseline scenario; Table **4**). The bivariate Re-NB model again provided the most reliable results for estimating the mean intervention effect for non-malignant tumors and the difference in mean intervention effect between malignant and non-malignant tumors. However, it also encountered more numerical issues. As for the residual heterogeneity parameters, all models provided poor estimates.

5. DISCUSSION

In meta-analysis of IR data, the classical inverse-variance weighting method fails to provide valid estimates when the event rate is low. One possible solution is to use the MH method, but it is only valid

Table 6: Models Performances under Scenario 3

True para-meter value	Model	Estimate	Relative bias (in %)	Coverage rate (in %)	CI's width	Converged runs (in %)
$\log IRR = -0.5$	Uni Re-Poi	-0.86	-71.15	84.03	2.83	91.4
	Bi Re-Poi	-0.85	-69.42	83.97	2.99	94.8
	Uni Re-NB	-0.82	-64.96	84.74	2.84	67.5
	Bi Re-NB	-0.53	-6.98	89.03	2.98	69.3
$\sigma^2_{\log IRR} = 2$	Uni Re-Poi	2.57	28.49	93.44	9.68	91.4
	Bi Re-Poi	2.62	31.00	80.06	9.94	94.8
	Uni Re-NB	0.72	-63.77	96.30	4.49	67.5
	Bi Re-NB	1.38	-30.90	30.30	0.59*	69.3
$\Delta_{\log IRR} = 0.5$	Uni Re-Poi	0.48	-3.24	66.19	2.25	91.4
	Bi Re-Poi	0.51	1.85	64.45	2.24	94.8
	Uni Re-NB	0.51	1.69	89.19	3.69	67.5
	Bi Re-NB	0.57	14.09	90.19	3.71	69.3

Note: Median values are provided for the estimate and CI's width. Confidence intervals were computed using the Wald method.
*Percentage value.

Table 7: Models Performances under Scenario 4

True para-meter value	Model	Estimate	Relative bias (in %)	Coverage rate (in %)	CI's width	Converged runs (in %)
$\log IRR = 0$	Uni Re-Poi	-0.18	-	86.29	2.09	81.7
	Bi Re-Poi	-0.24	-	83.33	2.14	86.8
	Uni Re-NB	-0.15	-	91.09	2.14	60.6
	Bi Re-NB	-0.10	-	93.53	2.38	58.7
$\sigma^2_{\log IRR} = 0.05$	Uni Re-Poi	1.00	1905.72	20.44	5.81	81.7
	Bi Re-Poi	1.04	1970.04	11.15	4.81	86.8
	Uni Re-NB	0.41	710.59	46.70	3.36	60.6
	Bi Re-NB	0.21	312.32	4.60	0.1*	58.7
$\Delta_{\log IRR} = 0$	Uni Re-Poi	-0.06	-	65.73	2.13	81.7
	Bi Re-Poi	-0.10	-	64.02	2.06	86.8
	Uni Re-NB	-0.02	-	89.93	3.22	60.6
	Bi Re-NB	0.03	-	92.33	3.35	58.7

Note: Median values are provided for the estimate and CI's width. Confidence intervals were computed using the Wald method.
*Percentage value.

under the assumption of a homogeneous intervention effect and it fails to include the information contained in DZ and SA studies. In this paper, we investigated the use of the univariate and bivariate Re-NB models to conduct a meta-analysis of heterogeneous incidence rates, in the presence of rare events and SA studies. Through simulations calibrated to mimic a real clinical dataset, we compared the performance of these two models to that of the univariate and bivariate Re-Poi models, which are based on the restrictive assumption of equi-dispersion.

The use of the Re-Poi model for the meta-analysis of IR data is not new and has already been discussed in the literature. For example, Spittal, Pirkis, and Gurrin showed that the univariate Re-Poi model generally outperformed the DerSimonian and Laird method, notably when the number of SZ or DZ studies was high [16]. Stijnen, Hamza, and Özdemir investigated the bivariate Poisson modelling [17]. However, we did not find any published study investigating the use of either the univariate or bivariate Re-NB models for the meta-analysis of IR data in the context of a heterogeneous intervention effect with both rare events and SA studies.

We found larger discrepancies between the univariate and bivariate versions of the Re-NB model than between the univariate and bivariate Re-Poi models. This suggested that taking into account a difference of over-dispersions between the intervention and control arms (i.e. $\eta_T \neq \eta_C$) was more crucial than taking into account a difference between the residuals heterogeneity of the log(IR) (i.e. $\sigma_T^2 \neq \sigma_C^2$).

Overall, we found that except for the scenario of no over-dispersion where all models yielded similar results, the univariate and bivariate Re-NB models were more performant than the univariate and bivariate Re-Poi models. This result was by no means obvious, given the greater complexity of the Re-NB models, which comprise more parameters to be estimated than the Re-Poi models, and the particular settings considered of rare events with many SZ, DZ, and SA studies.

Regarding the estimation of the mean intervention effect for non-malignant tumors (i.e. $\log IRR$), the bivariate Re-NB model was the only model to provide acceptable bias (never larger than 14% of the true parameter's value) and coverage rates (most of the time above 90%) across all scenarios. Due to extreme scarcity of the data (i.e. very few events and studies), results obtained for the residual heterogeneity of the intervention effect (i.e. $\sigma^2_{\log IRR}$) were poor across all scenarios investigated and whatever the model considered. Finally, biases in the difference in mean intervention effect parameter (i.e. $\Delta_{\log IRR}$) were acceptable and approximately the same for the four models, across the four scenarios investigated. Nevertheless, both Re-NB models provided CIs for this parameter with better coverage rates than the Re-Poi models.

To sum up, in settings of rare events, intervention effect heterogeneity, and SA studies, we highly recommend the use of the Re-NB models for the meta-analysis of incidence rate data. Indeed, count data often exhibit over-dispersion (as groups of individuals considered are heterogeneous and there are many unmeasured risk factors) and we showed that these models performed better than the univariate and bivariate Re-Poi models. Under the simulated scenario of equi-dispersion, the Re-NB models

provided similar results as the Re-Poi models. We would furthermore recommend the bivariate Re-NB model, as it allows more flexibility in modelling the IRs than its univariate counterpart.

Nevertheless, there are two limitations worth mentioning. First, convergence might be more difficult to achieve with the bivariate Re-NB model (i.e. proportion of converged runs achieved by this model was often below 60%, whereas it was most of the time above 80% for both Re-Poi models). We believe that convergence rates can be improved by selecting better starting values, which could be provided by the estimation of a less complex model such as the bivariate Re-Poi. Another solution could be to choose a conjugate distribution for the random effects to obtain a closed-form likelihood, which would be easier to maximize [24].

Second, results obtained for the residual heterogeneity parameter were poor, whatever the scenario considered. Notice that our simulations were calibrated to mimic a real clinical dataset where not only events were rare, but also few studies were included in the meta-analysis. Gathering more studies might improve the situation. Nevertheless, even the most sophisticated statistical method cannot compensate for extreme scarcity of the data and absence of information. A Bayesian approach could be adopted, but it is well known that in the setting of rare events, the selection of priors matters and results are subjective [25-26]. Still another option could be to investigate the use of Zero-Inflated models [27]. Finally, to improve the CIs obtained for this parameter, one could consider using the profile likelihood method [28] instead of the Wald method

ACKNOWLEDGEMENT

The authors are thankful to Arnaud Toussaint for providing the dataset that was used as illustrative example.

CONFLICT OF INTEREST

None declared.

FUNDING

The work of RpR was supported by the Swiss National Science Foundation (grant number: 200021_169504/1).

REFERENCES

[1] Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. BMJ Evid Based Med 2016; 21(4): 125-7. http://dx.doi.org/10.1136/ebmed-2016-110401

[2] Normand SLT. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 1999; 18: 321-59. https://doi.org/10.1002/(SICI)1097-0258(19990215)18:3<321::AID-SIM28>3.0.CO;2-P

[3] Nissen SE, Wolski K. Effect of Rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 2007; 356: 2457-71. https://dx.doi.org/10.1056/NEJMoa0727

[4] Niël-Weise BS, Stijnen T, van den Broek PJ. Anti-infective-treated central venous catheters: a systematic review of randomized controlled trials. Intensive Care Med 2007; 33: 2058-68. https://dx.doi.org/10.1007/s00134-007-0897-3

[5] DerSimonian R, Laird D. Meta-analysis in clinical trials. Control Clin Trials 1986; 7: 177-88. https://doi.org/10.1016/0197-2456(86)90046-2

[6] Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. Stat Med 2017; 36(5): 855-75. https://dx.doi.org/10.1002/sim.7141

[7] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004; 23: 1351-75. https://dx.doi.org/10.1002/sim.1761

[8] Friedrich JO, Adhikari NK and Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. BMC Med Res Methodol 2007; 7: 5. https://dx.doi.org/10.1186/1471-2288-7-5

[9] Kuss O, Gummert JF and Borgermann J. Meta-analysis with rare events should use adequate methods. J Thorac Cardiovasc Surg 2008; 136(1): 241. https://doi.org/10.1016/j.jtcvs.2007.12.046

[10] Keus F, Wetterslev J, Gluud C, *et al*. Robustness assessments are needed to reduce bias in meta-analyses that include zero-event randomized trials. Am J Gastroenterol 2009; 104(3): 546-51. https://dx.doi.org/10.1038/ajg.2008.22

[11] Piaget-Rossel R, Taffé P. Meta-analysis of rare events under the assumption of a homogeneous treatment effect. Biom J 2019; 1-18. https://dx.doi.org/10.1002/bimj.201800381

[12] Piaget-Rossel R, Taffé P. A pseudo-likelihood approach for the meta-analysis of homogeneous treatment effects: exploiting the information contained in single-arm and double-zero studies. J Stat Adv Theory Appl 2019; 21: 91-117. http://scientificadvances.co.in/admin/img_data/1365/images/JSATA7100122046RPiagetRossel.pdf

[13] Kuss O. Statistical methods for meta-analyses including information from studies without any events – add nothing to nothing and succeed nevertheless. Stat Med 2015; 34: 1097-116. https://doi.org/10.1002/sim.6383

[14] Guevara JP, Berlin JA, Wolf FM. Meta-analytic methods for pooling rates when follow-up duration varies: a case study. BMC Med Res Methodol 2004; 4: 17. https://doi.org/10.1186/1471-2288-4-17

[15] Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. Biom J 2015; 57: 633-48. https://dx.doi.org/10.1002/bimj.20140018

[16] Spittal JM, Pirkis J, Gurrin CL. Meta-analysis of incidence rate data in the presence of zero events. BMC Med Res Methodol 2015; 15: 42. https://doi.org/10.1186/s12874-015-0031-0

[17] Stijnen T. Hamza HT, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat Med 2010; 29: 3046-67. https://dx.doi.org/10.1002/sim.4040

[18] Hilbe JM. Negative binomial regression. Cambridge University Press; 2011. https://doi.org/10.1017/CBO978051197342

[19] Herbison P, Robertson MC, McKenzie JE. Do alternative methods for analysing count data produce similar estimates? Implications for meta-analyses. Syst Rev 2015; 4(1): 163. https://doi.org/10.1186/s13643-015-0144-x

[20] Mullins P, Sharplin P, Yki-Jarvinen H, Riddle MC, Haring HU. Negative binomial meta-regression analysis of combined glycosylated hemoglobin and hypoglycemia outcomes across eleven Phase III and IV studies of insulin glargine compared with neutral protamine Hagedorn insulin in type 1 and type 2 diabetes mellitus. Clin Ther 2007; 29(8): 1607-19. https://doi.org/10.1016/j.clinthera.2007.08.020

[21] Toussaint A. 'Width of margins in phyllodes tumors of the breast: the controversy drags on – a systematic review'. 2017. Diplôme interuniversitaire de sénologie et pathologie mammaire, François-Rabelais University and Rennes University, Tours and Rennes.

[22] Agresti A, Hartzel J. Tutorial in biostatistics: Strategies for comparing treatment on a binary response with multi-center data. Stat Med 2000; 19: 1115-39. https://doi.org/10.1002/(SICI)1097-0258(20000430)19:8<1115::AID-SIM408>3.0.CO;2-X

[23] StataCorp. Stata Statistical Software: Release 15. 2017. College Station, TX: StataCorp LP.

[24] Cai T, Parast L, Ryan L. Meta-analysis for rare events. Stat Med 2010; 29: 2078-89. https://dx.doi.org/10.1002/sim.3964

[25] Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague ? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med 2005; 24: 2401-28. https://dx.doi.org/10.1002/sim.2112

[26] Senn S. Trying to be precise about vagueness. Stat Med 2007; 26: 1417-30. https://dx.doi.org/10.1002/sim.2639

[27] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 1992; 34: 1-14. https://dx.doi.org/10.2307/1269547

[28] Böhning D, Kuhnert R, Rattanasiri S. Meta-analysis of binary data using profile likelihood. CRC Press; 2008. https://dx.doi.org/10.1201/9781420011333

Received on 01-09-2019 Accepted on 24-09-2019 Published on 10-10-2019

https://doi.org/10.6000/1929-6029.2019.08.07