Received: 27 January 2017 | Revised: 5 September 2017 | Accepted: 29 October 2017

DOI: 10.1002/sim.7569

RESEARCH ARTICLE

WILEY Statistics in Medicine

Data-generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis

Konstantinos Pateras | Stavros Nikolakopoulos | Kit Roes

Department of Biostatistics and Research Support, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Correspondence
Konstantinos Pateras, Department of Biostatistics and Research Support, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.
Email: kostas.pateras@gmail.com

Funding information
FP7-HEALTH-2013-INNOVATION-1 project Advances in Small Trials Design for Regulatory Innovation and Excellence (ASTERIX), Grant/Award Number: 603160

Simulation studies to evaluate performance of statistical methods require a well-specified data-generating model. Details of these models are essential to interpret the results and arrive at proper conclusions. A case in point is random-effects meta-analysis of dichotomous outcomes. We reviewed a number of simulation studies that evaluated approximate normal models for meta-analysis of dichotomous outcomes, and we assessed the data-generating models that were used to generate events for a series of (heterogeneous) trials. We demonstrate that the performance of the statistical methods, as assessed by simulation, differs between these 3 alternative data-generating models, with larger differences apparent in the small population setting. Our findings are relevant to multilevel binomial models in general.

KEYWORDS

dichotomous outcomes, data-generating model, heterogeneity, meta-analysis

1 | INTRODUCTION

Increasingly, simulation studies are used to assess properties of statistical methods in more complex settings. In addition to the statistical models used, the actual operational model and methods with which data are generated can impact the results. The data-generating model (DGM)[1] is essential to interpret the results, to arrive at proper conclusions and to compare between different simulation studies. More often than not, in our simulation work, we returned to the details of this DGM to understand results and sometimes correct to better fit the statistical model and realistic scenarios. A case in point is the random-effects meta-analysis of dichotomous outcomes, towards which recently several simulation based research papers addressed different questions, particularly for a few or small trials.[2-4]

The standard model for random-effects meta-analysis assumes approximately normal effect estimates $Y_i \sim N(\theta_i, s_i^2)$, for trial $i = 1, \ldots k$ for the study-specific effects θ_i and a normal-normal hierarchical model around the study effects $\theta_i \sim N(\theta, \tau^2)$, where s_i^2 are the study-specific within-study variances and τ^2 is the between-study variance. In the case of dichotomous outcomes, we can model the study-specific effects θ_i as the $log(OR) = logit(pT) - logit(pC)$, where pT is the experimental treatment arm event rate and pC the control arm event rate. Evidently, the normal approximation to the binomial distribution breaks down in the case of small samples or small number of events and this can have consequences for the DGM and its use in simulation studies.

Simulations of (individual) trial data in this setting, particularly for small trials, would typically generate numbers of events per trial arm according to binomial distributions, given pT and pC. However, the additional between-study variability implied by the (approximate) normal-normal model in this case should now be incorporated in modelling pT and pC, which a priori can be done in different ways. We reviewed a number of simulation studies that used the normal-normal model for dichotomous outcomes[2-11] and assessed the DGMs used to produce event rates (pT, pC) and generate events for a series of (heterogeneous) trials. In Section 2, we present and discuss the DGMs. In Section 3, we perform a comparison of the DGMs under 3 widely applied meta-analytical models via a simulation study. The manuscript concludes with a discussion in Section 4.

2 | DATA-GENERATING MODELS

In the literature, so far, at least 3 alternative DGMs were used for generating individual trial data. The first makes the assumption of homogeneity in the control arm and places all the between-study variance in the event rate pT of the treatment arm[9,11]; we refer to this as "*pCFixed*." The second is based on the assumption of a fixed average trial risk ($\boldsymbol{p_i}0 = (\boldsymbol{p_{iT}}+\boldsymbol{p_{iC}})/2$), with which we calculate the event probability in each arm, based on a simulated overall treatment effect[4,10]; we refer to this as "*pAverage*." The third is based on the incorporation of the between-study variance in both treatment arms via the use of logits[2,7]; we refer to this as "*pRandom*." The steps to generate events for each DGM are presented below.

Algorithm 1 Data Generating Model *pCfixed*

```
Set $\theta, \tau$, a range for $p_{iC}$ and a range for $m_i$, $i = 1, \ldots, k$ and $j = (C)ontrol, (T)reatment$.
$m_i \sim Uniform(m_{lo}, m_{up})$ - Generate study-arm sample sizes.
$n_{ij} = m_i$ - Set equal study-arm allocation ratios.
$\theta_i \sim Normal(\theta, \tau)$ - Generate study-specific treatment effects.
$p_{iC} \sim Uniform(\alpha, \beta)$ - Generate a study-specific control event probability.
$p_{iT} = p_{iC} \cdot exp(\theta_i)/(1 - p_{iC} + p_{iC} \cdot exp(\theta_i))$ - Compute the study-specific treatment event probability.
$r_{ij} \sim Binomial(p_{ij}, n_{ij})$ for $j = C$ and $T$ - Generate study events.
```

Algorithm 2 Data Generating Model *pAverage*

```
1: Set $\theta, \tau$, a range for $p_i0$ and a range for $m_i$, $i = 1, \ldots, k$ and $j = (C)ontrol, (T)reatment$.
2: $m_i \sim Uniform(m_{lo}, m_{up})$ - Generate study-arm sample sizes.
3: $n_{ij} = m_i$ - Set equal study-arm allocation ratios.
4: $\theta_i \sim Normal(\theta, \tau)$ - Generate study-specific treatment effects.
5: $p_i0 \sim Uniform(\alpha, \beta)$ - Generate a study-specific average event probability.
6: $p_i0 = \sum_{j=1}^{2} p_{ij}/2$
7: $\theta_i = log\left(\frac{(p_{iC})\cdot(1-p_{iT})}{(p_{iT})\cdot(1-p_{iC})}\right)$.
8: Solving (6) and (7) we acquire $p_{ij}$.
9: $r_{ij} \sim Binomial(p_{ij}, n_{ij})$ for $j = C$ and $T$ - Generate study events.
```

Algorithm 3 Data Generating Model *pRandom*

```
Set $\theta, \tau, p_{iC,Init}$ and a range for $m_i$, $i = 1, \ldots, k$ and $j = (C)ontrol, (T)reatment$.
$p_{iT,Init} = p_{iC,Init} \cdot exp(\theta)/(1 - p_{iC,Init} + p_{iC,Init} \cdot exp(\theta))$ - Compute the initial study-specific treatment event probability.
$m_i \sim Uniform(m_{lo}, m_{up})$ - Generate study-arm sample sizes.
$n_{ij} = m_i$ - Set equal study-arm allocation ratios.
$\mu_{ij} = log(p_{ij,Init}/1 - p_{ij,Init})$ - Compute mean logits given initial fixed event rates $p_{ij,Init}$.
$logit_{ij} \sim Normal(\mu_{ij}, \tau/\sqrt{2})$ - Generate study-specific control and treatment logits.
$p_{ij} = \frac{1}{1+e^{-logit_{ij}}}$ - Back-calculate the event rates for each trial arm.
$r_{ij} \sim Binomial(p_{ij}, n_{ij})$ for $j = C$ and $T$ - Generate study events.
```

Note that for 2 of the DGMs discussed, the use of uniform distributions is used (*pCFixed* and *pAverage*). This is done to replicate their use in the literature[4,9,10]. This adds an additional source of variability, not specifically modelled by the normal-normal hierarchical model. We keep using the term "fixed" and "homogeneous" for these DGMs, even if the probability of events is not kept fixed across studies. We retain the term "heterogeneous" for referring to heterogeneity resulting from the variance parameter of the random effects model.

As Figure 1 demonstrates, the primary difference among the 3 presented DGMs lies in the joint distribution of the 2 model event rate parameters, as used in generating data. Homogeneity of the control group event rates (*pCFixed*) has been discussed previously[12] and can be observed in the densities of Figure 1. The study-specific control event rates are homogeneous—coming from a $\mathit{Uniform}(0.1, 0.3)$— while the study-specific treatment event rates are heterogeneous. The

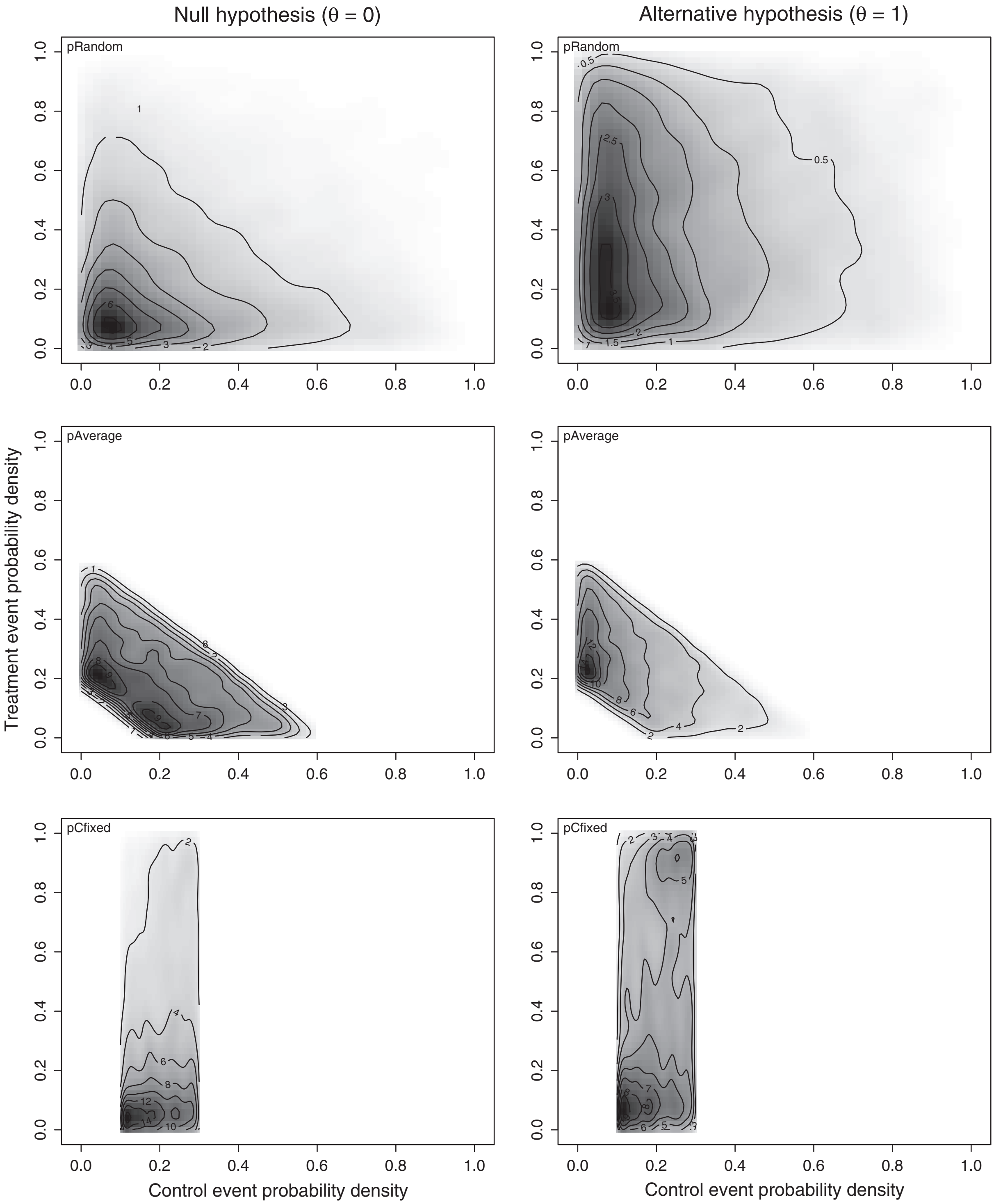

FIGURE 1 Empirical (numerically estimated) joint event probability densities for the control and treatment arm of the 3 data-generating models under the null and alternative hypothesis with substantial between-study standard-deviation ($\tau = 2$), small sample size ($n_{ij} = m_i \sim \mathit{Uniform}(20, 30), i = 1, 2; j = \mathit{Control}, \mathit{Treatment}$) and an (average) event rate, either as a fixed value of 0.20, or as a mean of 0.20 of a $\mathit{Uniform}(0.1, 0.3)$ distribution

pAverage approach makes an intuitively restrictive assumption since it constrains the simulated values of the control and treatment arm around an average true risk rate. The *pRandom* approach places the between-study variability in both treatment arms without imposing additional constraints. In this DGM, it is common to assume equal between-study standard-deviation ($\tau/\sqrt{2}$) in both arms, an assumption which might not always hold in practice but can be relaxed. For example, the standard of care—control treatment—might be less variable between studies in comparison to the experimental treatment; or more variable if the standard of care differs between regions or countries, a flexibility that is not straightforward to implement in the other 2 DGMs discussed here.

Indeed, in the *pCFixed* DGM, the probability of events in the 2 arms is not fixed but rather randomly generated via a joint distribution at the study-parameter level, where the control group rate is considered to be independent from the effect size. The *pRandom* DGM is largely the same, except that the range of pC is not restricted, and its distribution is skewed within its range. Naturally, after incorporating smaller heterogeneity in the control group, *pCFixed* can be considered a special case of *pRandom* if the 2 parameters of the uniform distribution generating event rates in *pCFixed* are equal (Algorithm 3—Step 6).

An important difference between the presented DGMs arises from their ability to accommodate ranges of event rates. The *pAverage* directly defines the average event rate ($p0$), the *pCFixed* directly defines the control group rate (pC), whereas the *pRandom* does not allow a direct impact on event rates. These fundamental characteristics of the 3 DGMs render their fair comparison through simulation less trivial. For the specific scenario studied here, where probabilities of events on the control groups are smaller than 50%, whenever the average effect size is positive, the control group event rate for *pAverage* is, on average, smaller than the competing DGM's simulations. This implies smaller numbers of events in the 2 arms. The total number of events is related to power. Therefore, differences in empirical power of the *pAverage* method may appear partially due to this difference in the average rate. Nonetheless, the constraints of the *pAverage* DGM inherently restrict the DGM from jointly exploring very low event rates (Figure 1), which minimizes the event rate's impact on power. Thus, the *pAverage* DGM makes (empirical) very large effect sizes less probable. This restriction becomes problematic particularly for studies in small populations where we usually seek to observe very large effect sizes.

Ideally, results and conclusions of simulation studies are expected to hold for the statistical model specified, and not to depend on the characteristics of the used DGM. Since they all generate the same true overall treatment effect, when we use statistical methods in the setting of many large trials and relatively frequent events, we may expect similar results under different DGMs. However, in the case of a small number of trials with small sample sizes, the normal approximation of the logOR might be insufficient and more sensitive to the choice of DGM. Thus, possible differences between the observed performance of methods may be enhanced.

Evidence from a few small trials would often become available or would be sufficiently similar to be synthesized, during a drug development and evaluation in rare diseases.[13-15] Until recently, the evaluation of meta-analytical methods in this rare disease context was not common. However, more attention has been drawn to this topic, especially since the initiation of 3 European projects, focused on characteristics of statistical methodologies in small populations **(ASTERIX, IDeAl, and InSPiRe)**. A number of articles have now been published that evaluate methods for a meta-analysis of a few and even 2 small trials.[2,3,6]

3 | SIMULATION STUDY AND RESULTS

To compare the implications of the 3 DGMs for both a meta-analysis of 2 large trials ($n_{ij} = m_i \sim Uniform(230, 240)$, $i = 1, 2; j = (C)ontrol, (T)reatment$) and a meta-analysis of 2 small trials ($n_{ij} = m_i \sim Uniform(20, 30)$, $i = 1, 2; j = C, T$), we follow Gonnermann et al[2] to evaluate the statistical properties of 3 meta-analysis methods; (1) a fixed-effect meta-analysis (FE), (2) a random-effects meta-analysis with the application of DerSimonian and Laird heterogeneity estimator[16] (DL) and (3) a random-effects meta-analysis with the Hartung and Knapp correction[7] (HK) for a meta-analysis of 2 trials. The (average) event rate is assumed to be 0.20. This is, however, interpreted and implemented differently between the DGMs used, ie, either as a fixed value of 0.20 (*pRandom*) or as a mean of 0.20 of $Uniform(0.1, 0.3)$ on either pC (*pCFixed*) or $p0$ (*pAverage*). We also apply a continuity correction (0.5) in all cells of a trial with zero cells. We assume equal allocation ratios within each trial. We present results under the null and the alternative hypothesis with varied levels of true between-study standard-deviation $\tau \in \{0.001, 0.5, 1, 2\}$, which corresponds to relative heterogeneity of $I^2 \approx \{0.01\%, 47\%, 63\%, 75\%\}$ for a small trial meta-analysis and $I^2 \approx \{0.05\%, 74\%, 84\%, 90\%\}$ for a large trial meta-analysis.

TABLE 1 Empirical type I error and empirical power based on 10^6 simulations

			Empirical type I error			**Empirical power**		
			($\theta = 0$)			**($\theta = 1$)**		
Method	**S.size**	τ	**PR**	**PA**	**PCF**	**PR**	**PA**	**PCF**
DL	Small	0.001	0.029	0.028	0.028	0.482	0.340	0.465
		0.5	0.062	0.058	0.060	0.417	0.311	0.420
		1	0.133	0.118	0.125	0.341	0.280	0.357
		2	0.224	0.210	0.205	0.297	0.278	0.298
	Large	0.001	0.037	0.038	0.038	0.998	0.986	0.997
		0.5	0.207	0.201	0.202	0.759	0.736	0.757
		1	0.267	0.264	0.263	0.487	0.476	0.489
		2	0.288	0.286	0.284	0.350	0.345	0.351
HK	Small	0.001	0.047	0.048	0.048	0.132	0.114	0.129
		0.5	0.047	0.049	0.049	0.106	0.098	0.107
		1	0.050	0.050	0.051	0.080	0.081	0.083
		2	0.056	0.059	0.059	0.065	0.078	0.066
	Large	0.001	0.052	0.052	0.052	0.400	0.339	0.394
		0.5	0.050	0.050	0.050	0.162	0.154	0.163
		1	0.050	0.049	0.049	0.092	0.089	0.092
		2	0.055	0.050	0.049	0.065	0.061	0.061
FE	Small	0.001	0.033	0.032	0.032	0.572	0.388	0.547
		0.5	0.075	0.070	0.075	0.532	0.369	0.539
		1	0.186	0.151	0.191	0.486	0.345	0.537
		2	0.373	0.270	0.391	0.480	0.346	0.553
	Large	0.001	0.049	0.049	0.049	1.000	0.999	1.000
		0.5	0.397	0.377	0.390	0.968	0.947	0.967
		1	0.627	0.570	0.632	0.857	0.800	0.883
		2	0.767	0.641	0.803	0.816	0.696	0.872

Abbreviations: PR, *pRandom*; PA, *pAverage*; PCF, *pCFixed*; FE, fixed-effect approach; HK, Hartung and Knapp approach; DL, DerSimonian Laird approach; θ, overall treatment effect (log odds ratio); τ, between-study standard-deviation; small sample size, $n_{ij} = m_i \sim Uniform(20, 30)$; large sample size, $n_{ij} = m_i \sim Uniform(230, 240)$; $i = 1, 2 j = Control, Treatment$.

Table 1 presents the differences between the empirical power curves when a treatment effect is present ($\theta = 1$) and under the null hypothesis ($\theta = 0$) for each DGM and each considered meta-analytical method. The *pAverage* DGM produces data that result in lower power than the other 2 DGMs, especially for the FE and DL approach. This can be explained by the constraints that are induced on the event probabilities (Figure 1), which are in turn influenced by the specific choice of $\alpha = 0.1$ and $\beta = 0.3$ for the uniform distributions. In addition, regarding high levels of true heterogeneity, the *pCFixed* tends to increase the empirical power of the FE approach. This could be expected, as when $\tau^2 > 0$ and heterogeneity is only applied to the treatment group event rates, larger effect sizes are produced compared to the other 2 DGMs. In terms of type I error, empirical values seem to be heavily dependent on the DGM when the FE approach is assessed. Evaluation of the Hartung and Knapp approach is less affected by the DGM, with small deviations in empirical power, and mostly for the *pAverage* DGM. A graphical representation of the empirical power curves can be found in the Appendix (Figure A1 and A2).

Figure 2 summarizes the performance of the 3 DGMs in terms of coverage of the 95% confidence intervals for small sample sizes ($m_i \sim Uniform(20, 30)$). Under heterogeneous conditions ($\tau = 1$), especially for the FE and DL approaches, the *pRandom* demonstrates lower coverage than the *pCFixed* and *pAverage*, across the considered levels of overall treatment effect. Regarding large sample sizes ($m_i \sim Uniform(230, 240)$), in terms of coverage of the 95% confidence intervals, the 3 DGMs show similar behavior for homogeneous conditions (Figure A3). On the contrary, for heterogeneous conditions the *pAverage* DGM starts to favor the FE and DL approaches when $\theta \geq 2$, bringing the 3 methods 95% coverage relatively closer than *pCFixed* and *pRandom*.

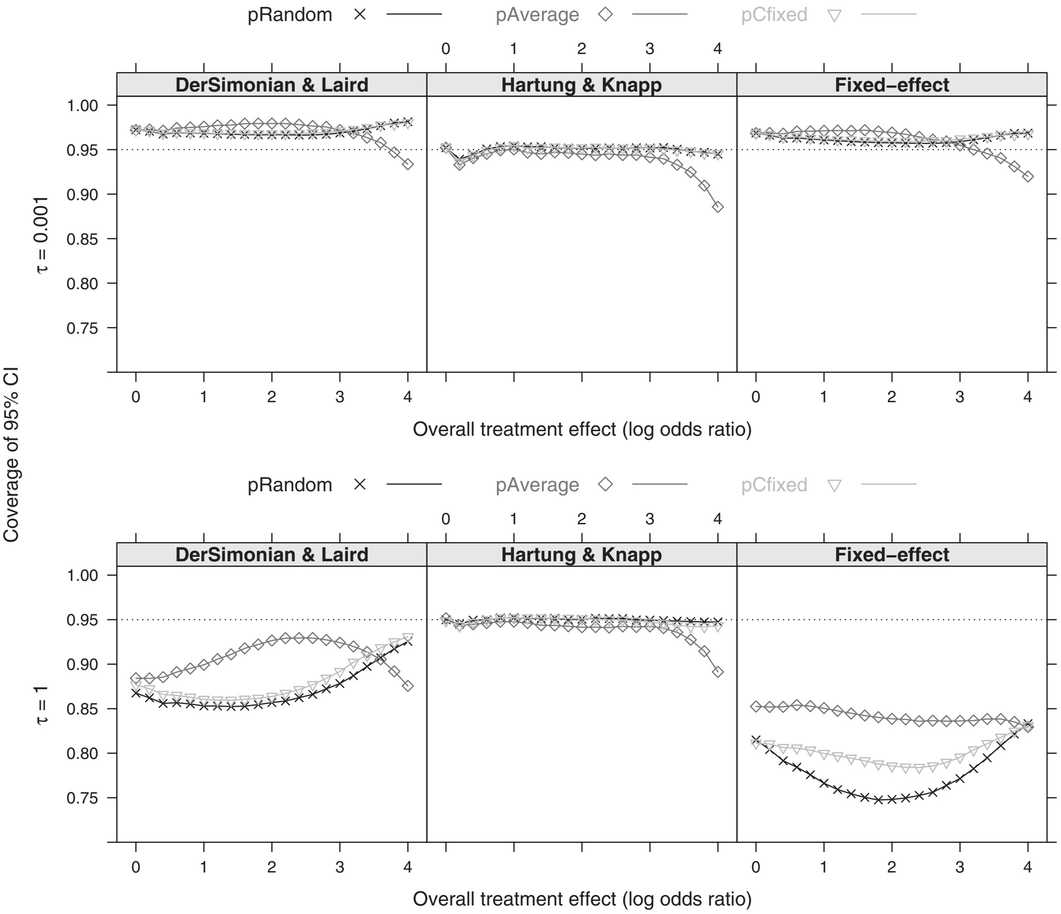

FIGURE 2 Impact of data-generating mechanism in a meta-analysis of two small studies ($n_{ij} = m_i \sim U(20, 30), i = 1, 2; j = \mathit{Control}$, *Treatment*) on coverage of the 95% confidence intervals. τ: between-study standard-deviation

4 | DISCUSSION

The choice of a DGM used in simulation studies is important and has to be consistent with the assumed statistical model under realistic assumptions related to the issue in question. Our simulations show that statistical methods perform differently across DGMs that were used to investigate properties of random-effects meta-analyses. Our simulation is not extensive and does not cover effects in other settings. Nonetheless, we noticed that the divergent behavior of DGMs is preserved when synthesizing many small trials but is reduced when synthesizing many large trials. In contrast to large study meta-analyses simulation studies, the choice of a DGM can impact the conclusions of small study meta-analyses simulation studies to a greater extent. The findings actually extend beyond the presented small population context and hold more generally for multilevel binomial data settings.

The elaboration on the DGM articulates one of the crucial conceptual difficulties of the random-effects model for meta-analysis. In all 3 random-effects DGM formulations and assessments of type I errors, in the presence of heterogeneity, there is also heterogeneity under the null hypothesis. Although all 3 DGMs are designed to produce the same true overall effect, the properties of the modelled joint empirical distribution of the control and treatment event rates can differ dramatically.

As a consequence, simulation studies that use different DGMs for essentially the same overall statistical model have the potential to result in different conclusions regarding performance of the statistical methods investigated. For this reason, methodological reviews for meta-analysis[17,18] have to report in detail the DGM of each study they include and potential consequences of the choice of DGM. If flexible assumptions on the event probability are needed, the use of *pRandom* DGM might be recommended. We believe that not enough emphasis is placed on the proper choice nor on the sufficient reporting of DGMs in both individual simulation studies and methodological reviews.

ACKNOWLEDGEMENTS

This work has been funded by the FP7-HEALTH-2013-INNOVATION-1 project Advances in Small Trials Design for Regulatory Innovation and Excellence (ASTERIX) Grant Agreement No. 603160. We would like to thank the two anonymous reviewers whose feedback improved the manuscript substantially.

ORCID

Konstantinos Pateras http://orcid.org/0000-0002-6005-9798
Stavros Nikolakopoulos http://orcid.org/0000-0002-9769-3725

REFERENCES

1. Cox DR, Hinkley DV. *Theoretical Statistics*. Florida, USA: CRC Press; 1979.
2. Gonnermann A, Framke T, Großhennig A, Koch A. No solution yet for combining two independent studies in the presence of heterogeneity. *Stat Med*. 2015;34(16):2476-2480.
3. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. *Res Synth Methods*. 2016;8(1):79-91.
4. IntHout J, Ioannidis JPa, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Method*. 2014;14(25):1-12.
5. Chung Y, Rabe-Hesketh S, Choi I-H. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat Med*. 2013;32(23):4071-4089.
6. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical J*. 2016;00:1-12.
7. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med*. 2001;20(24):3875-3889.
8. Inthout J, Ioannidis JPA, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. *J Clin Epidemiol*. 2015;68(8):860-869.
9. Novianti PW, Roes KCB, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemp Clin Trials*. 2014;37(1):129-138.
10. Veroniki AA, Mavridis D, Higgins JP, Salanti G. Characteristics of a loop of evidence that affect detection and estimation of inconsistency: A simulation study. *BMC Med Res Method*. 2014;14(106):1-12.
11. 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(15):2401-2428.
12. Senn S. Trying to be precise about vagueness. *Stat Med*. 2007;26(7):1417-1430.
13. Crins ND, Röver C, Goralczyk AD, Friede T. Interleukin-2 receptor antagonists for pediatric liver transplant recipients: a systematic review and meta-analysis of controlled studies. *Pediatric Transplantation*. 2014;18(8):839-850.
14. Zeng Y, Duan X, Ni X, Xu J. TPO receptor agonist for chronic idiopathic thrombocytopenic purpura. *Cochrane Database Syst Rev*. 2010;1:1-54.
15. Somaraju UR, Merrin M. Sapropterin dihydrochloride for phenylketonuria. *The Cochrane Database Syst Rev*. 2015;3(3):1-29.
16. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials*. 1986;7(3):177-188.
17. Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Res Synth Methods*. 2017;8(2):181-198. https//doi.org/10.1002/jrsm.1198.
18. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79.

How to cite this article: Pateras K, Nikolakopoulos S, Roes K. Data generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis. *Statistics in Medicine*. 2018;37:1115–1124. https://doi.org/10.1002/sim.7569

APPENDIX

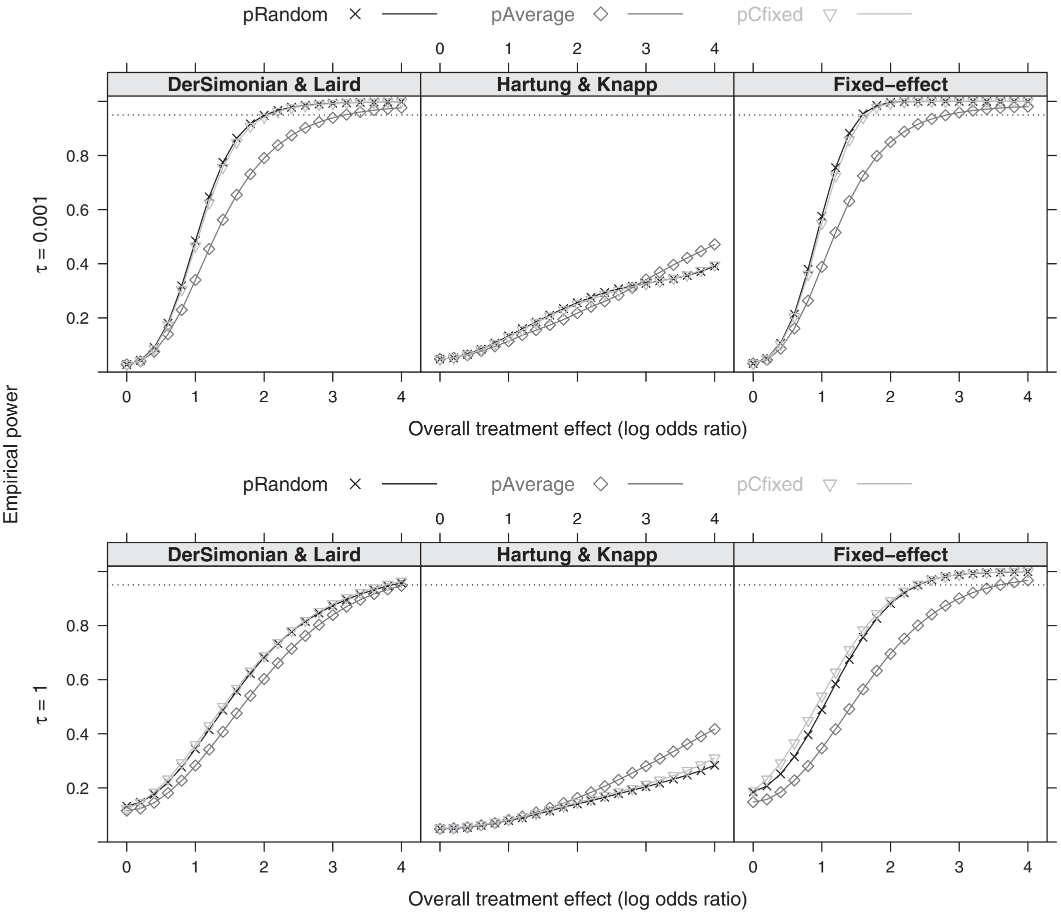

FIGURE A1 Impact of data-generating mechanism in a meta-analysis of 2 small studies ($n_{ij} = m_i \sim \mathit{Uniform}(20, 30)$, $i = 1, 2; j = \mathit{Control}, \mathit{Treatment}$) on empirical power. τ: between-study standard-deviation

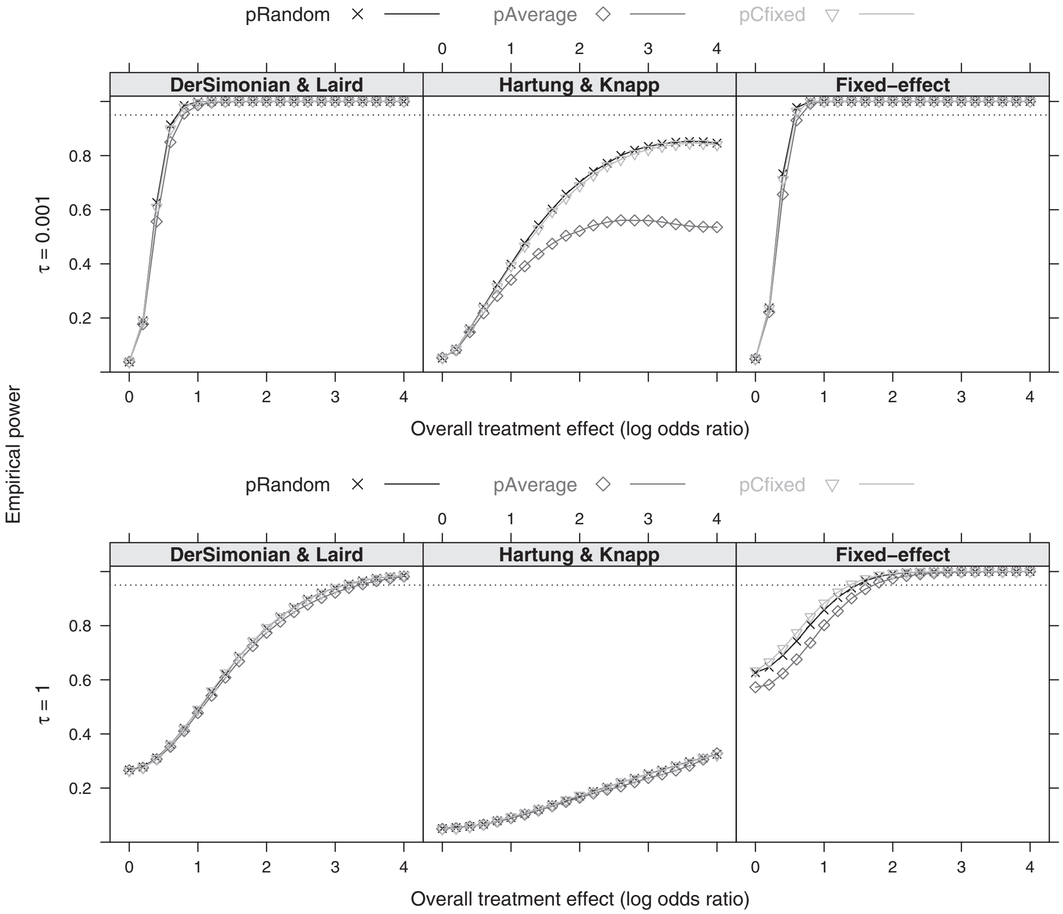

FIGURE A2 Impact of data-generating mechanism in a meta-analysis of 2 large studies ($n_{ij} = m_i \sim Uniform(230, 240)$, $i = 1, 2; j = Control, Treatment$) on empirical power. τ: between-study standard-deviation

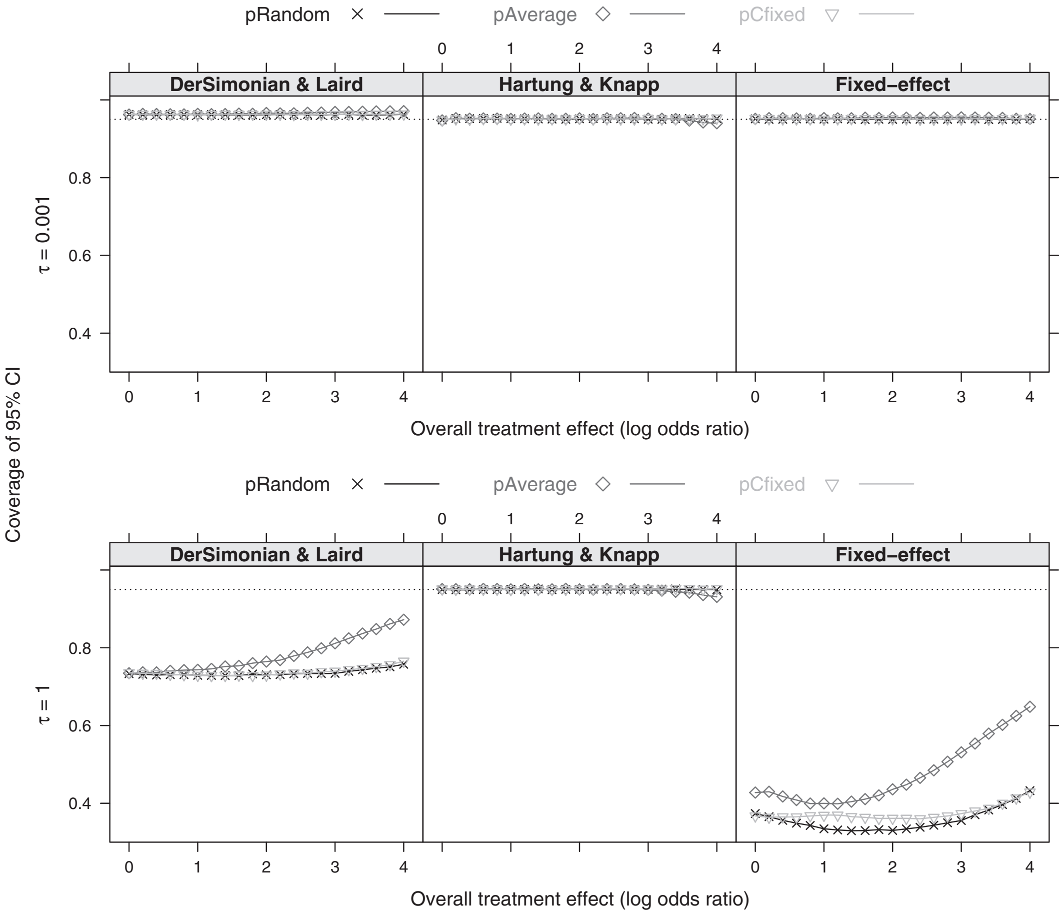

FIGURE A3 Impact of data-generating mechanism in a meta-analysis of 2 large studies ($n_{ij} = m_i \sim Uniform(230, 240)$, $i = 1, 2; j = Control, Treatment$) on coverage of the 95% confidence intervals. τ: between-study standard-deviation