Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression

I. Kosmidis
Department of Statistical Science
University College London
London, WC1E 6BT, U.K
i.kosmidis@ucl.ac.uk

A. Guolo
Department of Statistical Sciences
University of Padova
Via Cesare Battisti, 241, 35121 Padova, Italy
annamaria.guolo@unipd.it

C. Varin
Department of Environmental Sciences, Informatics and Statistics,
Ca' Foscari University of Venice
Via Torino 150, 30170 Venezia Mestre, Italy
cristiano.varin@unive.it

January 10, 2017

Abstract

Random-effects models are frequently used to synthesise information from different studies in meta-analysis. While likelihood-based inference is attractive both in terms of limiting properties and of implementation, its application in random-effects meta-analysis may result in misleading conclusions, especially when the number of studies is small to moderate. The current paper shows how methodology that reduces the asymptotic bias of the maximum likelihood estimator of the variance component can also substantially improve inference about the mean effect size. The results are derived for the more general framework of random-effects meta-regression, which allows the mean effect size to vary with study-specific covariates.
Keywords: *Bias reduction; Heterogeneity; Meta-analysis; Penalized likelihood; Random effect; Restricted maximum likelihood*

1 Introduction

Meta-analysis is a widely applicable approach to combining information from different studies about a common effect of interest. A popular framework for accounting for the heterogeneity between studies is the random-effects specification in DerSimonian & Laird (1986). There is ample evidence that frequentist inference for this specification can result in misleading conclusions, especially if inference is carried out by relying on first-order asymptotic arguments in the common setting of small or moderate number of studies (e.g., van Houwelingen et al., 2002; Guolo & Varin, 2015). The same considerations apply to the random-effects meta-regression model, which is a direct extension of random-effects meta-analysis allowing for study-specific covariates. Proposals presented to account for the finite number of studies include modification of the limiting distribution of test statistics (Knapp & Hartung, 2003), restricted maximum likelihood (Viechtbauer, 2005) and second-order asymptotics (Guolo, 2012). Recently, Zeng & Lin (2015) suggested a double resampling approach that outperforms several alternatives in terms of empirical coverage probability of confidence intervals for the mean effect size.

The current paper studies the extent of the bias of the maximum likelihood estimator of the random-effect variance and introduces a bias-reducing penalized likelihood that yields a substantial improvement in the estimation of the random-effect variance. The bias-reducing penalized likelihood is related to the approximate conditional likelihood of Cox & Reid (1987) and the restricted maximum likelihood for inference about the random-effects variance. The order of the penalty function allows the derivation of a χ^2 approximation of the distribution of the logarithm of the penalized likelihood ratio statistic, which can be used for inference about the fixed-effect parameters. Real-data examples and two simulation studies illustrate the improvement in finite-sample performance against alternatives from the recent literature.

2 Random-effects meta-regression and meta-analysis

Suppose there are K studies about a common effect of interest, each of them providing pairs of summary measures $(y_i, \hat{\sigma}_i^2)$, where y_i is the study-specific estimate of the effect, and $\hat{\sigma}_i^2$ is the associated estimation variance $(i = 1, \ldots, K)$. In some situations, the pairs $(y_i, \hat{\sigma}_i^2)$ may be accompanied by study-specific covariates $x_i = (x_{i1}, \ldots, x_{ip})^\top$, which describe the heterogeneity across studies. In the meta-analysis literature, it is usually assumed that the within-study variances $\hat{\sigma}_i^2$ are estimated well enough to be considered as known and equal to the values reported in each study. Under this assumption, the random-effects meta-regression model postulates that $y_1, \ldots, y_K$ are realizations of random variables $Y_1, \ldots, Y_K$, respectively, which are independent conditionally on independent random effects $U_1, \ldots, U_K$, and the conditional distribution of Y_i given $U_i = u_i$ is $N(u_i + x_i^\top \beta, \hat{\sigma}_i^2)$, where β is an unknown p-vector of effects. The random effect U_i is typically assumed to be distributed according to $N(0, \psi)$, where ψ accounts for the between-study heterogeneity.

In matrix notation, and conditionally on $(U_1, \ldots, U_K)^\top = u$, the random-effects meta-regression model is

$$Y = X\beta + u + \epsilon, \tag{1}$$

where $Y = (Y_1, \ldots, Y_K)^\top$, X is the model matrix of dimension $K \times p$ with $x_i^\top$ in its ith row, and $\epsilon = (\epsilon_1, \ldots, \epsilon_K)^\top$ is a vector of independent errors each with a $N(0, \hat{\sigma}_i^2)$ distribution. Under this specification, the marginal distribution of Y is multivariate normal with mean $X\beta$ and variance $\hat{\Sigma} + \psi I_K$, where I_K is the $K \times K$ identity matrix and $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2)$. The random-effects meta-analysis model is a meta-regression model where X is a column of ones.

The random-effects meta-regression model is used here as a working model for theoretical development. In light of the recent criticisms of the assumption of known within-study variances (see, for example Hoaglin, 2015), § 4 and the Supplementary Material illustrate the good performance of the derived procedures under more realistic scenarios, where the estimation variances are directly related to the estimates of the summary measure.

The parameter β is naturally estimated by weighted least squares as

$$\hat{\beta}(\psi) = \{X^\top W(\psi) X\}^{-1} X^\top W(\psi) Y, \tag{2}$$

with $W(\psi) = (\hat{\Sigma} + \psi I_K)^{-1}$. Then, inference about β can be based on the fact that under model (1), $\hat{\beta}(\psi)$ has an asymptotic normal distribution with mean β and variance $\left\{X^\top W(\psi) X\right\}^{-1}$.

In this case, the reliability of the associated inferential procedures critically depends on the availability of an accurate estimate of the between-study variance ψ. A popular choice is the DerSimonian & Laird (1986) estimator $\hat{\psi}_{\mathrm{DL}} = \max\{0, (Q-n+p)/A\}$, where $Q = (Y - X\hat{\beta}_{\mathrm{F}})^\top \hat{\Sigma}^{-1}(Y - X\hat{\beta}_{\mathrm{F}})$ is the Cochran statistic, with $\hat{\beta}_{\mathrm{F}} = \hat{\beta}(0)$ and $A = \mathrm{tr}(\hat{\Sigma}^{-1}) - \mathrm{tr}\{(X^\top\hat{\Sigma}^{-1}X)^{-1}X^\top\hat{\Sigma}^{-2}X\}$. Viechtbauer (2005) presents evidence of the loss of efficiency of $\hat{\psi}_{\mathrm{DL}}$, which can impact inference; see also Guolo (2012).

Inference about β can alternatively be based on the likelihood function. The log-likelihood function for $\theta = (\beta^\top, \psi)^\top$ in model (1) is

$$\ell(\theta) = -\frac{1}{2}\log|W(\psi)| - \frac{1}{2}R(\beta)^\top W(\psi)R(\beta), \tag{3}$$

where $|W(\psi)|$ denotes the determinant of $W(\psi)$ and $R(\beta) = y - X\beta$. A calculation of the gradient $s(\theta)$ of $\ell(\theta)$ shows that the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}} = (\hat{\beta}_{\mathrm{ML}}^\top, \hat{\psi}_{\mathrm{ML}})^\top$ for θ results from solving the equations

$$\begin{cases} s_\beta(\theta) = X^\top W(\psi)R(\beta) = 0_p\,, \\ s_\psi(\theta) = R^\top(\beta)W(\psi)^2R(\beta) - \mathrm{tr}\,[W(\psi)\}] = 0\,, \end{cases} \tag{4}$$

where 0_p denotes a p-dimensional vector of zeros, and $s_\beta(\theta) = \nabla_\beta \ell(\theta)$ and $s_\psi(\theta) = \partial\ell(\theta)/\partial\psi$, so that $\hat{\beta}_{ML} = \hat{\beta}(\hat{\psi}_{ML})$. As observed in Guolo (2012) and Zeng & Lin (2015), inferential procedures that rely on first-order approximations of the log-likelihood, e.g., likelihood-ratio and Wald statistics, perform poorly when the number of studies K is small to moderate.

3 Bias reduction

3.1 Bias-reducing penalized likelihood

From the results in Kosmidis & Firth (2009, 2010), the first term in the expansion of the bias function of the maximum likelihood estimator is found to be $b(\theta) = \{0_p^\top, b_\psi(\psi)\}^\top$, where

$$b_\psi(\psi) = -\frac{\mathrm{tr}\{W(\psi)H(\psi)\}}{\mathrm{tr}\{W(\psi)^2\}}\,, \tag{5}$$

with $H(\psi) = X\{X^\top W(\psi)X\}^{-1}X^\top W(\psi)$. A sketch derivation for (5) is given in the Appendix. In what follows $b(\theta)$ is called the first-order bias.

The non-zero entries of $W(\psi)$ and the diagonal entries of $H(\psi)$ are all necessarily positive, so the maximum likelihood estimator of ψ is subject to downward bias, which, as also noted in Viechtbauer (2005), affects inference about β, by over-estimating the non-zero entries of $W(\psi)$, and hence over-estimating the information matrix

$$F(\theta) = -E_\theta\left\{\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta^\top}\right\} = \begin{bmatrix} X^\top W(\psi)X & 0_p \\ 0_p^\top & \frac{1}{2}\mathrm{tr}\left\{W(\psi)^2\right\} \end{bmatrix}. \tag{6}$$

This over-estimation of $F(\theta)$ can result in hypothesis tests with large Type I error and confidence intervals or regions with actual coverage appreciably lower than the nominal level.

An estimator that corrects for the first-order bias of $\hat{\theta}_{\mathrm{ML}}$ results from solving the adjusted score equations $s^*(\theta) = s(\theta) - F(\theta)b(\theta) = 0_{p+1}$ (Firth, 1993; Kosmidis & Firth, 2009). Substituting of (4), (5) and (6) in the expression for $s^*(\theta)$ gives that the adjusted score functions for β and ψ are $s_\beta^*(\theta) = s_\beta(\theta)$ and

$$s_\psi^*(\theta) = R^\top(\beta)W(\psi)^2R(\beta) - \mathrm{tr}\,[W(\psi)\{I_K - H(\psi)\}] = 0\,, \tag{7}$$

respectively. The expression for the differential of the log-determinant can be used to show that $s^*_\beta(\theta)$ and $s^*_\psi(\theta)$ are the derivatives of the penalized log-likelihood function

$$\ell^*(\theta) = \ell(\theta) - \frac{1}{2}\log\left|F_{(\beta\beta)}(\psi)\right| , \tag{8}$$

where $\ell(\theta)$ is as in (3), $F_{(\beta\beta)}(\psi) = X^\top W(\psi) X$ is the β-block of the information matrix $F(\theta)$, and $|\cdot|$ denotes determinant, so the solution of the adjusted score equations is the maximum penalized likelihood estimator $\hat{\theta}_{\mathrm{MPL}}$.

For $\beta = \hat{\beta}(\psi)$, expression (8) reduces to both the logarithm of the approximate conditional likelihood of Cox & Reid (1987) for inference about ψ, when β is treated as a nuisance component, and to the restricted log-likelihood function of Harville (1977). Hence, maximising the bias-reducing penalized log-likelihood (8) is equivalent to calculating the maximum restricted likelihood estimator for ψ. The latter estimator was originally constructed to reduce underestimation of variance components in finite samples as a consequence of failing to account for the degrees of freedom that are involved in the estimation of the fixed effects β. Smyth & Verbyla (1996) and Stern & Welsh (2000) have shown the equivalence of the restricted log-likelihood with approximate conditional likelihood in the more general context of inference about variance components in normal linear mixed models.

3.2 Estimation

Given a starting value $\psi^{(0)}$ for ψ, the following iterative process has a stationary point that maximizes (8). At the jth iteration ($j = 1, 2, \ldots$), a new candidate value $\beta^{(j+1)}$ for β is obtained as the weighted least squares estimator (2) at $\psi = \psi^{(j)}$; a candidate value for $\psi^{(j+1)}$ is then computed by a line search for solving the adjusted score equation (7) evaluated at $\beta = \beta^{(j+1)}$. The iteration is repeated until either the candidate values do not change across iterations or the adjusted score functions are sufficiently close to zero.

3.3 Penalized likelihood inference

The profile penalized likelihood function can be used to construct confidence intervals and regions, and carry out hypothesis tests for β. If $\beta = (\gamma^\top, \lambda^\top)^\top$, and $\hat{\lambda}_{\mathrm{MPL},\gamma}$ and $\hat{\psi}_{\mathrm{MPL},\gamma}$ are the estimators of λ and ψ, respectively, from maximising (8) for fixed γ, then the logarithm of the penalized likelihood ratio statistic $2\{\ell^*(\hat{\gamma}_{\mathrm{MPL}}, \hat{\lambda}_{\mathrm{MPL}}, \hat{\psi}_{\mathrm{MPL}}) - \ell^*(\gamma, \hat{\lambda}_{\mathrm{MPL},\gamma}, \hat{\psi}_{\mathrm{MPL},\gamma})\}$ has the usual limiting χ^2_q distribution, where $q = \dim(\gamma)$. To derive this limiting result, note that the adjustment to the scores in (4) is additive and $O(1)$, so the extra terms depending on it and its derivatives in the asymptotic expansion of the penalized likelihood disappear as information increases.

The impact of using the penalized likelihood for estimation and inference in random-effects meta-analysis and meta-regression is more profound for a small to moderate number of studies. As the number of studies increases, the log-likelihood derivatives dominate the bias-reducing adjustment in (7) in terms of asymptotic order. As a result, inference based on the penalized likelihood becomes indistinguishable from likelihood inference.

In § 4, the performance of penalized likelihood inference is compared with that of alternative methods under the random-effects meta-analysis model (1), and under a more realistic model for individual-within-study data.

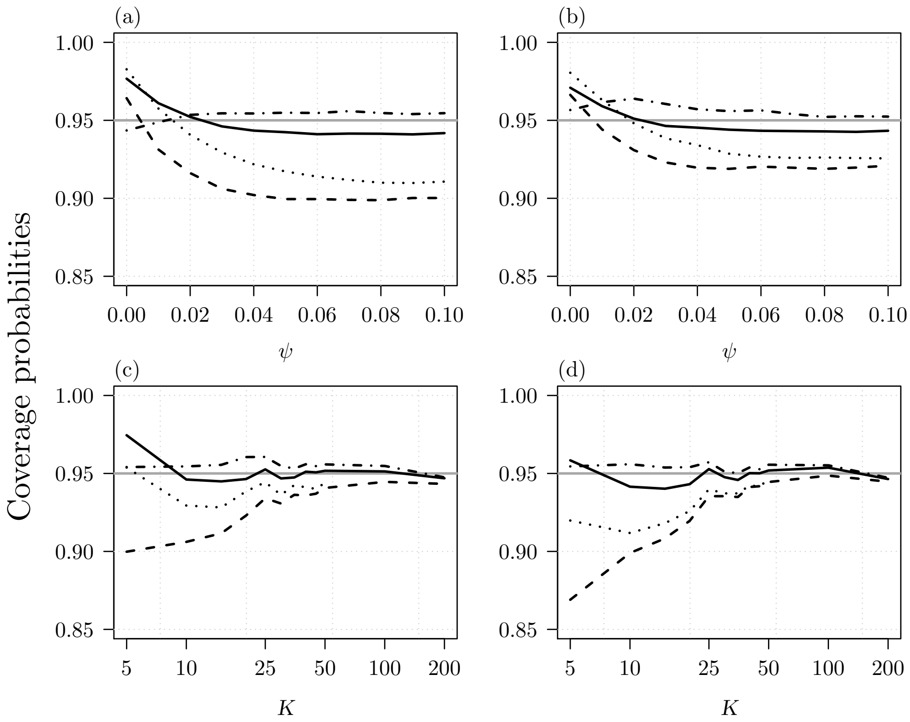

Figure 1: Empirical coverage probabilities of two-sided confidence intervals for β for increasing ψ, when (a) $K = 10$ and (b) $K = 20$, and for increasing K (in log scale) when (c) $\psi = 0.03$ and (d) $\psi = 0.07$. The curves correspond to profile penalized likelihood (solid), DerSimonian & Laird method (dashed), Zeng & Lin double resampling (dotted; available only for $K \leq 50$), and Skovgaard's statistic (dotted-dashed). The grey horizontal line is the target 95% nominal level.

4 Simulation studies

4.1 Random-effects meta-analysis

The simulation studies under the random-effects meta-analysis model (1) are performed using the design in Brockwell & Gordon (2001). Specifically, the study-specific effects y_i are simulated from the random-effect meta-analysis with true effect $\beta = 0.5$ and variance $\hat{\sigma}_i^2 + \psi$, where $\hat{\sigma}_i^2$ are independently generated from a χ_1^2 distribution multiplied by 0.25 and then restricted to the interval $(0.009, 0.6)$. The between-study variance ψ ranges from 0 to 0.1 and the number of studies K from 5 to 200. For each combination of ψ and K considered, 10 000 data sets are simulated using the same initial state for the random number generator.

Zeng & Lin (2015, Section 5) show that their double resampling approach outperforms several existing methods in terms of the empirical coverage probabilities of confidence intervals for β at nominal level 95%. The methods considered in Zeng & Lin (2015) include profile likelihood (Hardy & Thompson, 1996), modified DerSimonian & Laird (see Sidik & Jonkman, 2002; Knapp & Hartung, 2003; Copas, 2003), quantile approximation (Jackson & Bowden, 2009) and the approach described in Henmi & Copas (2010). The present simulation study takes advantage of these previous simulation results, and Figure 1 compares the performance of double resampling with that of the profile penalized likelihood. In order to avoid long computing times, empirical coverage for double-resampling has been calculated only for $K \leq 50$.

The profile penalized likelihood confidence interval has empirical coverage that is appreciably closer to the nominal level than double resampling.

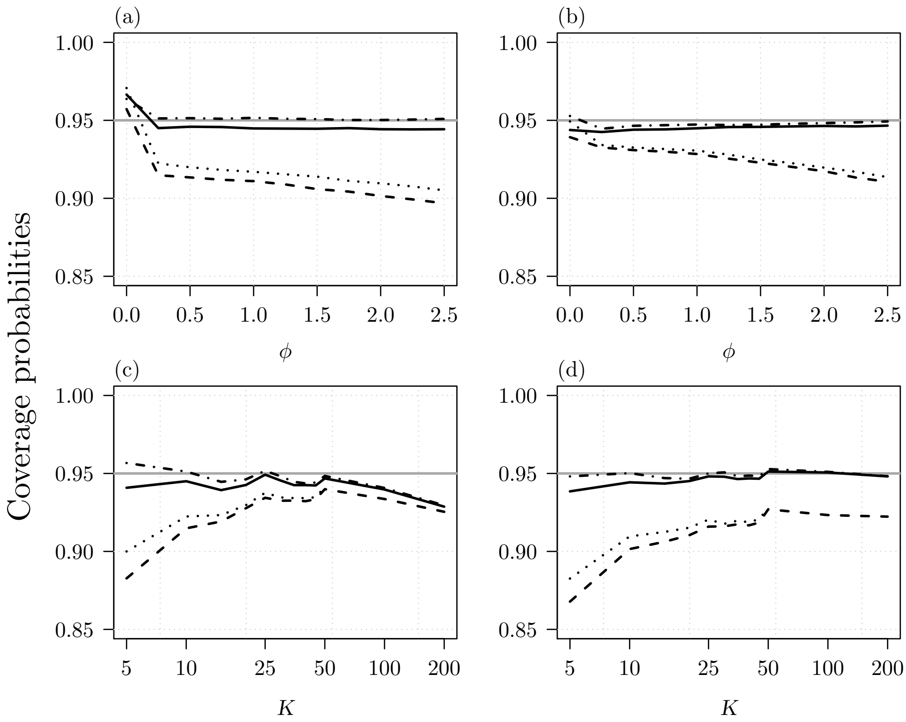

Figure 2: Empirical coverage probabilities of two-sided confidence intervals for δ for increasing ϕ, when (a) $K = 10$ and (b) $K = 35$, and for increasing K (in log scale) when (c) $\phi = 0.25$ and (d) $\phi = 2$. The curves correspond to profile penalized likelihood (solid), DerSimonian & Laird method (dashed), Zeng & Lin double resampling (dotted), and Skovgaard's statistic (dotted-dashed). The grey horizontal line is the target 95% nominal level.

Figure 1 also includes results for two alternative confidence intervals. The first uses the classical DerSimonian & Laird estimator $\hat{\beta}(\psi_{\text{DL}})$ and its estimated variance $1/\sum_{i=1}^{K} 1/(\hat{\sigma}_i+\hat{\psi}_{\text{DL}})$. Not surprisingly, the empirical coverage of this confidence interval is grossly smaller than the nominal confidence level. The second interval is used for reference and results from the numerical inversion of Skovgaard's statistic, which is designed to produce second-order accurate p-values for tests on the mean effect size (Guolo, 2012; Guolo & Varin, 2012). The profile penalized likelihood interval has comparable performance to that based on Skovgaard's statistic, with the latter having empirical coverage slightly closer to the nominal level for a wider range of values for ψ. In general, though, the numerical inversion of Skovgaard's statistic can be unstable due to the discontinuity of the statistic around the maximum likelihood estimator. In contrast, the calculation of profile penalized likelihood intervals is not prone to such instabilities. The penalized likelihood also results in a bias-reduced estimator of ψ, whose reliable estimation is often of interest in medical studies (Veroniki et al., 2016).

4.2 Standardized mean differences from two-arm studies

The profile penalized likelihood and all other methods in Figure 1 have been developed under the validity of the random-effects meta-analysis model. This assumption may be unrealistic, especially in settings where the estimation variances are directly related to the summary measure (Hoaglin, 2015). Here, we examine the performance of the methods under an alternative specification of the data generating process, where the study-specific effects and their variances are calculated by simulating individual-within-study data. Specifically, we assume that the ith study consists of two arms with n_i individuals each, and that $n_1, \dots, n_K$ are independent uniform draws from the integers $\{30, 31, \dots, 100\}$. Then, conditionally on a random effect $\alpha_i \sim N(0, \phi)$, we assume that the observation $z_{i,rj}$ for the jth individual in the rth arm is the realisation of a $N(\mu + I_r(\delta + \alpha_i)\sigma, \sigma^2)$ random variable, where $I_1 = 0$ and $I_2 = 1$. The difference between the marginal variances of the arms increases with ϕ. The true effects are set to $\mu = 0$, $\sigma = 1$ and $\delta = -2$. The study-specific effect of interest is δ, estimated using the standardized mean difference $y_i = J_i(\bar{z}_{i,2} - \bar{z}_{i,1})/s_i$, where s_i^2 is the pooled variance from the two arms of the ith study, and $J_i = 1 - 3/\{8(n_i - 1) - 1\}$ is the Hedges correction (see, e.g., Borenstein, 2009, Chapter 4). The corresponding estimated variance for y_i is $\hat{\sigma}_i^2 = 2J_i/n_i + J_i y_i^2/(4n_i)$, which is a quadratic function of y_i. The between-study variance ϕ ranges from 0 to 2.5 and the number of studies K from 5 to 200. For each combination of ϕ and K considered, 10 000 data sets are simulated using the same initial state for the random number generator.

Figure 2 shows the empirical coverage of the confidence intervals for δ based on the methods that were examined in Figure 1. Empirical coverage for double-resampling has again been calculated only for $K \leq 50$. The good performance of the profile penalized likelihood interval and the interval based on the Skovgaard's statistic persists for small and moderate number of studies, even under the alternative data generating process. The performance of the intervals based on the DerSimonian & Laird estimator and double resampling is, again, poor.

Figure 2 also illustrates the effect of increasing the number of studies under the alternative specification of the data generating process. As the number of studies increases, the inadequacy of the assumptions of the working random-effects meta-regression model becomes more notable. Model mis-specification will eventually result in loss of coverage for all methods examined here, including the intervals based on profile penalized likelihood and the Skovgaard's statistic.

The Supplementary Material provides the full results from this study and two other simulation studies, where the summary measures are log-odds-ratios from a case-control study.

5 Case study: meat consumption data

Larsson & Orsini (2014) investigate the association between meat consumption and relative risk of all-cause mortality. The data include 16 prospective studies, eight of which are about unprocessed red meat consumption and eight about processed meat consumption. We consider meta-regression with a covariate taking value 1 for processed red mean and 0 for unprocessed. The DerSimonian & Laird estimate of ψ is $\hat{\psi}_{\text{DL}} = 0.57 \times 10^{-2}$, the maximum likelihood estimate is $\hat{\psi}_{\text{ML}} = 0.85 \times 10^{-2}$ and the maximum penalized likelihood estimate is the largest with $\hat{\psi}_{\text{MPL}} = 1.18 \times 10^{-2}$. The estimates of β are $\hat{\beta}_{\text{ML}} = (0.10, 0.11)^\top$, $\hat{\beta}_{\text{MPL}} = (0.09, 0.11)^\top$ and $\hat{\beta}_{\text{DL}} = (0.11, 0.10)^\top$, where the first element in each vector corresponds to the intercept and the second to meat consumption.

The DerSimonian & Laird method indicates some evidence for a higher risk associated to the consumption of red processed meat with a p-value of 0.027. In contrast, the penalized likelihood ratio and Skovgaard's statistic suggest that there is rather weak evidence for higher risk, with p-values of 0.066 and 0.073, respectively.

The Supplementary Material contains a simulation study under the maximum likelihood fit that illustrates that the maximum likelihood estimator of ψ is negatively biased. The other estimators almost fully compensate for that bias, but $\hat{\psi}_{\text{MPL}}$ is appreciably more efficient than $\hat{\psi}_{\text{DL}}$. The simulation study therein is also used to illustrate the good performance of the penalized likelihood ratio test in terms of size.

6 Supplementary material

The Supplementary Material provides R (R Core Team, 2016) code to reproduce the case study in § 5, and another analysis. The full results of the simulations are provided including the performance of confidence intervals based on alternative methods.

Acknowledgements

This work was partially supported by grants 'IRIDE', Ca' Foscari University of Venice and 'Progetti di Ricerca di Ateneo 2015', University of Padova. The authors thank David Firth for discussions and feedback on an early version of the paper. The authors are also grateful to two anonymous Referees, the Associate Editor and the Editor, for helpful comments and suggestions.

Appendix A: derivation of the expression for the first-order bias

The first-order bias of $\hat{\theta}_{(ML)}$ has the form $b(\theta) = -\{F(\theta)\}^{-1} A(\theta)$ (Kosmidis & Firth, 2010), where $A(\theta)$ has components

$$A_t(\theta) = -\frac{1}{2}\mathrm{tr}\left[\{F(\theta)\}^{-1}\{P_t(\theta) + Q_t(\theta)\}\right] \quad (t = 1, \ldots, p+1)\,.$$

There, $P_t(\theta) = E_\theta\{s(\theta)s(\theta)^\top s_t(\theta)\}$ and $Q_t(\theta) = E_\theta\{-I(\theta)s_t(\theta)\}$.

The model assumptions imply that $E_\theta\{R_i(\beta)^m\}$ is 0 if m is odd and $(m-1)!!/w_i(\psi)^{m/2}$ if m is even, where $w_i(\psi) = 1/(\hat{\sigma}_i^2 + \psi)$, and $(m-1)!!$ denotes the double factorial of $m-1$ $(m = 1, 2, \ldots; i = 1, \ldots, K)$. Direct matrix calculations give

$$P_t(\theta) + Q_t(\theta) = 0_{(p+1)\times(p+1)} \quad (t = 1, \ldots, p)\,; \quad P_{p+1}(\theta) + Q_{p+1}(\theta) = \begin{bmatrix} X^\top W(\psi)^2 X & 0_p \\ 0_p^\top & 0 \end{bmatrix},$$

where $0_{p\times p}$ is the $p \times p$ zero matrix. So, $A_t(\theta) = 0$ for $t \in \{1, \ldots, p\}$ and

$$A_{p+1}(\theta) = \mathrm{tr}\left[\left\{X^\top W(\psi) X\right\}^{-1} X^\top W(\psi)^2 X\right] = \mathrm{tr}\left\{W(\psi)H(\psi)\right\}\,.$$

Inserting the expressions for the components of $A(\theta)$ into the expression for $b(\theta)$ gives $b(\theta) = \{0_p^\top, b_\psi(\psi)\}^\top$, where $b_\psi(\psi)$ is as in (5).

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2009). *Introduction to Meta-Analysis.* Chichester, West Sussex: Wiley.

Brockwell, S. E. & Gordon, I. N. (2001). A comparison of statistical methods for meta-analysis. *Stat. Med.* **20**, 825–40.

Copas, J. (2003) Letters to the editor: a simple confidence interval for meta-analysis. Sidik K., Jonkman J. N. (2002) *Statistics in Medicine* **21**, 3153–3159. *Stat. Med.* **22**, 2667–68.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc. B* **49**, 1–39.

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–88.

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

GUOLO, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Stat. Med.* **31**, 313–27.

GUOLO, A. & VARIN, C. (2012). The R package metaLik for likelihood inference in meta-analysis. *J. Stat. Softw.* **50**, 1–14.

GUOLO, A. & VARIN, C. (2015). Random-effects meta-analysis: The number of studies matters. *Stat. Methods Med. Res.*, to appear.

HARDY, R. J. & THOMPSON, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Stat. Med.* **15**, 619–29.

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Ass.* **72**, 320–38.

HENMI, M. & COPAS, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Stat. Med.* **29**, 2969–83.

HOAGLIN, D. C. (2015). We know less than we should about methods of meta-analysis. *Res. Syn. Meth.* **6**, 287–9.

VAN HOUWELINGEN, H. C., ARENDS, L. R. & STIJNEN, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat. Med.* **21**, 589–624.

JACKSON, D. & BOWDEN, J. (2009). A re-evaluation of the 'quantile approximation method' for random effects meta-analysis. *Stat. Med.* **28**, 338–48.

KNAPP, G. & HARTUNG, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Stat. Med.* **22**, 2693–710.

KOSMIDIS, I. & FIRTH, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96**, 793–804.

KOSMIDIS, I. & FIRTH, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electron. J. Stat.* **4**, 1097–112.

LARSSON, S. C. & ORSINI, N. (2014) Red meat and processed meat consumption and all-cause mortality: A meta-analysis. *Am. J. Epidemiol.* **179**, 282–89.

R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

SMYTH, G. K. & VERBYLA, A. P. (1996). A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. R. Statist. Soc. B* **58**, 565–72.

STERN, S. E. & WELSH, A. H. (2000). Likelihood inference for small variance components. *Can. J. Stat.* **83**, 319–30.

SIDIK, K. & JONKMAN, J. N. (2002). A simple confidence interval for meta-analysis. *Stat. Med.* **21**, 3153–9.

VERONIKI, A. A., JACKSON, D., VIECHTBAUER, W., BENDER, R., BOWDEN, J., KNAPP, G., KUSS, O., HIGGINS, J. P. T., LANGAN, D., and SALANTI, G. (2016) Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Syn. Meth.* **7**, 55–79.

VIECHTBAUER, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* **30**, 261–93.

ZENG, D. & LIN, D. Y. (2015). On random-effects meta-analysis. *Biometrika* **102**, 281–94.

Supplementary material for

Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression

I. Kosmidis
Department of Statistical Science
University College London
London, WC1E 6BT, U.K
i.kosmidis@ucl.ac.uk

A. Guolo
Department of Statistical Sciences
University of Padova
Via Cesare Battisti, 241/243, 35121 Padova, Italy
annamaria.guolo@unipd.it

C. Varin
Department of Environmental Sciences, Informatics and Statistics,
Ca' Foscari University of Venice
Via Torino 150, 30170 Venezia Mestre, Italy
cristiano.varin@unive.it

January 8, 2017

1 R version, package details and other functions

The current report reproduces the real-data analysis and provides the full results of the simulation studies described in the paper "Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression" by I. Kosmidis, A. Guolo and C. Varin. The report also enriches the paper with an extra case-study and the results from two extra simulation scenarios. The outputs in the current report have been produced using R version 3.3.0 (R Development Core Team, 2016), and the R package `metaLik` version 0.42.0 (Guolo & Varin, 2012).

The file `functionsMPL.R` that accompanies the current report (at the time of writing the file is available from `http://www.ucl.ac.uk/~ucakiko/files/functionsMPL.R`) provides:

- an R function to maximize the penalized likelihood in expression (8) of the paper for general meta-regression settings (see the function `BiasFit`);
- an R function to perform hypothesis tests for the parameters of a meta-regression model using the profiles of the penalized likelihood as in §3.3 of the paper (see the function `lrtest` in `functionsMPL.R`);
- an R implementation of the double resampling approach in Zeng & Lin (2015) for hypothesis testing in meta-analysis; and
- other helper functions for the above.

The following chunk of code loads the required packages and the functions in `functionsMPL.R`:

```
library(metaLik)
library(metatest)
library(parallel)
## REPLACE THIS WITH LINK TO URL
source(url("http://www.ucl.ac.uk/~ucakiko/files/functionsMPL.R"))
```

2 Case studies

2.1 Meat consumption data

The data for the analysis in § 5 of the main text is

```
# Meat consumption data
########################
larsson <- data.frame(logRR = c(-0.3425,  0.2546,  0.1740,  0.1655,
                               -0.0834,  0.0953,  0.2151,  0.3988,
                               0.0488,  0.1484,  0.2231,  0.2390,
                               0.1823,  0.3577,  0.0583,  0.1484),
                         sigma2 = c(0.017224, 0.001271, 0.000663, 0.005027,
                                    0.003383, 0.003603, 0.062186, 0.118504,
                                    0.071613, 0.000310, 0.000501, 0.001160,
                                    0.000759, 0.005087, 0.031266, 0.023078),
                         type = c(rep("non-p", 8), rep("p", 8)))
```

The variable `type` is the meat type (`p` for processed and `non-p` for non-processed), `logRR` is the logarithm of the relative risk of all-cause mortality for the highest versus the lowest consumption category, and `sigma2` is the variance of the logarithm of the relative risk.

The code chunk below fits the random-effects meta-regression model with response `logRR`, explanatory variable `type` and summary variances `sigma2`. The model includes an intercept parameter β_1, the parameter β_2 for `type` and the heterogeneity parameter `psi`. The model parameters are estimated using maximum likelihood, the DerSimonian & Laird estimator, and maximum penalized likelihood.

```
m1 <- metaLik(logRR ~ type, data = larsson, sigma2 = sigma2)
estimates1 <- BiasFit(m1)
estimates1 <- with(estimates1, data.frame(ML = ML[1:3], DL = DL[1:3], MPL = MPL[1:3]))
rownames(estimates1)[3] <- "psi"
round(estimates1, 4)

##                 ML     DL    MPL
## (Intercept) 0.0994 0.1060 0.0947
## typep       0.1064 0.1004 0.1098
## psi         0.0085 0.0057 0.0118
```

The following code chunk calculates the p-value for testing $\beta_2 < 0$ using the DerSimonian & Laird method, the penalized likelihood ratio and the Skovgaard's statistic.

```
pvalue1_dl <- pnorm(m1$DL[2] / sqrt(m1$vcov.DL[2, 2]), lower.tail = FALSE)
pvalue1_pd <- lrtest(m1, what = 2, type = "penloglik", null = 0.0, optMethod = "BFGS",
                     alternative = "greater")$pvalue
pvalue1_Skovgaard <- test.metaLik(m1, param = 2, value = 0, print = FALSE,
                                  alternative = "greater")$pvalue.rskov
pvalues1 <- c(pvalue1_dl, pvalue1_pd, pvalue1_Skovgaard)
names(pvalues1) <- c("DerSimonian Laird", "Penalized likelihood", "Skovgaard")
round(pvalues1, 3)

##    DerSimonian Laird Penalized likelihood            Skovgaard
##                0.027                0.066                0.073
```

The chunk of code below simulates 10 000 independent samples of the 16 logarithms of relative risks under the maximum likelihood fit `m1`, conditionally on `type` and `sigma2`. For each simulated sample, ψ is estimated using maximum likelihood, maximum penalized likelihood and the DerSimonian & Laird estimator. In addition, p-values are computed for the hypothesis $\beta_2 < 0$, using the DerSimonian & Laird statistic, the likelihood ratio statistic, the penalized likelihood ratio statistic and Skovgaard's statistic.

```
# Number of cores to be used for the simulations (does /not/ work in Windows machines)
ncores <- 4
nsimu <- 10000
simudata <- simulate(m1, nsim = nsimu, seed = 123)
tau2ind <- length(coef(m1)) + 1
typep <- coef(m1)[2]
results <- mclapply(seq.int(ncol(simudata)), function(i) {
    mod <- update(m1, data = within(larsson, logRR <- simudata[, i]))
    out <- BiasFit(mod)
    estimates <- data.frame(estimate = with(out, c(ML[tau2ind], DL[tau2ind], MPL[tau2ind])),
                            method = c("ML", "DL", "MPL"))
    pvalues <- perform_tests(y = simudata[, i], X = mod$X, sigma2 = mod$sigma2,
                             what = 2, null = typep, B = 1)
    pvalues <- data.frame(pvalue = pvalues[c("DL", "LR", "PLR", "Skovgaard"), "pvalues_g"])
    pvalues$method <- c("DerSimonian Laird", "Likelihood", "Penalized likelihood", "Skovgaard")
    list(estimates = estimates, pvalues = pvalues)
}, mc.cores = ncores)
```

Figure 3 shows boxplots of the estimators of ψ calculated from the 10 000 simulated samples. The dashed lines correspond to the parameters values used for the simulation and the point inside each box is the average of the estimates for the corresponding method. As expected from expression (5) of the main text, the maximum likelihood estimator of ψ is negatively biased, while the other estimators almost fully compensate for that bias. The distribution of the DerSimonian & Laird estimator of ψ appears to have a heavier right tail than the maximum penalized likelihood estimator, which links to the findings of past studies on the loss of efficiency of the former (e.g., Viechtbauer, 2005).

The simulated samples are also used below to calculate the empirical p-value distribution (%) for the tests based on the DerSimonian & Laird statistic, the likelihood ratio statistic, the penalized likelihood ratio statistic and Skovgaard's statistic. The empirical p-value distribution for the penalized likelihood ratio statistic and Skovgaard's statistic are markedly closer to uniform than for the other methods.

```
pvalues <- do.call("rbind", lapply(results, function(x) x$pvalues))
pvalues$method <- factor(pvalues$method)
alphas <- c(1.0, 2.5, 5.0, 10.0, 25.0, 50.0, 75.0, 90.0, 95.0, 97.5, 99.0)/100
distr <- with(pvalues, {
    sizes <- sapply(alphas, function(alpha) {
        tapply(pvalue, method, function(ps) mean(ps < alpha))
    })
    colnames(sizes) <- format(alphas * 100, digits = 2)
    round(sizes * 100, 1)
})
distr

##                       1.0  2.5  5.0 10.0 25.0 50.0 75.0 90.0 95.0 97.5 99.0
## DerSimonian Laird     2.5  4.5  7.6 12.4 26.6 49.9 73.4 87.6 92.3 95.2 97.3
## Likelihood            2.1  4.5  7.6 13.0 27.9 50.1 71.7 86.6 92.2 95.3 97.7
## Penalized likelihood  1.3  3.0  5.6 11.0 25.9 49.9 73.8 89.0 94.2 96.9 98.6
## Skovgaard             1.1  2.7  5.0 10.2 25.2 49.9 74.6 89.7 94.9 97.4 98.9
```

2.2 Local anesthesia data

Ambulatory hysteroscopy is a useful instrument to diagnose intrauterine pathologies. Cooper et al. (2010) perform a meta-analysis about the efficacy of different types of local anesthesia

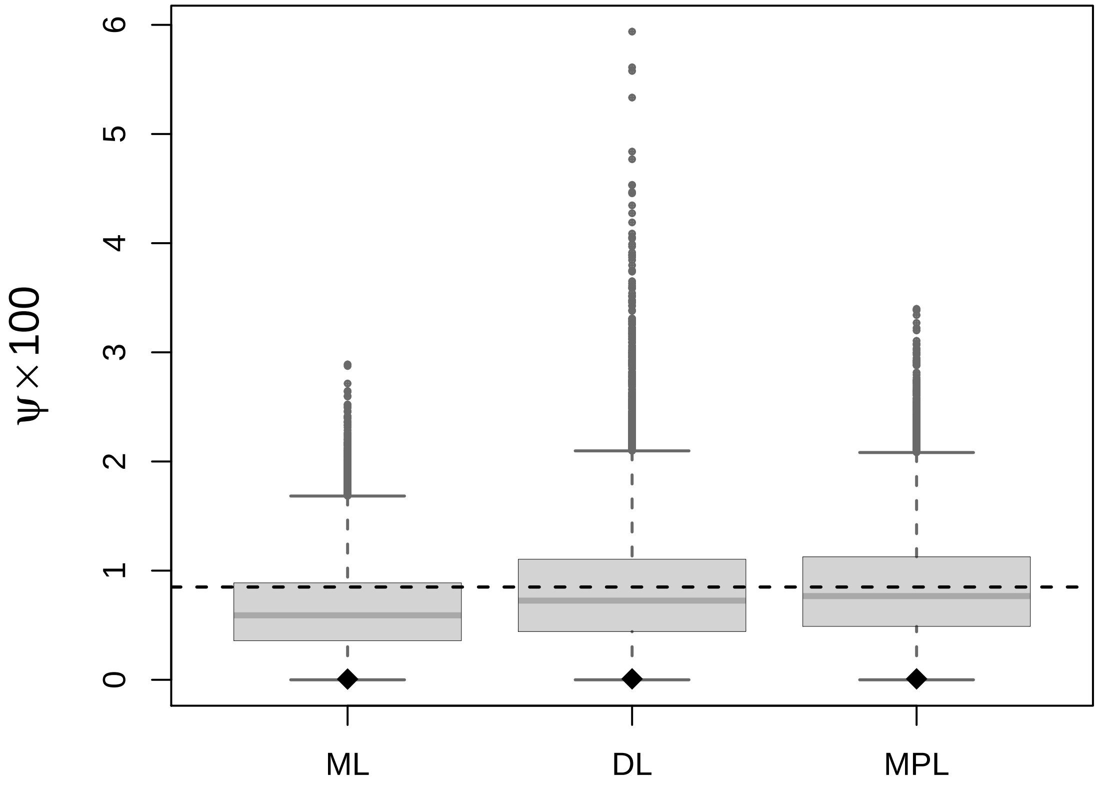

Figure 3: Boxplot for the DerSimonian and Laird (DL), maximum likelihood (ML) and maximum penalized likelihood (MPL) estimators of ψ as calculated from 10 000 simulated samples under the maximum likelihood fit in the case study of §5. The horizontal dashed line is the parameter value used for the simulation and the point inside each box is the average of the estimates for the corresponding method.

used to control pain during hysteroscopy. The following code chunks provide the analysis for an investigation on the use of paracervical anesthesia.

The available data consist of 5 standardized mean differences (`size`) of pain scores measured at the time of hysteroscopy from five randomized controlled trials, and the corresponding estimated variances (`sigma2`).

```
cooper <- data.frame(y = c(0.00, -1.71, -0.19, -0.58, -4.27),
                     sigma2 = c(0.03959288, 0.07731804, 0.02265332, 0.01759683, 0.16040842))
```

The code chunk below fits the random-effects meta-analysis model with vector of responses `y` and vector of summary variances `sigma2`. The model includes the overall standardized mean difference β and the heterogeneity parameter `psi`, which are estimated using maximum likelihood, the DerSimonian & Laird estimator, and maximum penalized likelihood.

```
m2 <- metaLik(y ~ 1, data = cooper, sigma2 = sigma2)
estimates2 <- BiasFit(m2)
estimates2 <- with(estimates2, data.frame(ML = ML[1:2], DL = DL[1:2], MPL = MPL[1:2]))
rownames(estimates2)[2] <- "psi"
round(estimates2, 4)

##                  ML      DL     MPL
## (Intercept) -1.3168 -1.2829 -1.3236
## psi          2.3055  1.0808  2.9273
```

The DerSimonian & Laird estimate of ψ is 1.08, while the maximum likelihood estimate is 2.31, which is appreciably larger. The maximum penalized likelihood estimate takes the even larger value 2.93.

The following code chunk calculates the p-value for testing $\beta = 0$ using the DerSimonian & Laird method, the double resampling method of Zeng & Lin (2015), the penalized likelihood ratio statistic and the Skovgaard's statistic.

```
pvalue2_dl <- 2*pnorm(-abs(m2$DL / sqrt(m2$vcov.DL)))
pvalue2_dr <- double.resampling(0.0,  m2, B = 1000, myseed = 123)
pvalue2_pd <- lrtest(m2, what = 1, type = "penloglik", null = 0.0, optMethod = "BFGS")$pvalue
pvalue2_Skovgaard <- test.metaLik(m2, param = 1, value = 0, print = FALSE)$pvalue.rskov
pvalues2 <- c(pvalue2_dl, pvalue2_dr, pvalue2_pd, pvalue2_Skovgaard)
names(pvalues2) <- c("DerSimonian Laird", "Double resampling", "Penalized likelihood", "Skovgaard")
round(pvalues2, 3)

##    DerSimonian Laird    Double resampling Penalized likelihood
##                0.007                0.018                0.137
##            Skovgaard
##                0.158
```

The DerSimonian & Laird method supports the effectiveness of paracervical local anesthesia with a p-value of 0.007, as does the double resampling approach with a p-value of 0.018. The opposite conclusion is obtained using the penalized likelihood ratio statistic, which returns a p-value equal to 0.137. The Skovgaard's statistic confirms the result, with a p-value of 0.158.

3 Simulation studies

3.1 Random-effects meta-analysis

Figure 4 and Figure 5 show the full results from the simulation study in § 4.1 of the paper. Specifically, Figure 4 and Figure 5 include empirical coverage probabilities of two-sided confidence intervals for increasing values of ψ for all $K \in \{5, 10, \ldots, 50, 100, 200\}$. The empirical coverage for the double resampling method is computed only for $K \leq 50$ because of the long computing times involved. The left column of each figure shows the empirical coverage for the confidence intervals examined in Figure 1 of the main text. The right column shows the empirical coverage of confidence intervals based on the Wald statistic, the profile likelihood-ratio statistic, and the Bartlett-corrected likelihood-ratio statistic (see, for example, Huizenga et al., 2011).

3.2 Standardized mean differences from two-arm studies

Figure 6 and Figure 7 show the full results from the simulation study in § 4.2 of the paper. Specifically, Figure 6 and Figure 7 include empirical coverage probabilities of two-sided confidence intervals for increasing values of ϕ for all $K \in \{5, 10, \ldots, 50, 100, 200\}$. The layout of the results in Figure 6 and Figure 7 is the same as that in Figure 4 and Figure 5.

3.3 Case-control study

The current subsection complements § 4 of the main text with results from an additional simulation study under a more realistic data generating process than the working random-effects meta-analysis model. We assume that the ith study consists of n_i individuals, and that $n_1, \ldots, n_K$ are independent uniform draws from the integers $\{30, 31, \ldots, 100\}$. Then, conditionally on random effects u_{i1} and u_{i2}, and covariates $x_{i1}, \ldots, x_{in_j}$, we assume that the individual measurements $z_{i1}, \ldots, z_{in_i}$ in the ith study are realizations of independent Bernoulli random variables with probabilities $\exp(\eta_{ij})/\{1 + \exp(\eta_{ij})\}$, where $\eta_{ij} = \beta_0 + u_{i1} + (\beta_1 + u_{i2})x_{ij}$ $(j = 1, \ldots, n_i)$. The random effects are assumed to be independent with u_{it} having a $N(0, v_t)$ distribution. The covariates are realizations of independent Bernoulli random variables with probability p. The study-specific effect y_i is, then, the estimate of γ_2 from a logistic regression with linear predictor $\gamma_1 + \gamma_2 x_{ij}$ using data $(z_{i1}, x_{i1})^\top, \ldots, (z_{in_i}, x_{in_i})^\top$. The corresponding estimated variance $\hat{\sigma}_i^2$ is based on the evaluation of the expected information matrix at the estimates. The above setting has been inspired by the simulation study in Abo-Zaid et al. (2013, Appendix B). In order to avoid the incidence of infinite estimates, the estimation of the study-specific effects is carried out using the `brglm` R package (Kosmidis, 2013).

We use $\upsilon_1 = 0.1$, $p = 0.5$ and consider a set of "moderate" true effects where $\beta_0 = -1.27$ and $\beta_1 = 0.9$, and another set of "larger" effects where $\beta_0 = -2$ and $\beta_1 = 1.5$. The between-study variance υ_2 ranges from 0 to 2.5, and the number of studies K ranges from 5 to 200. For each combination of υ_2 and K, 10 000 data sets are simulated. The random number generator is initialised to have the same state for each combination of υ_2 and K.

Figure 8 and Figure 9 show the empirical coverage of various 95% confidence intervals for β_1 under the set of moderate true effects ($\beta_0 = -1.27$ and $\beta_1 = 0.9$) for increasing values of υ_2 for all $K \in \{5, 10, ..., 50, 100, 200\}$. Again, the empirical coverage for the double resampling method is computed only for $K \leq 50$. Figure 10 and Figure 11 are the corresponding plots for the set of larger true effects ($\beta_0 = -2$ and $\beta_1 = 1.5$).

The good performance of the confidence intervals based on the profile penalized likelihood, Skovgaard's statistic and the Bartlett-corrected likelihood-ratio statistic is apparent for small to moderate number of studies. In contrast, the intervals based on the DerSimonian & Laird estimator, the double bootstrap, the profile likelihood and the Wald statistic illustrate poor performance, particularly for small values of K. As in Figure 7, Figure 9 and Figure 11 illustrate that as the number of studies grows, model mis-specification eventually results in loss of coverage for all methods examined here, including the one based on the penalized likelihood. The loss of coverage is most severe for small values of the between-variance component υ_2. The intervals based on the DerSimonian & Laird method seem to be the ones most severely affected by model mis-specification.

References

Abo-Zaid, G., Guo, B., Deeks, J. J., Debray, T. P., Steyerberg, E. W., Moons, K. G. & Riley, R. D. (2013). Individual participant data meta-analyses should not ignore clustering. *J. Clin. Epidemiol.* **66**, 865–73.

Cooper, N. A. M., Khan, K. S. & Clark T. J. (2010). Local anaesthesia for pain control during outpatient hysteroscopy: systematic review and meta-analysis. *Brit. Med. J.* **340**, c1130.

Guolo, A. & Varin, C. (2012). The R package metaLik for likelihood inference in meta-analysis. *J. Stat. Softw.* **50**, 1–14.

Huizenga, H. M., Visser, I. & Dolan, C. V. (2011), Testing overall and moderator effects in random effects meta-regression. *Brit. J. Math. Stat. Psy.* **64**, 1–19.

Kosmidis, I. (2013) brglm: Bias reduction in binary-response Generalized Linear Models. R package version: 0.5-9, http://www.ucl.ac.uk/ ucakiko/software.html

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* **30**, 261–93.

Zeng, D. & Lin, D. Y. (2015). On random-effects meta-analysis. *Biometrika* **102**, 281–94.

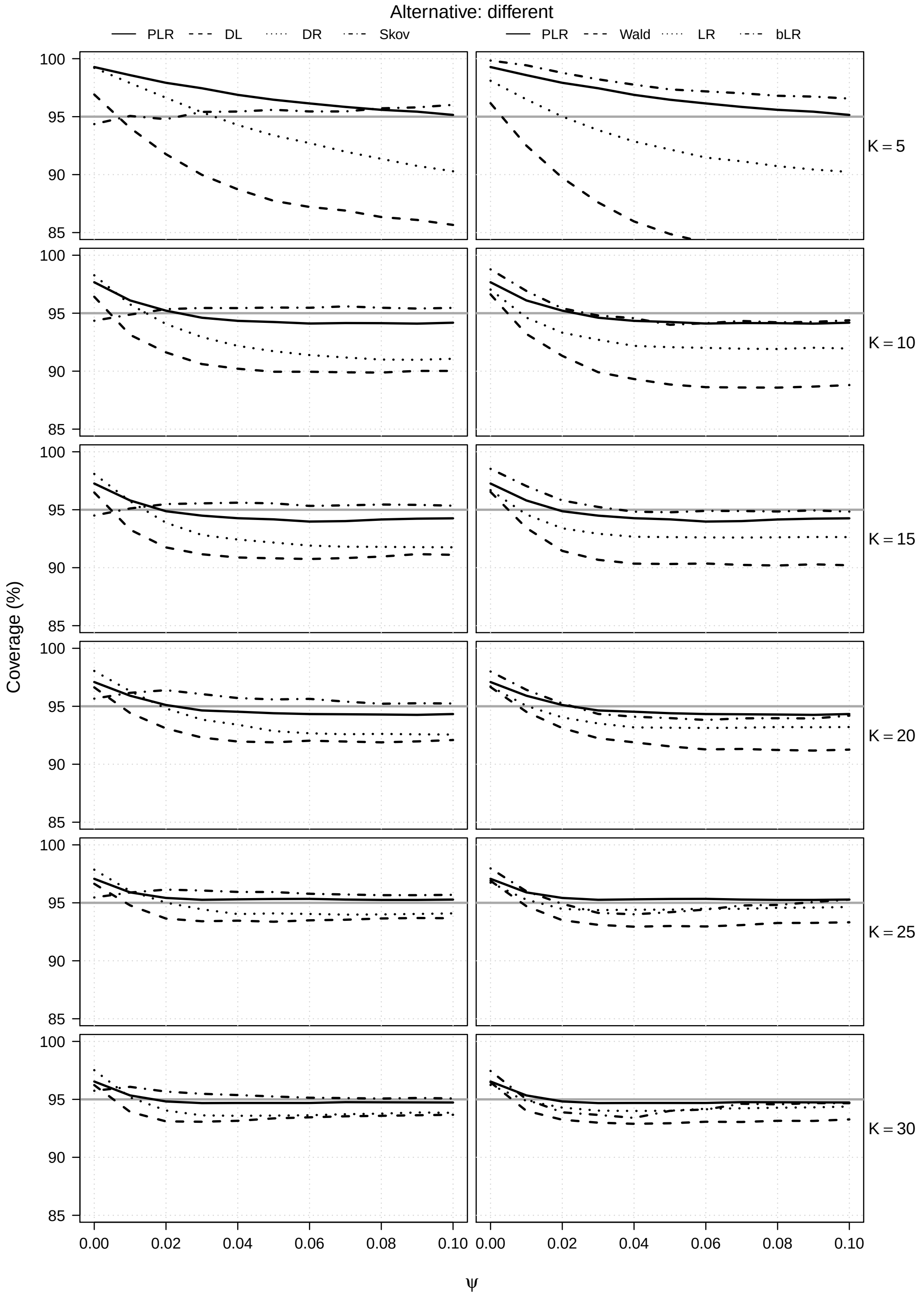

Figure 4: Empirical coverage probabilities of two-sided confidence intervals for random-effects meta-analysis. The empirical coverage is calculated for increasing values of ψ and for $K \in \{5, 10, 15, 20, 25, 30\}$. The curves correspond to the proposed penalized likelihood method (PLR), the DerSimonian & Laird method (DL; left column), the Zeng & Lin double resampling method (DR; left column), the Skovgaard statistic (Skov; left column), the Wald statistic (Wald; right column), the likelihood-ratio statistic (LR; right column), and the Bartlett-corrected likelihood-ratio statistic (bLR; right column). The grey horizontal line is the target 95% nominal level.

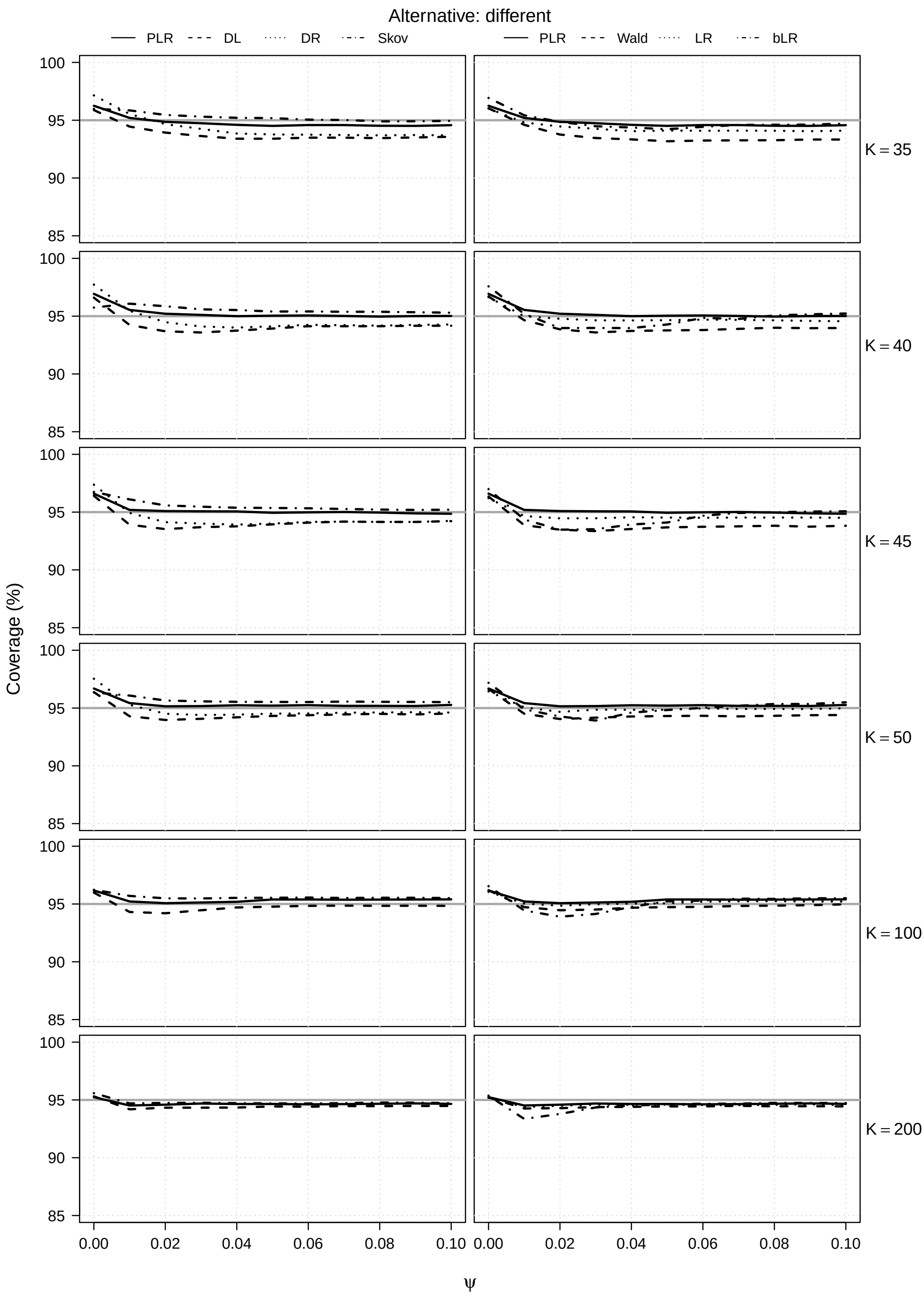

Figure 5: Empirical coverage probabilities of two-sided confidence intervals for random-effects meta-analysis. The empirical coverage is calculated for increasing values of ψ and for $K \in \{35, 40, 45, 50, 100, 200\}$. The curves correspond to the proposed penalized likelihood method (PLR), the DerSimonian & Laird method (DL; left column), the Zeng & Lin double resampling method (DR; left column), the Skovgaard statistic (Skov; left column), the Wald statistic (Wald; right column), the likelihood-ratio statistic (LR; right column), and the Bartlett-corrected likelihood-ratio statistic (bLR; right column). The grey horizontal line is the target 95% nominal level.

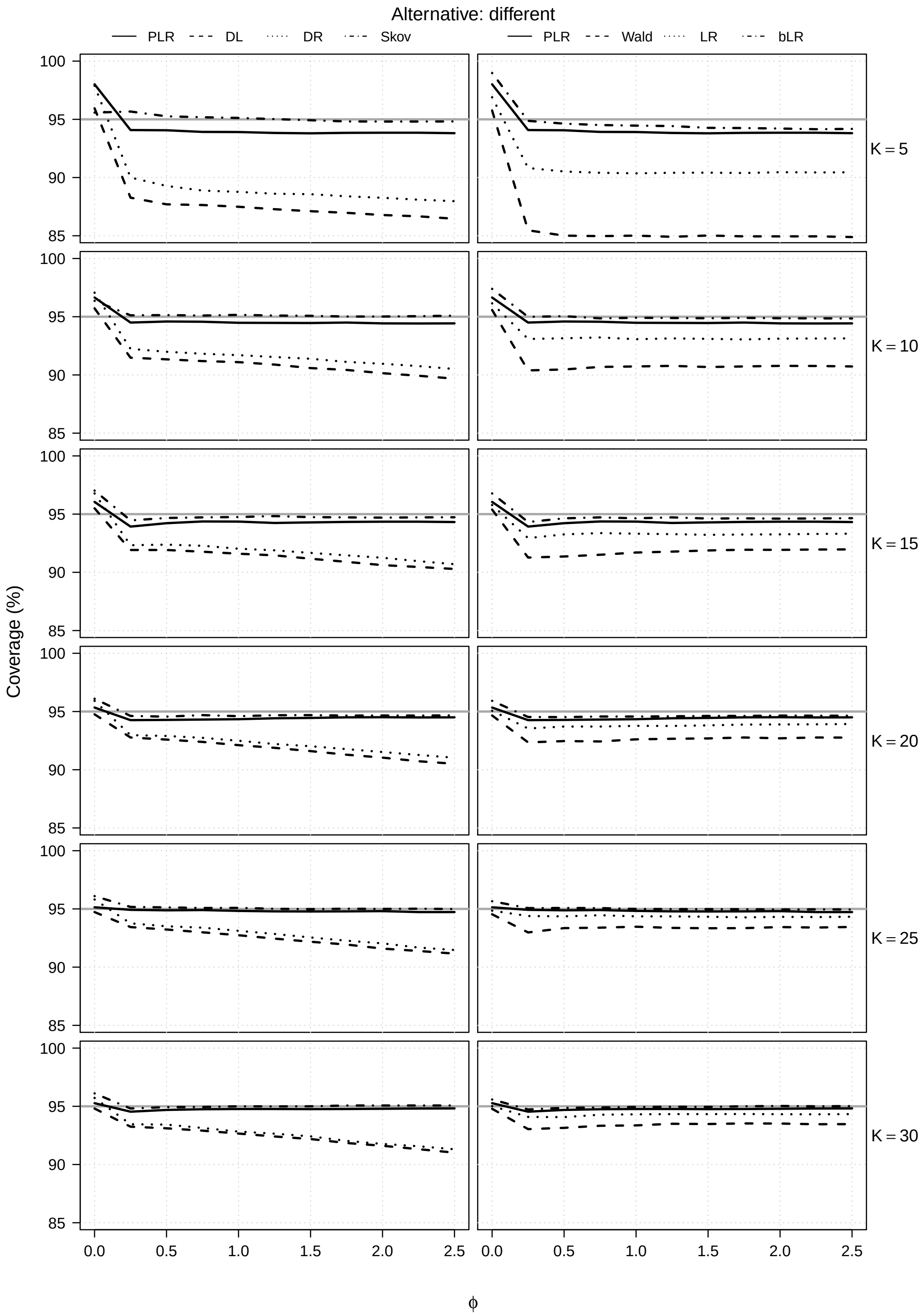

Figure 6: Empirical coverage probabilities of two-sided confidence intervals for standardized mean differences from two-arm studies. The empirical coverage is calculated for increasing values of ϕ and for $K \in \{5, 10, 15, 20, 25, 30\}$. The curves correspond to the proposed penalized likelihood method (PLR), the DerSimonian & Laird method (DL; left column), the Zeng & Lin double resampling method (DR; left column), the Skovgaard statistic (Skov; left column), the Wald statistic (Wald; right column), the likelihood-ratio statistic (LR; right column), and the Bartlett-corrected likelihood-ratio statistic (bLR; right column). The grey horizontal line is the target 95% nominal level.

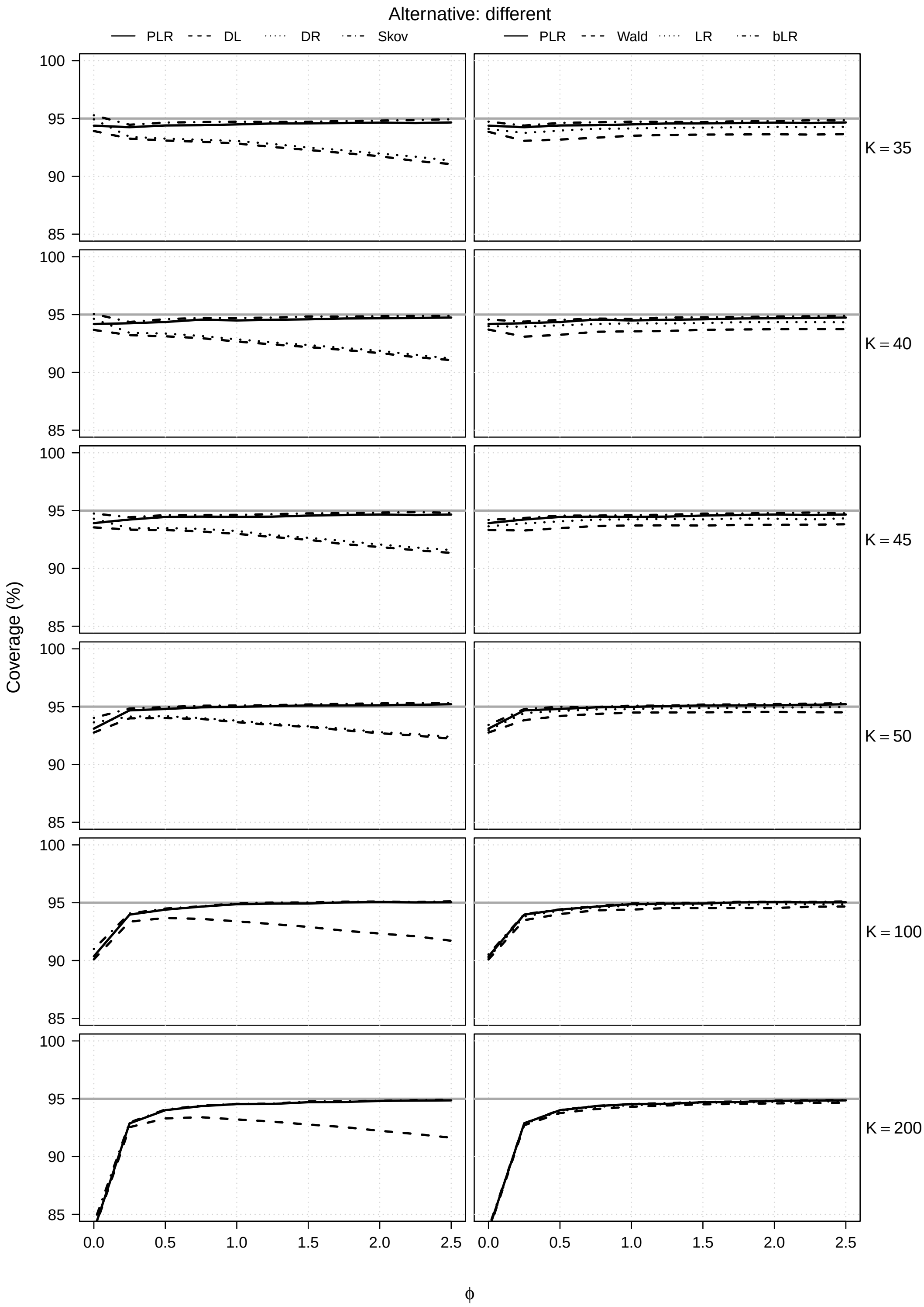

Figure 7: Empirical coverage probabilities of two-sided confidence intervals for standardized mean differences from two-arm studies. The empirical coverage is calculated for increasing values of ϕ and for $K \in \{35, 40, 45, 50, 100, 200\}$. The curves correspond to the proposed penalized likelihood method (PLR), the DerSimonian & Laird method (DL; left column), the Zeng & Lin double resampling method (DR; left column), the Skovgaard statistic (Skov; left column), the Wald statistic (Wald; right column), the likelihood-ratio statistic (LR; right column), and the Bartlett-corrected likelihood-ratio statistic (bLR; right column). The grey horizontal line is the target 95% nominal level.

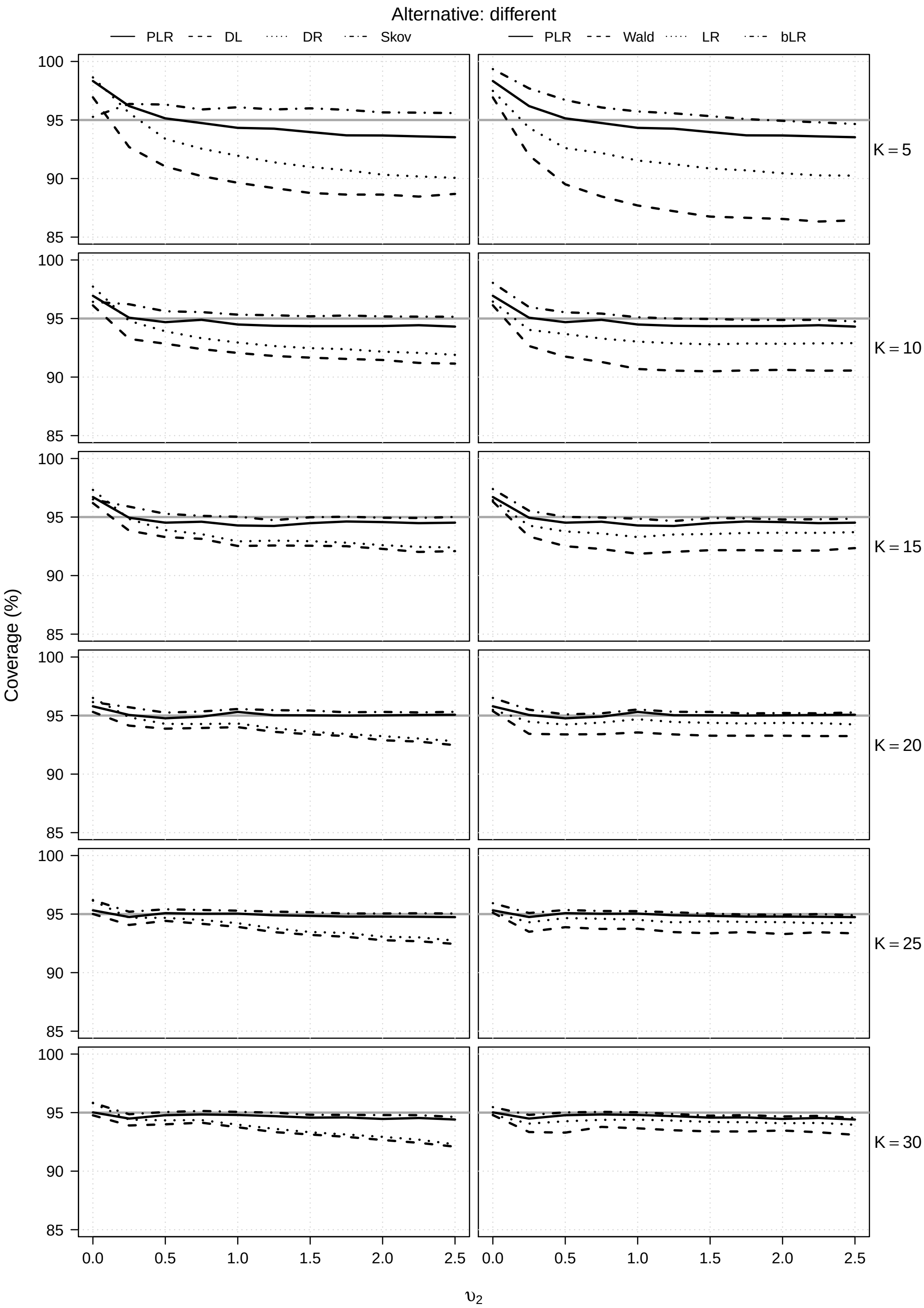

Figure 8: Empirical coverage probabilities of two-sided confidence intervals for log-odds ratios from a case-control study design with moderate effects. The empirical coverage is calculated for increasing values of υ_2 and for $K \in \{5, 10, 15, 20, 25, 30\}$. The curves correspond to the proposed penalized likelihood method (PLR), the DerSimonian & Laird method (DL; left column), the Zeng & Lin double resampling method (DR; left column), the Skovgaard statistic (Skov; left column), the Wald statistic (Wald; right column), the likelihood-ratio statistic (LR; right column), and the Bartlett-corrected likelihood-ratio statistic (bLR; right column). The grey horizontal line is the target 95% nominal level.

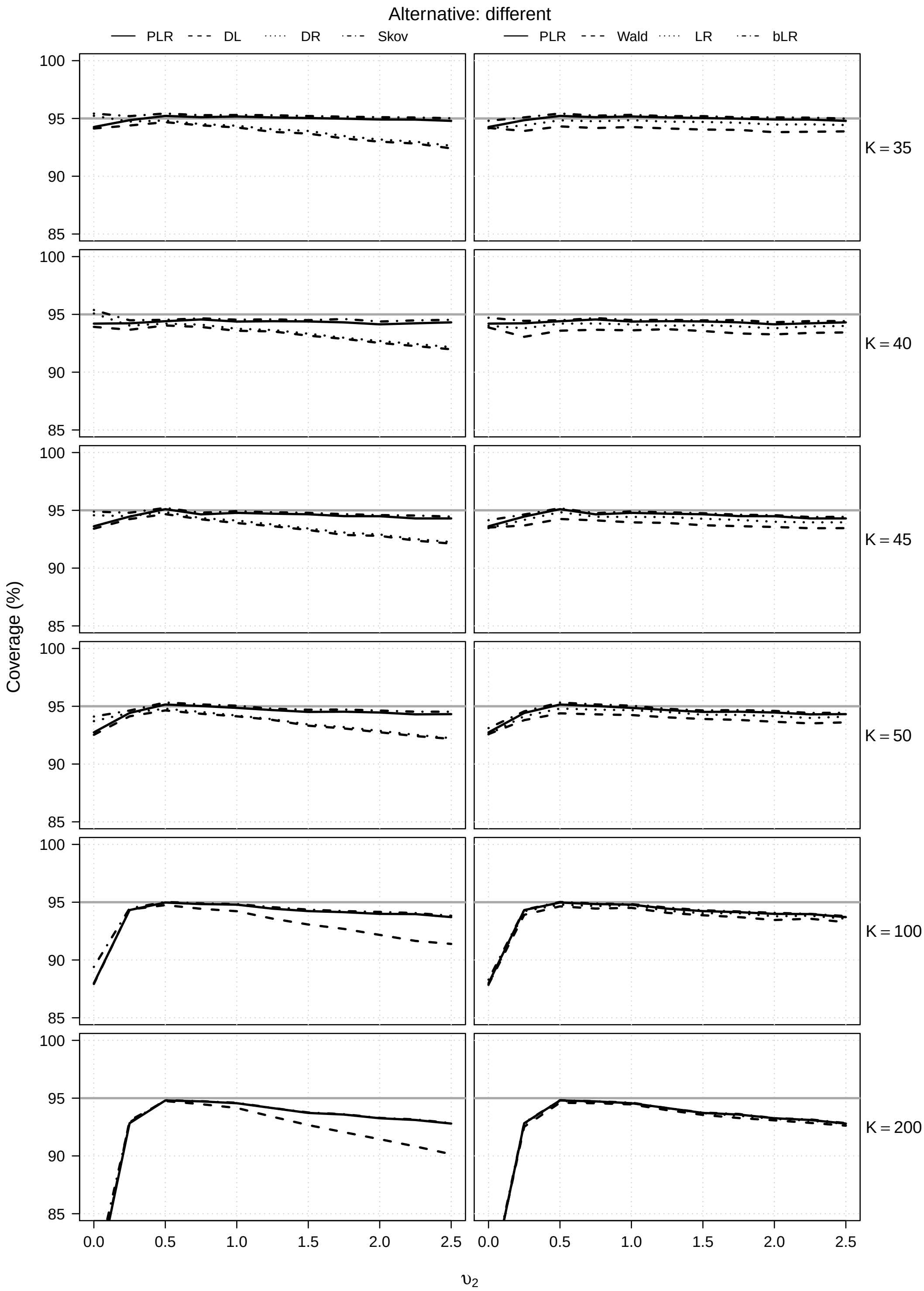

Figure 9: Empirical coverage probabilities of two-sided confidence intervals for log-odds ratios from a case-control study design with moderate effects. The empirical coverage is calculated for increasing values of υ_2 and for $K \in \{35, 40, 45, 50, 100, 200\}$. The curves correspond to the proposed penalized likelihood method (PLR), the DerSimonian & Laird method (DL; left column), the Zeng & Lin double resampling method (DR; left column), the Skovgaard statistic (Skov; left column), the Wald statistic (Wald; right column), the likelihood-ratio statistic (LR; right column), and the Bartlett-corrected likelihood-ratio statistic (bLR; right column). The grey horizontal line is the target 95% nominal level.

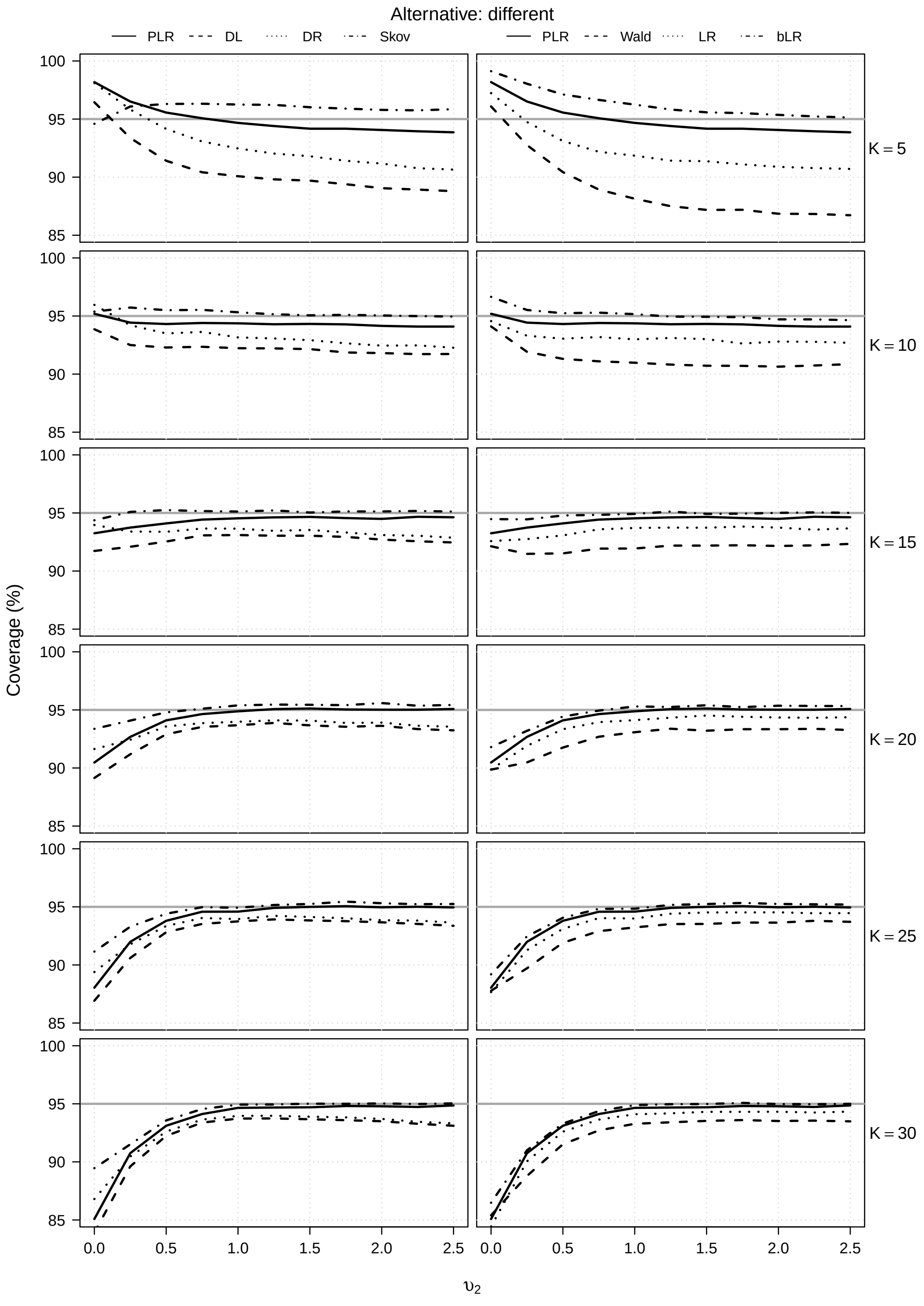

Figure 10: Empirical coverage probabilities of two-sided confidence intervals for log-odds ratios from a case-control study design with larger effects. The empirical coverage is calculated for increasing values of υ_2 and for $K \in \{5, 10, 15, 20, 25, 30\}$. The curves correspond to the proposed penalized likelihood method (PLR), the DerSimonian & Laird method (DL; left column), the Zeng & Lin double resampling method (DR; left column), the Skovgaard statistic (Skov; left column), the Wald statistic (Wald; right column), the likelihood-ratio statistic (LR; right column), and the Bartlett-corrected likelihood-ratio statistic (bLR; right column). The grey horizontal line is the target 95% nominal level.

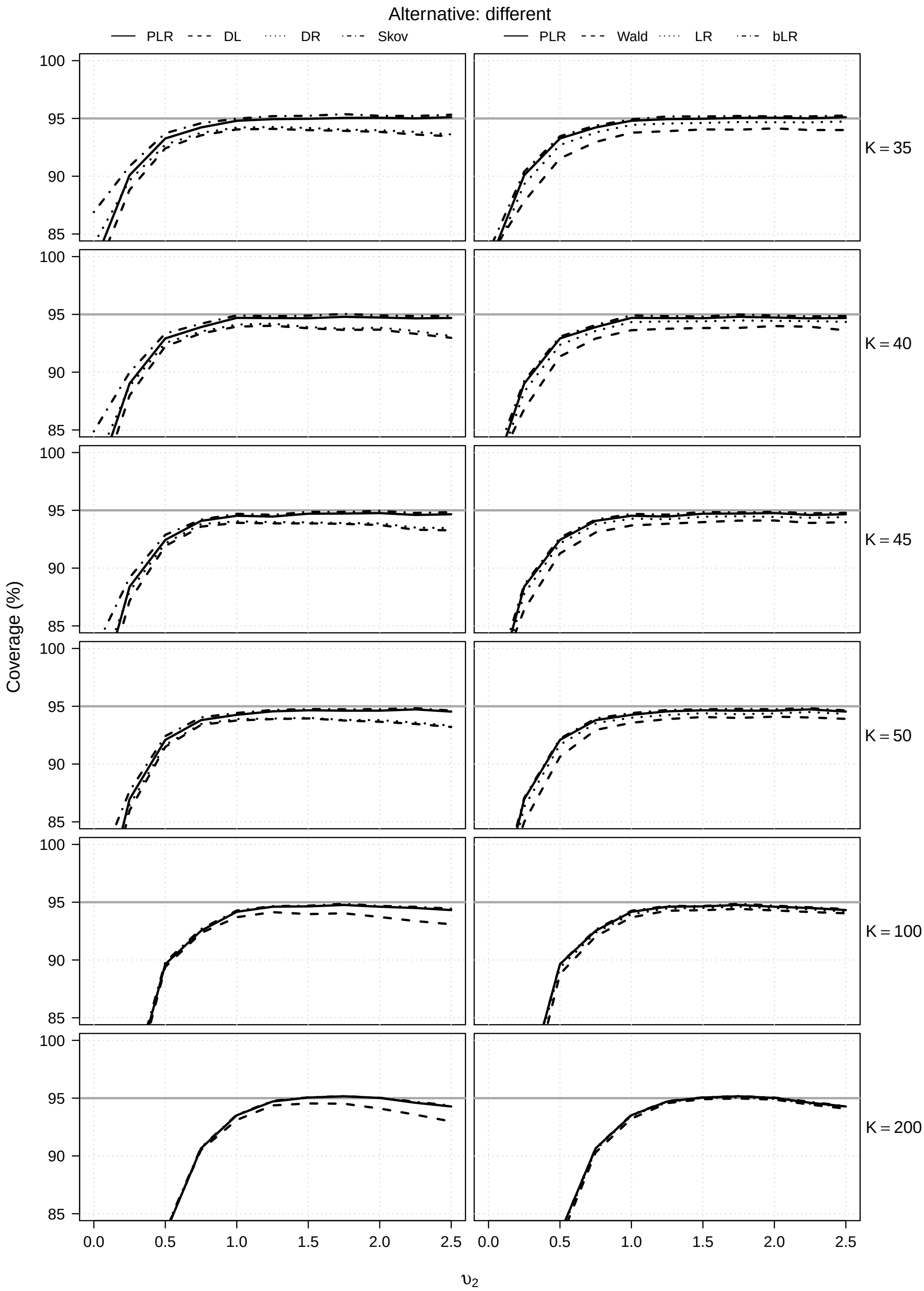

Figure 11: Empirical coverage probabilities of two-sided confidence intervals for log-odds ratios from a case-control study design with larger effects. The empirical coverage is calculated for increasing values of υ_2 and for $K \in \{35, 40, 45, 50, 100, 200\}$. The curves correspond to the proposed penalized likelihood method (PLR), the DerSimonian & Laird method (DL; left column), the Zeng & Lin double resampling method (DR; left column), the Skovgaard statistic (Skov; left column), the Wald statistic (Wald; right column), the likelihood-ratio statistic (LR; right column), and the Bartlett-corrected likelihood-ratio statistic (bLR; right column). The grey horizontal line is the target 95% nominal level.