
Article:
Laud, P.J. orcid.org/0000-0002-3766-7090 (2017) Equal-tailed confidence intervals for comparison of rates. Pharmaceutical Statistics, 16 (5). pp. 334-348. ISSN 1539-1604

https://doi.org/10.1002/pst.1813

Supporting information related to "Equal-tailed confidence intervals for comparison of rates"

Peter J Laud

APPENDIX S1
COMPUTATIONAL DETAILS OF ROOT-FINDING METHODS, INDETERMINATE AND NON-MONOTONIC SCORE FUNCTIONS

This appendix contains full details of a number of minor technical issues that were omitted from the main manuscript for the sake of brevity.

S1.1 Root-finding

The bisection root-finding method has an obvious difficulty for RR and OR, which are unbounded above. This can be surmounted by mapping the $[0, \infty]$ interval onto a finite interval (such as $[-1, 1]$ for consistency with the calculations for RD) using an inverse tangent transformation. A similar solution can be used for Poisson RD, for which θ is also unbounded below. The alternative root-finding method using the secant method can be unreliable because it is not guaranteed to converge.

S1.2 Problematic score functions

I do not wish to hide the unsettling fact that there are some circumstances where the skewness-corrected score function $Z(\theta)$ in Equation (2) is non-monotonic, and even for the uncorrected score it can be indeterminate at a single point value for θ, both of which can lead to issues with root-finding.

For example, when the skewness correction is omitted, $Z(\theta)$ can be indeterminate at one of the four corners of the parameter space (such as at $\theta = 0$ when $\widehat{p}_1 = 0$ and $\widehat{p}_2 = 0$). Note that this is not unique to the parameterisation presented here, but occurs for both the Miettinen-Nurminen and Gart-Nam versions of the score function. In order to avoid a root-finding algorithm failure in such a case, a simple solution is to set $Z(\theta)$ to zero when $S(\theta)$ is zero.

When the skewness correction is included, the score function is non-monotonic for binomial RD when both $\widehat{p}_1 = 0$ or 1 and $\widehat{p}_2 = 0$ or 1, and in a range of other boundary cases when one group is much larger than the other. Similarly for Poisson RD when $\widehat{p}_1 = 0$ or $\widehat{p}_2 = 0$. For RR and OR, there may be no solution to one of the equations $Z(\theta) = \pm z$, in which case the corresponding limit would naturally be 0 or $+\infty$ as appropriate. Non-monotonic score functions for all contrasts are also induced by the continuity correction. The bisection root-finding algorithm for identifying confidence limits is generally unaffected by the non-monotonic behaviour of the score function, as long as the formulation in Equation (3) is used instead of Equation (2). However, identification of the point estimate (i.e. where $Z(\theta) = 0$) can be problematic, as there may be more than one solution. It is curious to note that in general the point estimate can be shifted by the skewness corrections (this feature can be confirmed by setting a confidence level close to zero), so it may not always be appropriate just to use the crude point estimate. However, this seems the only sensible option for RD when $\widehat{p}_1 = 0$ and $\widehat{p}_2 = 0$. For the continuity-corrected interval, the point estimate should be calculated without the correction.

For the Gart-Nam parameterisation, the issue with indeterminate scores is more widespread and less trivial, in particular resulting in the solution not being found for one of the confidence limits for any case where $\widehat{p}_1 = 0$ or 1, even without the skewness correction term.

S1.3 Double-zero cells

For the stratified RD method with IVS weights, if there is a stratum containing no events (or no non-events) on both arms, 0.5 may be added to each cell, only to avoid the stratum being given 100% weight. For RR and OR, a stratum with no events may be excluded altogether[1] (similarly for any stratum containing 100% events for OR), because it essentially contributes no information to the estimate of θ.

(Objections to this practice have recently been made.[2,3] However, the claim that 'double-zero studies point to no differences in treatment effects' is only true for RD, not for RR and OR. It may be true that a double-zero trial shows that the event rate in both arms is low, but not that their ratio is close to unity: the point estimate for such a trial is indeterminate. For Poisson RR, and similarly for binomial RR with rare events, the absence of information in double-zero strata can be crudely demonstrated by doubling the magnitude of the denominators, and observing no change in the confidence interval. Having said that, there may therefore be a counter-argument in favour of including double-zero trials somehow if their sample size allocation ratio differs greatly from that of the other trials.)

In the unlikely event that no events occurred whatsoever in any strata, then there is no information about RR or OR, so the confidence interval would be $(0, \infty)$.

APPENDIX S2
'CONTINUITY CORRECTED' AND OTHER CONSERVATIVE METHODS

There may be circumstances in which it is desirable to select a 'strictly conservative' confidence interval, which constrains coverage probabilities to be strictly above the nominal confidence level at all times. This logically implies the more stringent requirement for one-sided non-coverage probabilities to be strictly below $\alpha/2$. For such applications, a number of computationally intensive so-called 'exact' methods are available, which are beyond the scope of this paper.

'Exact' methods are also often thought to be necessary when sample sizes are small, perhaps for fear that violation of the assumptions underlying the asymptotic methods could result in severely erroneous type I error rates. This is certainly true of the **AN** methods, but it is found not to be the case for **SCAS**, which actually becomes more conservative (but not excessively so) with smaller sample sizes (not shown).

When large sample sizes preclude the use of an 'exact' method, an approximation may be used to achieve conservative coverage, using a 'continuity correction', in which case a very small degree of under-coverage might be considered acceptable.

It has been said that to describe such an adjustment as a continuity 'correction' is something of a misnomer, as is the term 'exact'.[4] To some extent, this depends on one's preference regarding strictly conservative versus proximate coverage. In any case, it is very important to note that some 'exact' methods, and many 'continuity corrected' methods, fail to achieve strictly conservative *one-sided* coverage,[5] and 'continuity corrected' methods can also have regions of very poor two-sided coverage.[6]

A new continuity corrected **SCAS** method ('**SCAS-cc**') for both binomial and Poisson RD can be obtained by using a modified score function: $S'(\theta) = S(\theta) - \text{sign}(S(\theta)) \times cc$ where $cc = \gamma(\min(n_1, n_2))^{-1}$, with $\gamma = 0.5$ based on Hauck and Anderson.[7]

For RR and OR, there is a common practice of adding 0.5 to cell counts, which was originally designed to reduce bias due to infinite observed values.[8,9] Often this is used selectively in meta-analyses, within strata where empty cells occur (otherwise the offending strata are given zero weight). This adjustment is often referred to as a 'continuity correction',[10] but this is quite misleading, as it does not have the usual effect of boosting coverage probabilities above the nominal level: this 'correction' would perhaps be better labelled as a 'sparse data adjustment', to distinguish it from the continuity correction by Yates.[11] The sparse data adjustment avoids uninformative $(0, \infty)$ intervals, but it does not generally enhance one-sided or two-sided coverage (not shown).

For a proper continuity correction for RR, the adjustment for **SCAS-cc** described above produces uneven coverage, and a tentatively suggested alternative is $cc = \gamma(1/n_1 + \theta/n_2)$.

A continuity correction for OR is possible by reinstating the Cornfield correction[12] to the alternative formulation given in Miettinen and Nurminen[13] (p.217). With the formulae specified here, this adjustment is represented by

$$cc = \gamma\left[\frac{1}{n_1\widetilde{p}_1(1-\widetilde{p}_1)} + \frac{1}{n_2\widetilde{p}_2(1-\widetilde{p}_2)}\right].$$

For the single proportion case, $cc = \gamma/n$.

In each case, the conventional value for γ is 0.5, but if very rare under-coverage is tolerable, then a compromise may be reached between over-conservative coverage and minimal under-coverage, by using a smaller γ such as 0.25.

An alternative for achieving the 'minimum coverage' criterion might be to apply the MOVER approach using 'exact' confidence intervals for the individual group rates (e.g. Clopper-Pearson for binomial rates[14] or Garwood intervals for Poisson rates,[15] which can be obtained as quantiles of a Beta or Gamma distribution respectively, as described in Section 2.2, but using $a_i = 0$, $b_i = 1$ for the lower limit, and $a_i = 1$, $b_i = 0$ for the upper limit). This '**MOVER-E**' method is quite successful, although it narrowly fails to achieve the strict minimum coverage criterion for RD. Note that **MOVER-J** and **MOVER-E** are both based on intervals using Beta or Gamma parameters of the form $a_i = 0.5 - \gamma$, $b_i = 0.5 + \gamma$ for the lower limit for p_i, and $a_i = 0.5 + \gamma$, $b_i = 0.5 - \gamma$ for the upper limit (i.e. $\gamma = 0$ for **MOVER-J** and $\gamma = 0.5$ for **MOVER-E**). This leads to a

more adaptable '**MOVER-cc**' method which, like **SCAS-cc**, allows a compromise to be reached using an intermediate value of γ.

Selected surface plots in Appendix S3.4 demonstrate the potential of the **SCAS-cc** and **MOVER-cc** methods for researchers wishing to align minimum coverage with the nominal confidence level, either strictly or with a limited degree of under-coverage. In these examples, with $\gamma = 0.5$ strictly conservative one-sided coverage is achieved for **SCAS-cc** but not **MOVER-cc**. The 'compromise' methods with $\gamma = 0.25$ are conservative in the vast majority of the parameter space, but slightly less so for RD using the MOVER method. In all cases the corresponding two-sided coverage is above the nominal value in more than 99.5% of the parameter space.

Further exploration of the coverage properties of both of the above methods with different values of γ at different sample sizes is a subject for further research, as is the question of continuity correction for stratified methods.

APPENDIX S3
EXTENDED GRAPHICAL EVALUATION FOR THE SINGLE STRATUM CASE

For reference, the precise calculation of RNCP for each chosen sample size pair n_1, n_2 proceeds as follows at every parameter space point (p_1, p_2):

(1) Identify the set of 'observable' integer values of X_i for the given p_i, $i = 1, 2$. For the Poisson case, and if evaluating a subset of the binomial parameter space, for practical purposes this is restricted to the upper 99.999th percentile of the relevant probability distribution. For the full binomial parameter space it is simply $X_i \in \{0, 1, ..., n_i\}$.

(2) Calculate the bivariate probability of every observable outcome pair (X_1, X_2) as:

$\binom{n_1}{X_1} p_1^{X_1}(1-p_1)^{n_1-X_1}\binom{n_2}{X_2} p_2^{X_2}(1-p_2)^{n_2-X_2}$ for the binomial case, or

$(p_1 n_1)^{X_1} e^{-p_1 n_1}(p_2 n_2)^{X_2} e^{-p_2 n_2}(X_1!X_2!)^{-1}$ for the Poisson case (the 'lfactorial' function in R is useful here when X_i is large).

(3) Calculate the $100(1-\alpha)\%$ intervals for every observable outcome, and ascertain in each case whether the upper limit falls below the true value of θ (i.e. $p_1 - p_2$ for RD, p_1/p_2 for RR, etc). Incomputable **AN** intervals for RR and OR are considered to have an infinite upper limit.

(4) Sum the probabilities from (2) for all outcomes that satisfy (3).

After this exercise has been repeated for every parameter space point, the moving average RNCP is calculated as the mean of the RNCP at all points within a square 'window' $[p_1 \pm \xi, p_2 \pm \xi]$, where ξ may be varied depend on the desired amount of surface smoothing.

The surface plots in the following sections provide a more extensive evaluation in support of those contained in the main paper.

Section S3.1 displays one-sided non-coverage probabilities for sample sizes of $(n_1, n_2) = (30, 30)$, $(45, 15)$, and $(100, 100)$, first for binomial and Poisson RD, then binomial and Poisson RR, and lastly OR. Plots with $(n_1, n_2) = (50, 150)$ are also included, because with p_1 and p_2 transposed they can be interpreted as an assessment of *left-sided* non-coverage for $(n_1, n_2) = (150, 50)$.

Section S3.2 explores the coverage in the extreme lower left portion of the parameter space, for assessment of the methods for the analysis of rare events. Some of the Bayesian methods are compared in Section S3.3, continuity corrected methods for more conservative coverage are plotted in Section S3.4, and intervals for the single rate are evaluated in Section S3.5.

S3.1 Primary methods under different sample size conditions

Binomial and Poisson Rate Difference: $(n_1, n_2) = (30, 30)$

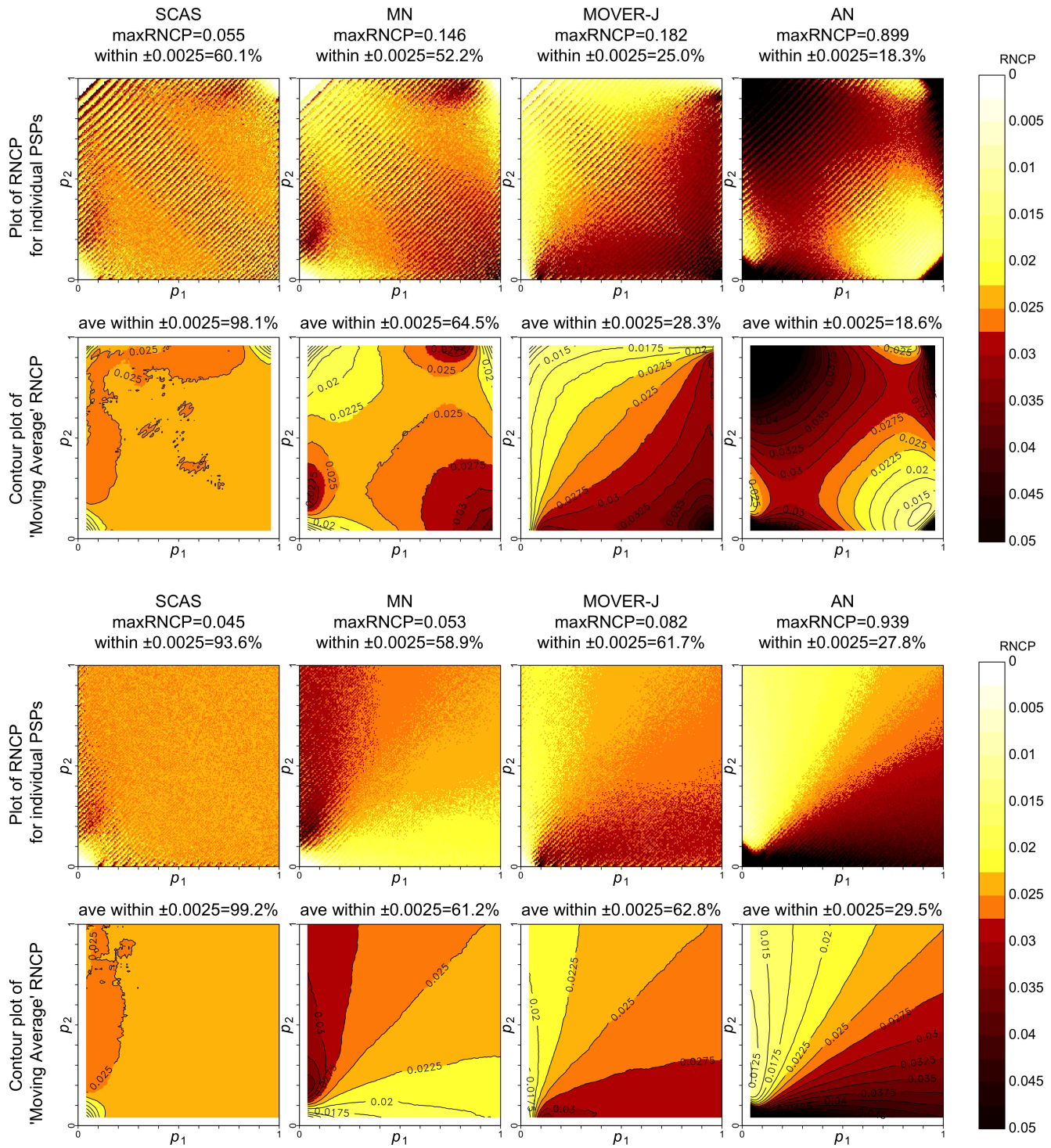

Figure S1. Rate Difference: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 30$, $n_2 = 30$. Top: Binomial, Bottom: Poisson.

Binomial and Poisson Rate Difference: $(n_1, n_2) = (45, 15)$

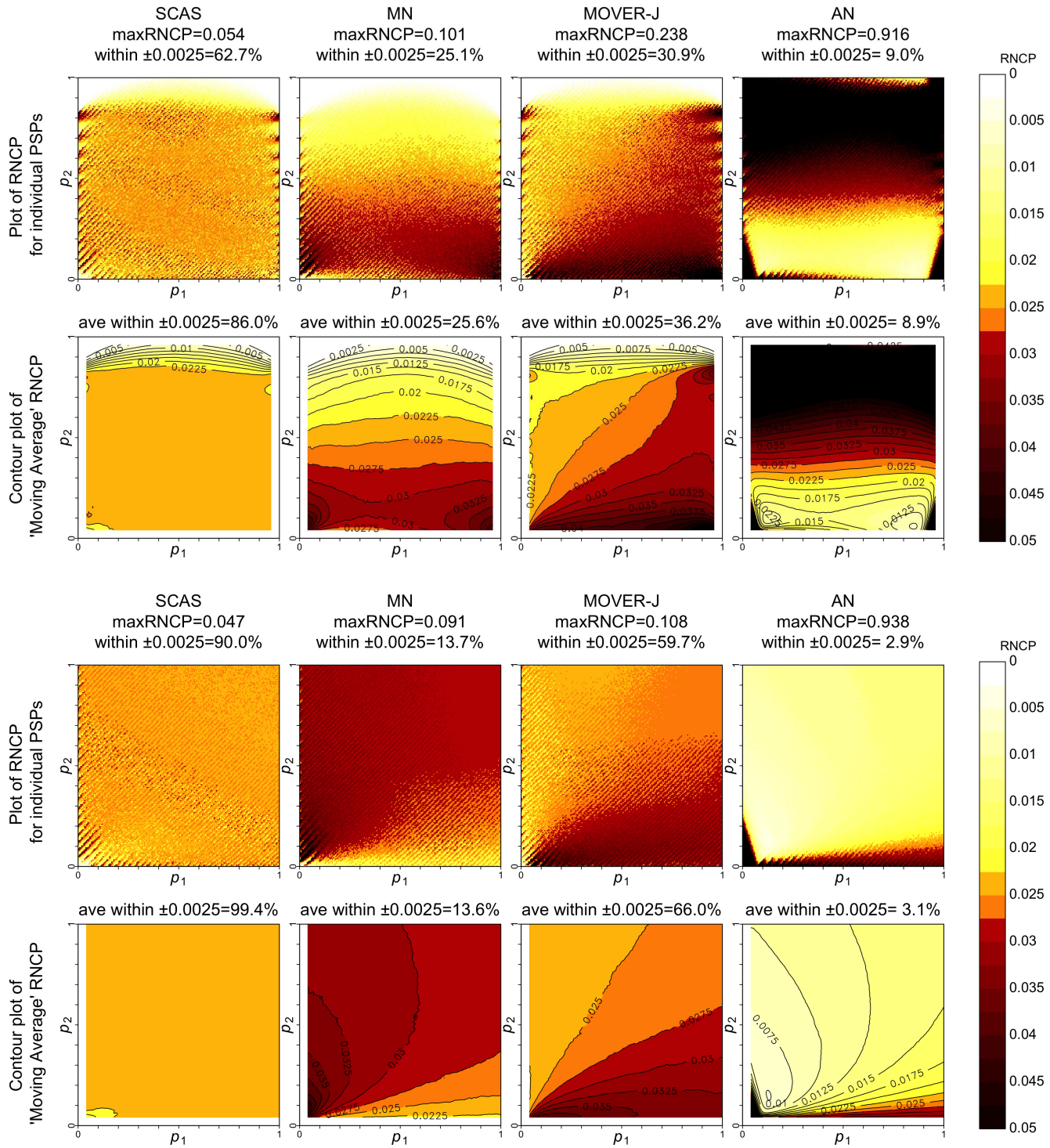

Figure S2. Rate Difference: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 45$, $n_2 = 15$. Top: Binomial, Bottom: Poisson.

Binomial and Poisson Rate Difference: $(n_1, n_2) = (100, 100)$

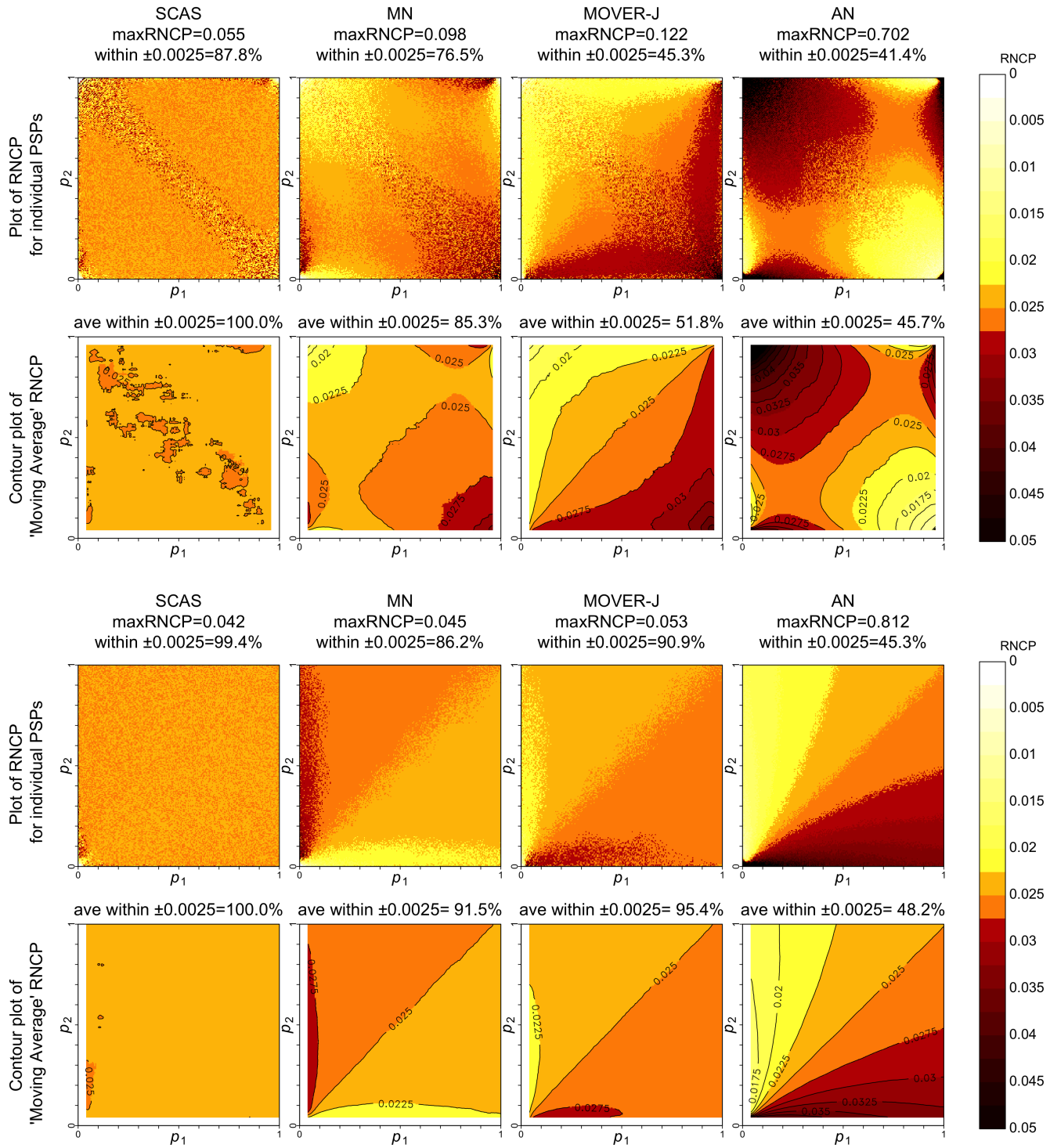

Figure S3. Rate Difference: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 100$, $n_2 = 100$. Top: Binomial, Bottom: Poisson.

Binomial and Poisson Rate Difference: $(n_1, n_2) = (50, 150)$

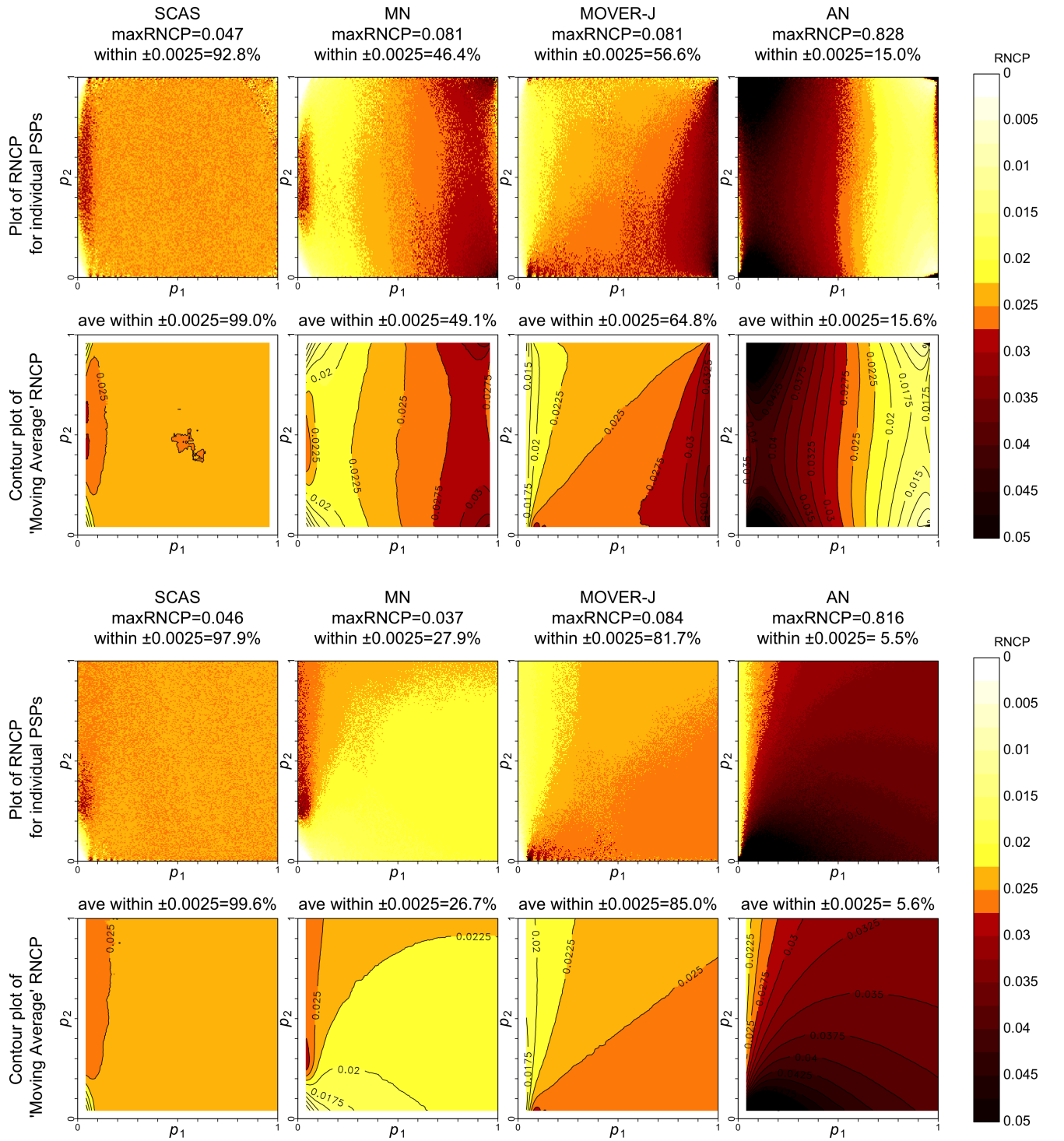

Figure S4. Rate Difference: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 50$, $n_2 = 150$. Top: Binomial, Bottom: Poisson.

Binomial and Poisson Rate Ratio: $(n_1, n_2) = (30, 30)$

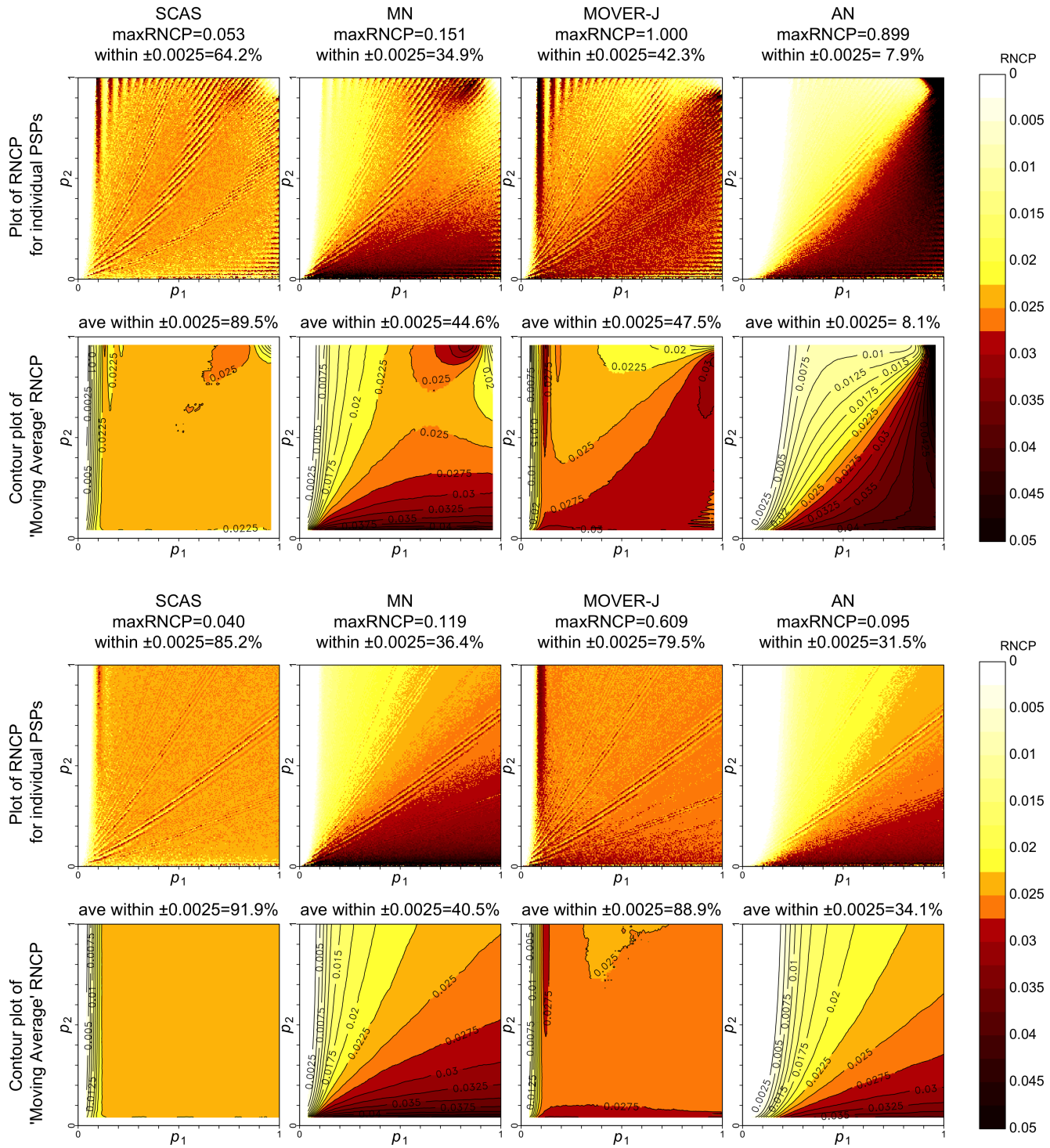

Figure S5. Rate Ratio: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 30$, $n_2 = 30$. Top: Binomial, Bottom: Poisson.

Binomial and Poisson Rate Ratio: $(n_1, n_2) = (45, 15)$

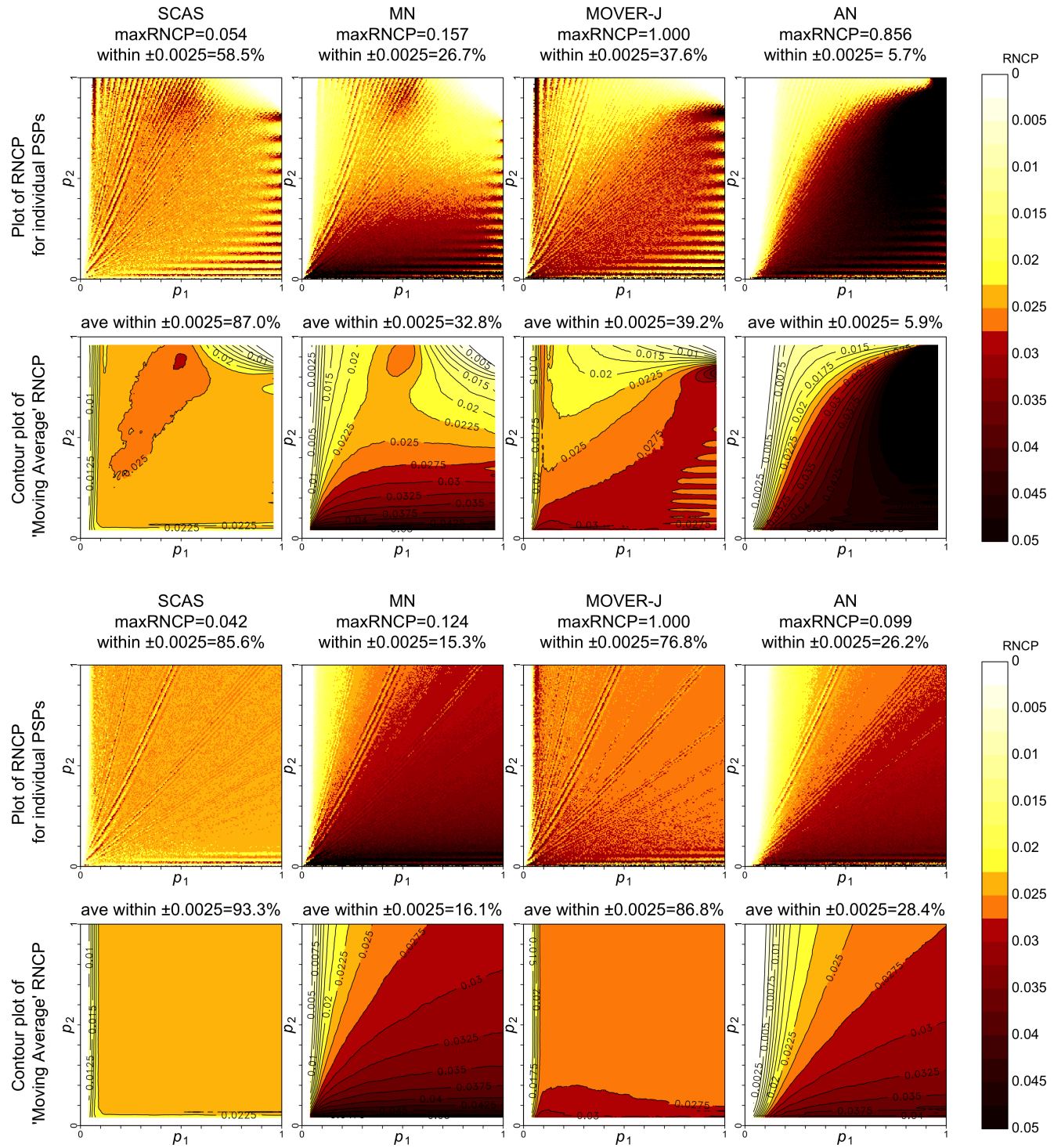

Figure S6. Rate Ratio: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 45$, $n_2 = 15$. Top: Binomial, Bottom: Poisson.

Binomial and Poisson Rate Ratio: $(n_1, n_2) = (100, 100)$

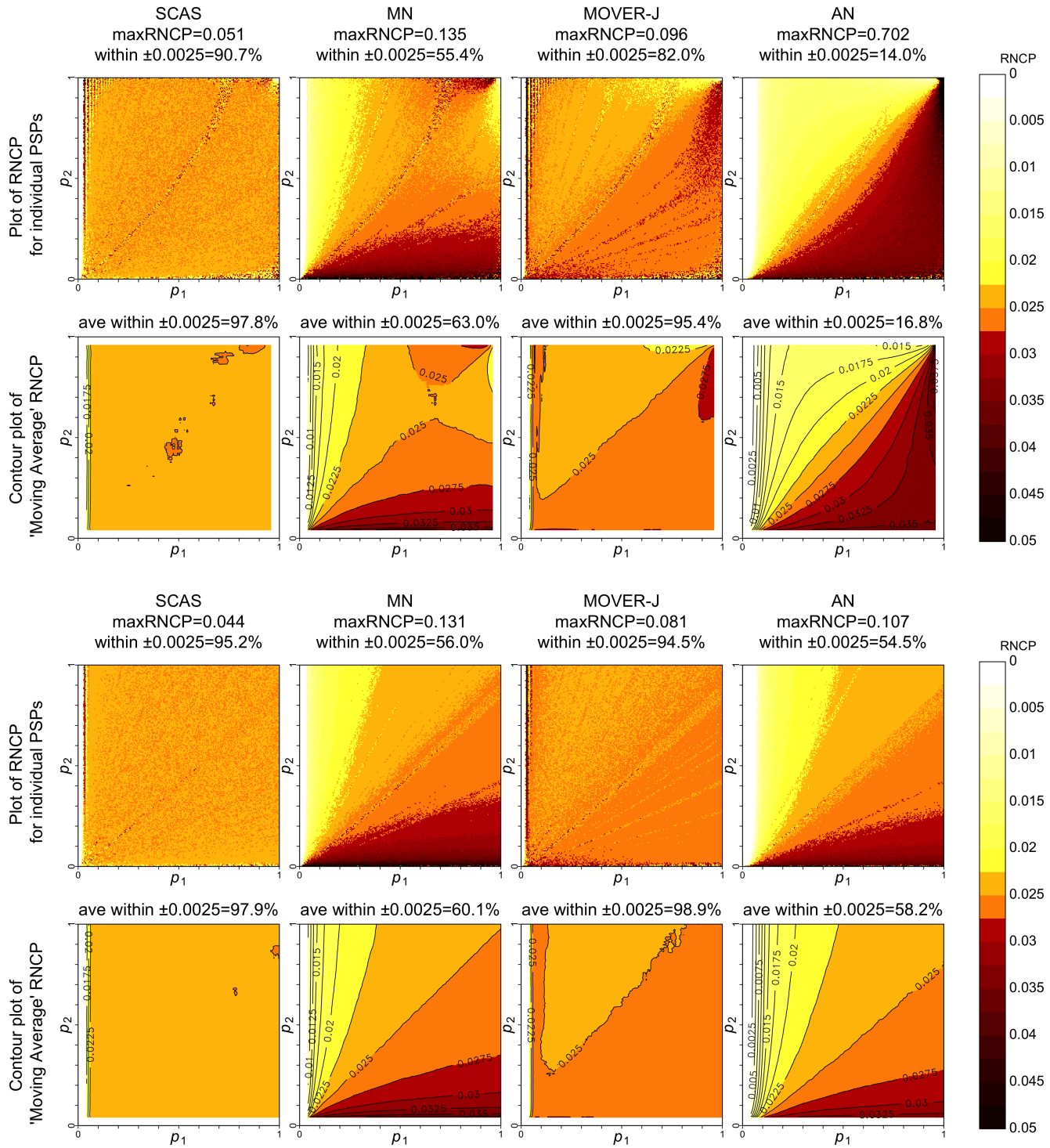

Figure S7. Rate Ratio: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 100$, $n_2 = 100$. Top: Binomial, Bottom: Poisson.

Binomial and Poisson Rate Ratio: $(n_1, n_2) = (50, 150)$

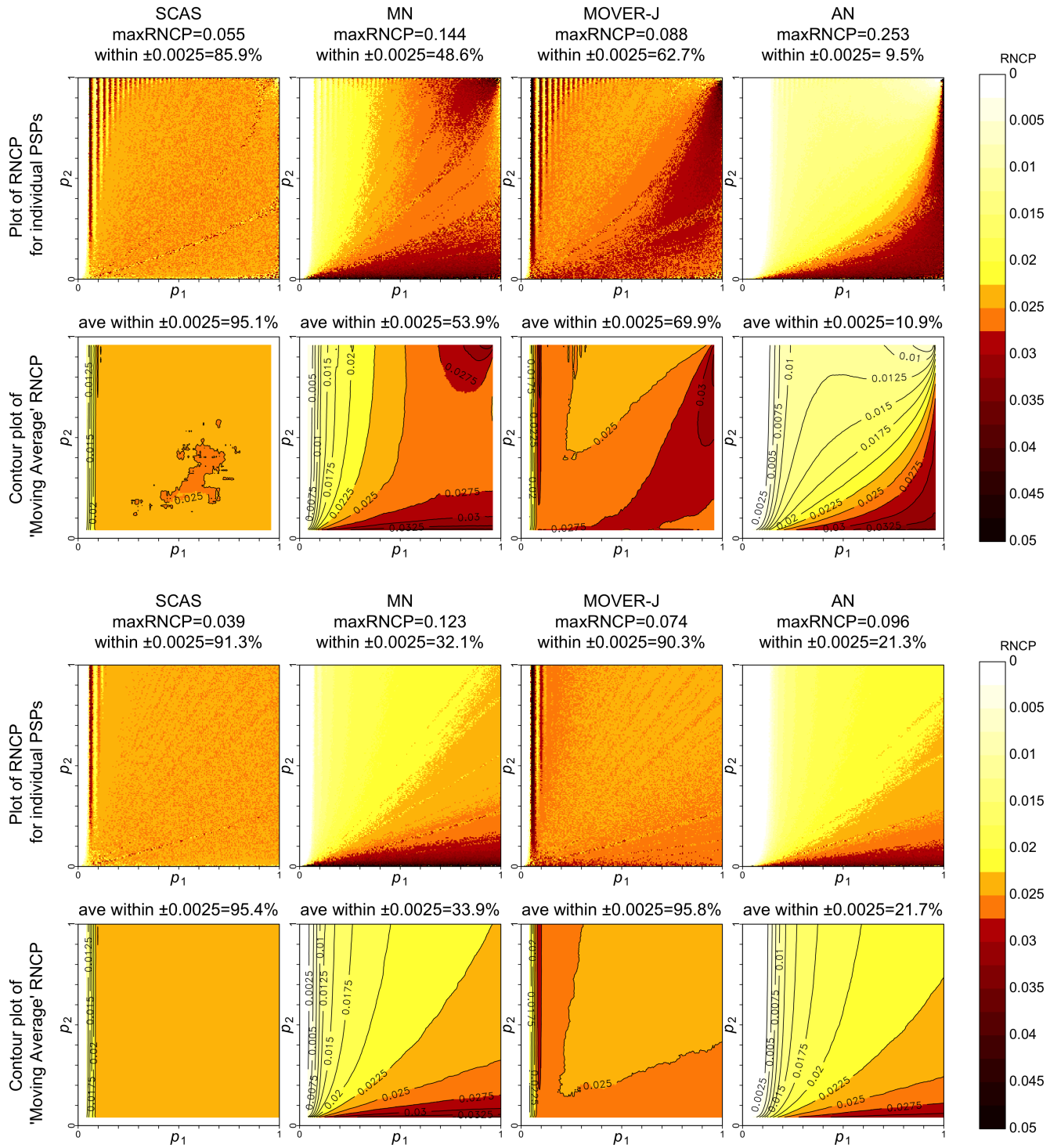

Figure S8. Rate Ratio: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 50$, $n_2 = 150$. Top: Binomial, Bottom: Poisson.

Odds Ratio: $(n_1, n_2) = (30, 30)$ and $(45, 15)$

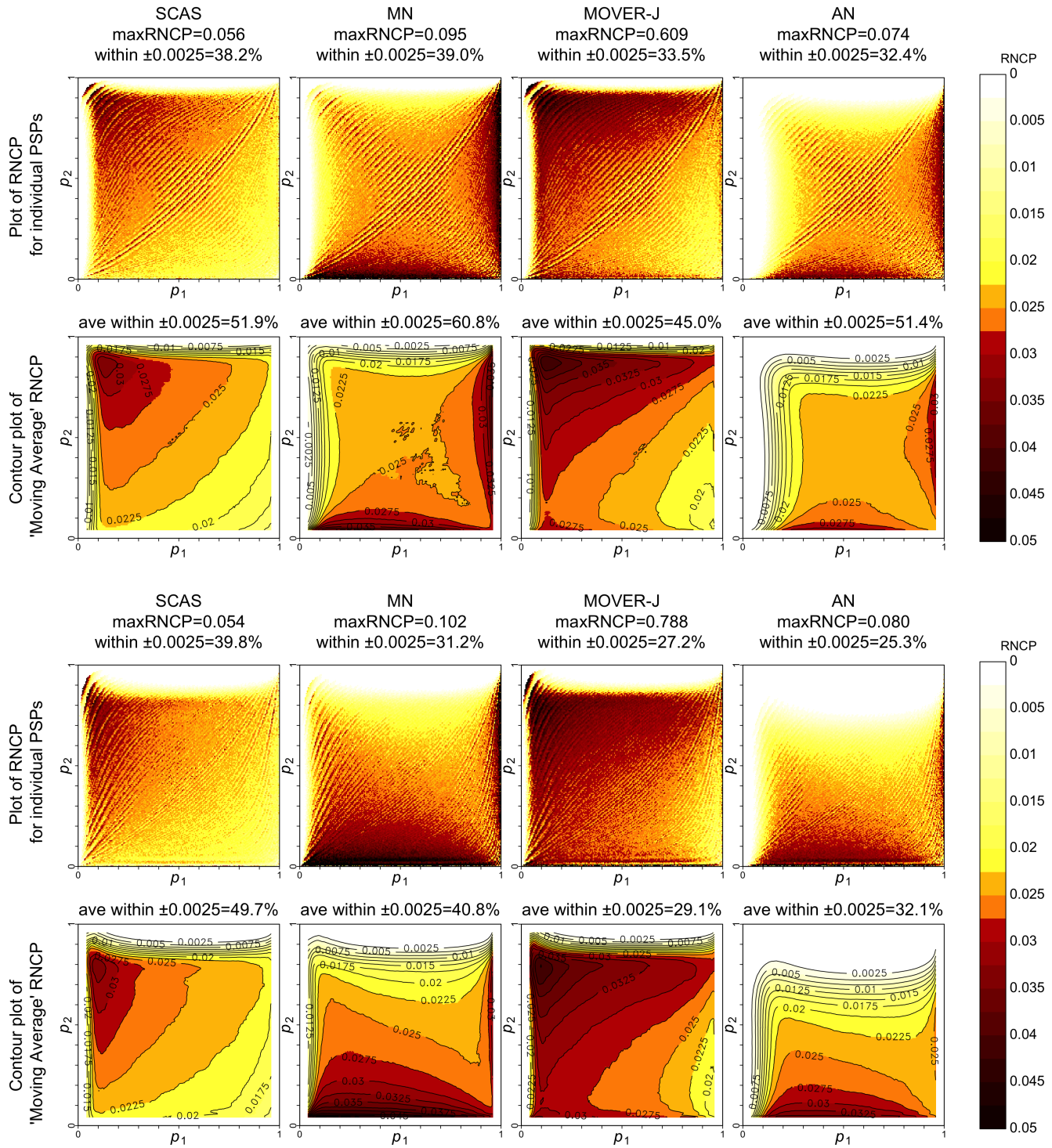

Figure S9. Odds Ratio: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with (top) $n_1 = 30$, $n_2 = 30$, and (bottom) $n_1 = 45$, $n_2 = 15$.

Odds Ratio: $(n_1, n_2) = (100, 100)$ and $(50, 150)$

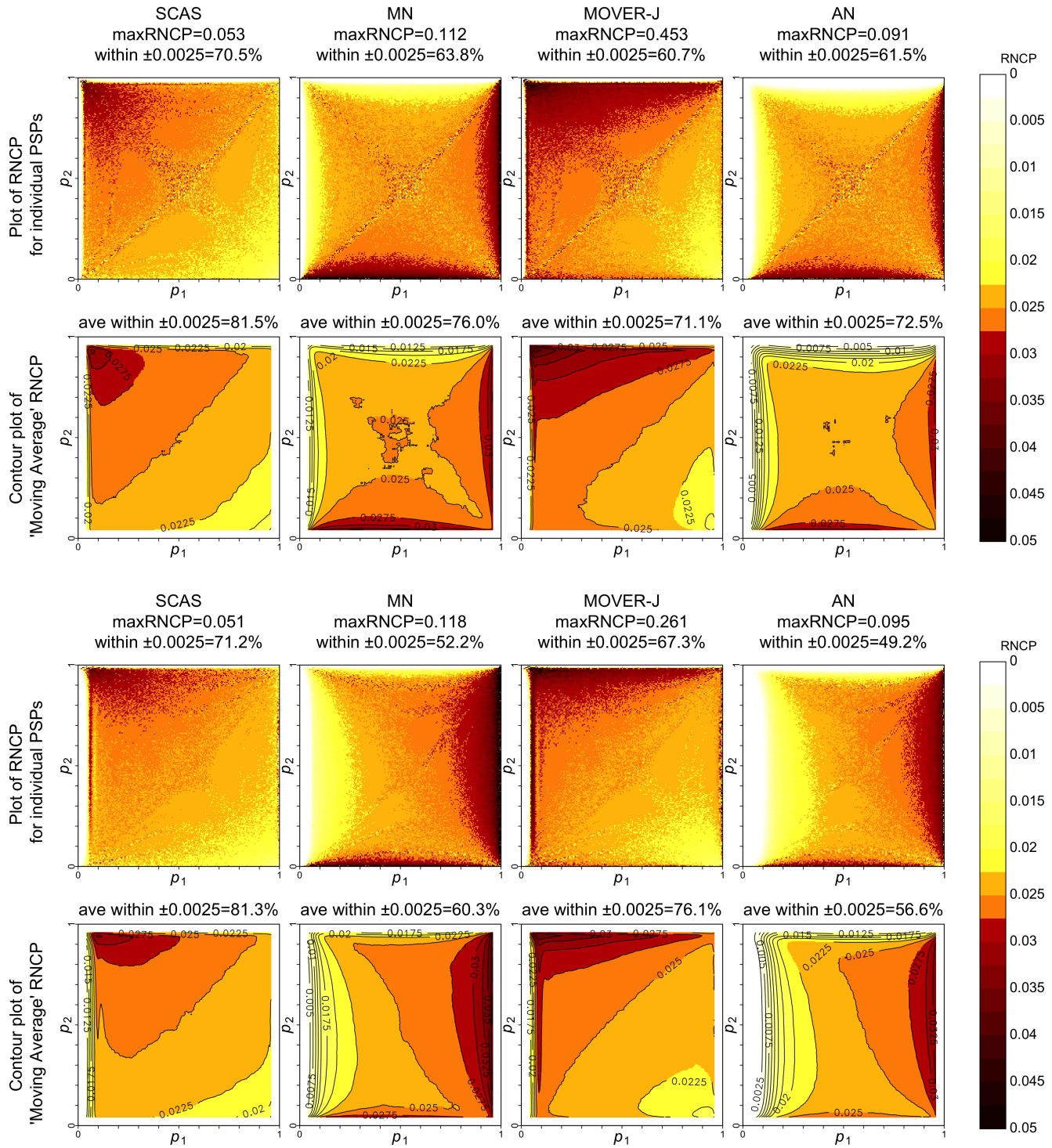

Figure S10. Odds Ratio: Right-sided non-coverage probability (RNCP) and moving average RNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with (top) $n_1 = 100$, $n_2 = 100$, and (bottom) $n_1 = 50$, $n_2 = 150$.

S3.2 Rare events: p_1 and p_2 <0.05

Rate Difference: $(n_1, n_2) = (600, 200)$

Coverage properties for Poisson and binomial RD in this region of the parameter space are very similar, so only the binomial case is shown.

Right- and Left-sided non-coverage:

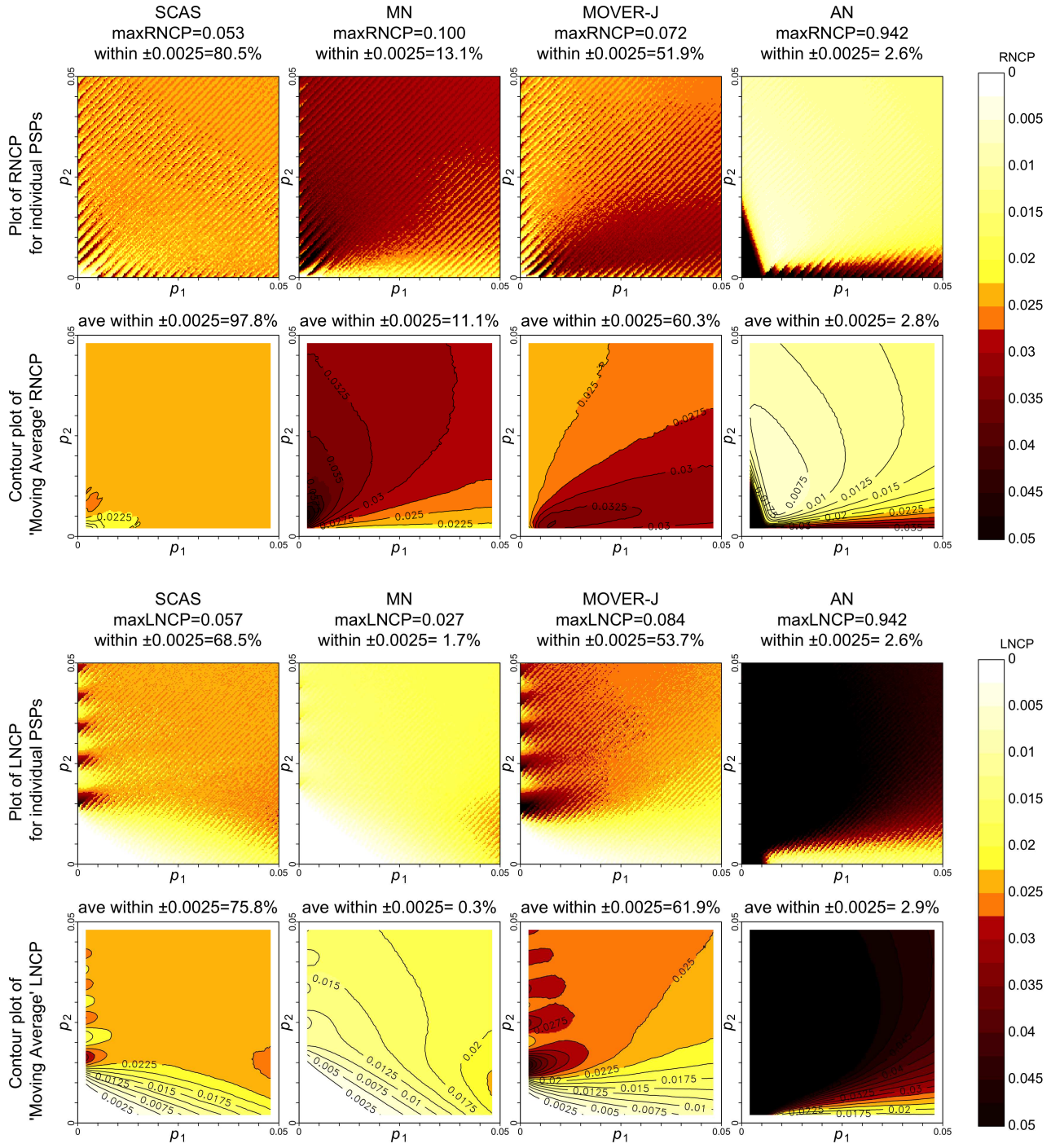

Figure S11. Rate Difference: Right- and Left-sided non-coverage probability (RNCP, LNCP) and moving average RNCP, LNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 600$, $n_2 = 200$ and $p_1 < 0.05$, $p_2 < 0.05$. Top: RNCP, Bottom: LNCP.

Rate Ratio and Odds Ratio: $(n_1, n_2) = (600, 200)$

Coverage properties for Poisson and binomial RR (not shown) in this region of the parameter space are very similar to those shown below for OR.

Right- and Left-sided non-coverage:

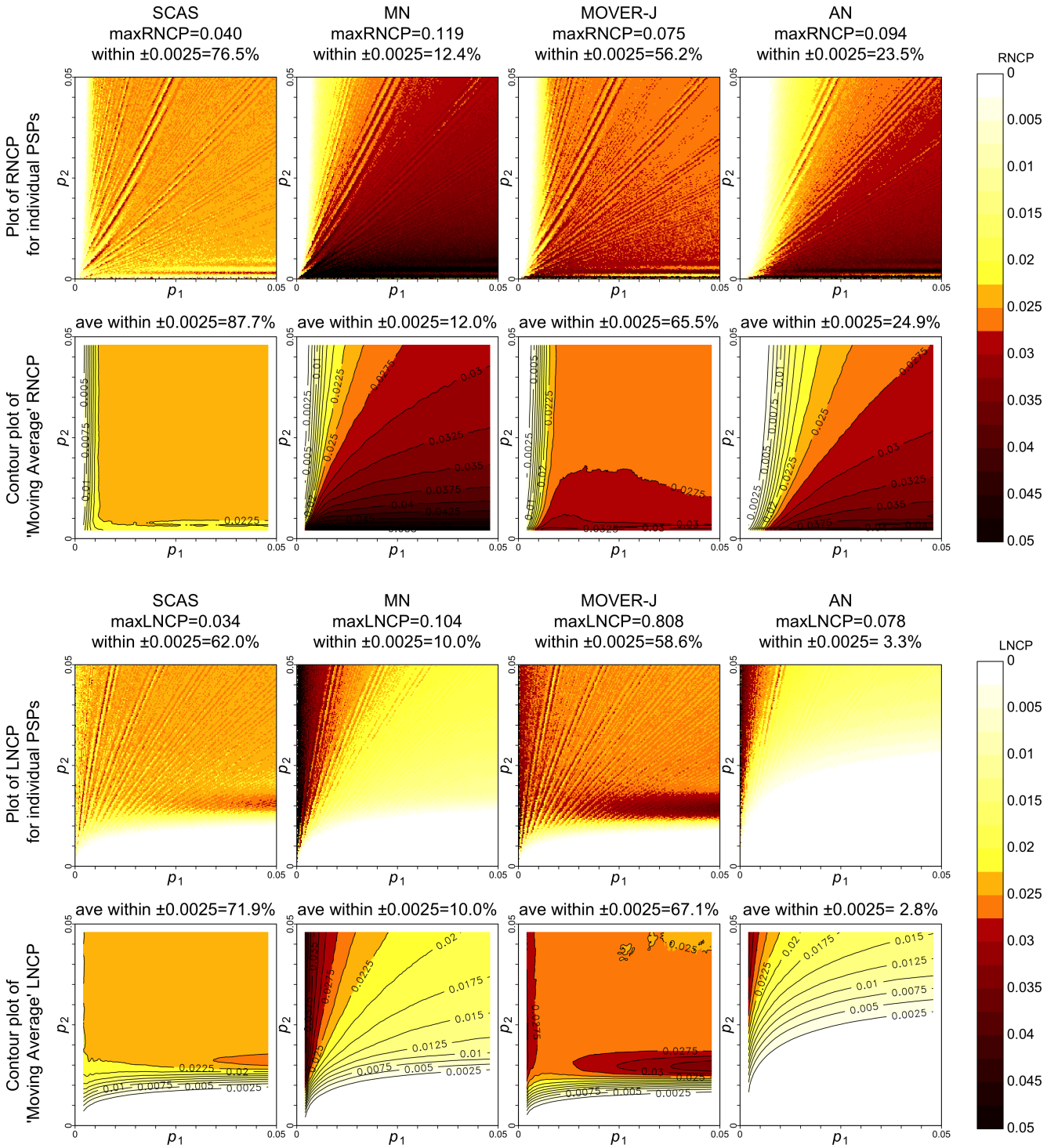

Figure S12. Odds ratio: Right- and Left-sided non-coverage probability (RNCP, LNCP) and moving average RNCP, LNCP, for **SCAS**, **MN**, **MOVER-J** and **AN**, with $n_1 = 600$, $n_2 = 200$ and $p_1 < 0.05$, $p_2 < 0.05$. Top: RNCP, Bottom: LNCP.

S3.3 Bayesian methods

Here the other Bayesian methods for RD are seen to be similar to **MOVER-J** when Jeffreys priors are used, either using the exact Bayesian method (Agresti-Min '**AMB-exact**') or the approximation using simulations ('**AMB-approx**'),[16] and relatively poor with Uniform priors (Nurminen-Mutanen '**NM**').[17] Similar observations are made regarding other contrasts for binomial rates (not shown), and for the Barker-Cadwell methods for Poisson RR ('**BC-J**', '**BC-U**').[18]

Binomial RD and Poisson RR: $(n_1, n_2) = (150, 50)$

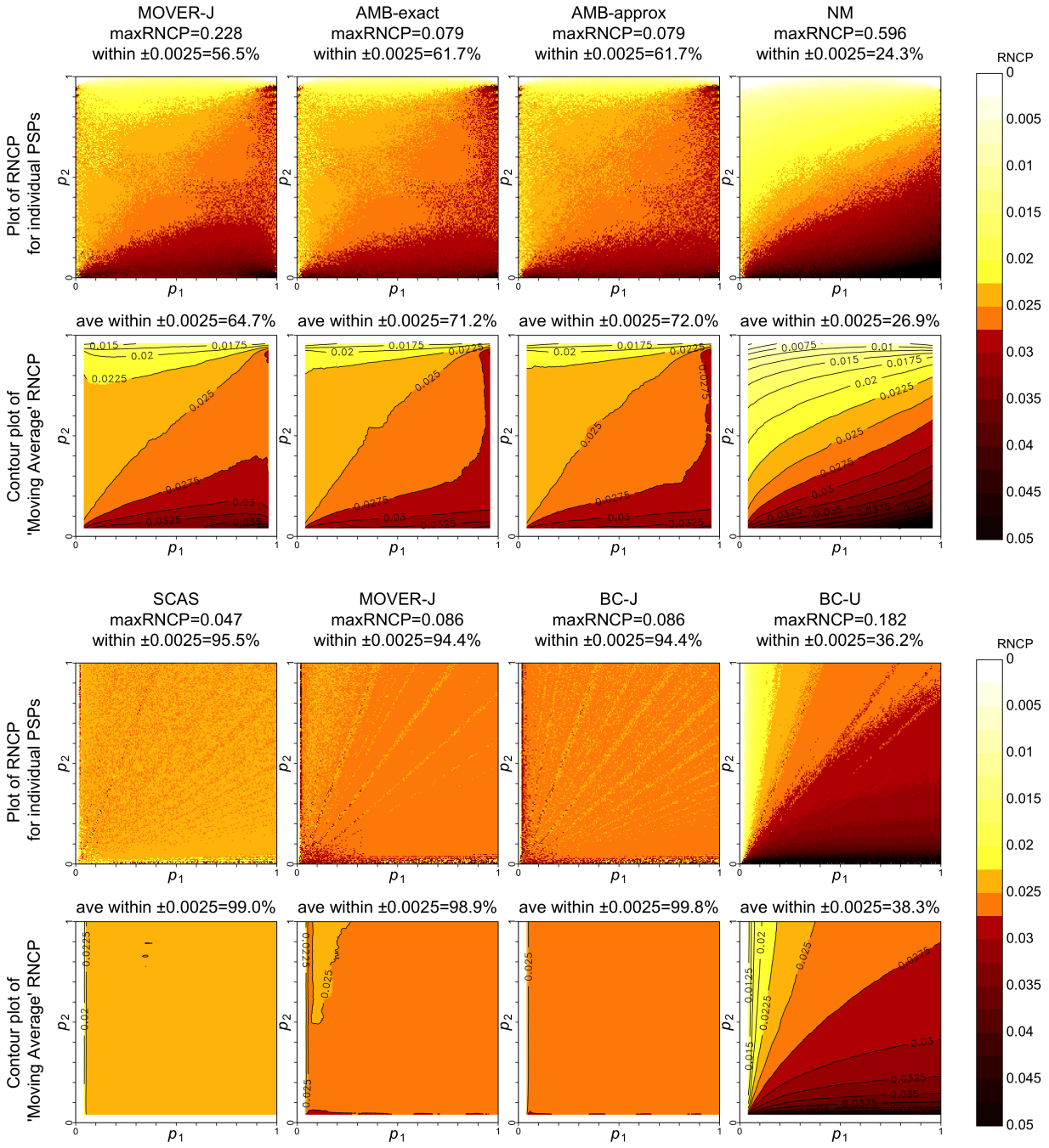

Figure S13. Right-sided non-coverage probability (RNCP) and moving average RNCP, for approximate and 'exact' Bayesian methods, with $n_1 = 150$, $n_2 = 50$. Top: Binomial RD, Bottom: Poisson RR.

S3.4 Continuity corrected methods

Figure S14 illustrates the coverage properties of the continuity corrected '**SCAS-cc**' and '**MOVER-cc**' methods, using the default value of $\gamma = 0.5$, and an experimental 'compromise' value of $\gamma = 0.25$. In the context of aligning minimum coverage with the nominal significance level, a more relevant summary measure (pctCons) is shown, indicating the percentage of the parameter space where RNCP is below $\alpha/2$. Table S1 contains example intervals calculated with these methods.

Binomial RD and Poisson RR: $(n_1, n_2) = (150, 50)$

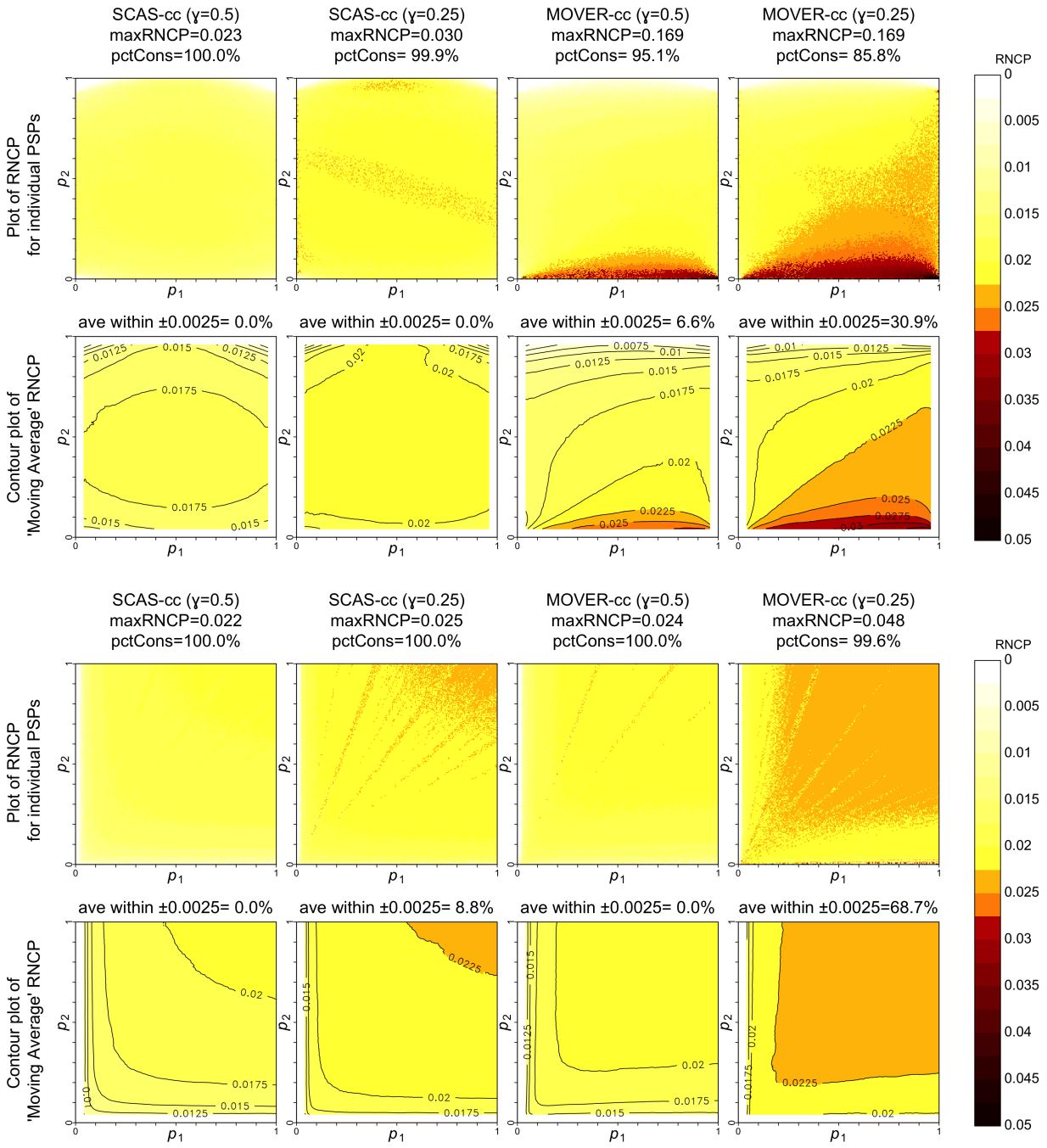

Figure S14. Right-sided non-coverage probability (RNCP) and moving average RNCP, for continuity-corrected methods, with $n_1 = 150$, $n_2 = 50$. Top: Binomial RD, Bottom: Poisson RR.

Table S1. Example continuity-corrected 95% Confidence Intervals

	Success rate: 12/16 (cisapride) vs 1/16 (placebo) (Milo 1984)				
	Binomial RD	Poisson RD	Binomial RR	Poisson RR	Binomial OR
SCAS-cc					
$\gamma = 0.5$	(0.348, 0.897)	(0.252, 1.260)	(2.133, 29123)	(1.718, 72534)	(3.819, 163689)
$\gamma = 0.25$	(0.367, 0.888)	(0.269, 1.240)	(2.366, 647.609)	(1.918, 736.308)	(4.588, 3447.613)
MOVER-cc					
$\gamma = 0.5$	(0.319, 0.868)	(0.224, 1.241)	(2.246, 471.307)	(1.830, 483.351)	(4.497, 2038.097)
$\gamma = 0.25$	(0.346, 0.859)	(0.249, 1.221)	(2.447, 199.968)	(2.003, 207.752)	(5.129, 915.275)
	Adverse event rate: 5/56 vs 0/29 (Goodfield 1992)				
	Binomial RD	Poisson RD	Binomial RR	Poisson RR	Binomial OR
SCAS-cc					
$\gamma = 0.5$	(-0.048, 0.209)	(-0.055, 0.221)	(0.463, ∞)	(0.432, ∞)	(0.435, ∞)
$\gamma = 0.25$	(-0.034, 0.198)	(-0.039, 0.209)	(0.585, ∞)	(0.549, ∞)	(0.561, ∞)
MOVER-cc					
$\gamma = 0.5$	(-0.044, 0.189)	(-0.051, 0.201)	(0.552, ∞)	(0.515, ∞)	(0.521, ∞)
$\gamma = 0.25$	(-0.028, 0.183)	(-0.033, 0.194)	(0.666, 1.02×10^7)	(0.626, 1.02×10^7)	(0.643, 1.13×10^7)

S3.5 Equal-tailed intervals for a single rate

Figure S15 illustrates the one-sided coverage properties of the **SCAS** and Jeffreys methods for a single binomial proportion or Poisson rate, compared with the Wilson score and approximate normal (Wald) methods, with $n = 30$. The dashed line shows RNCP, the solid line is moving average RNCP, and reference lines are drawn at $\alpha/2 \pm 0.1\alpha/2$. The systematic bias in the latter two methods was pointed out by Cai.[19] Table S2 contains example intervals calculated with these methods. RNCP plots for continuity-corrected methods for the single rate are shown for reference in Figure S16.

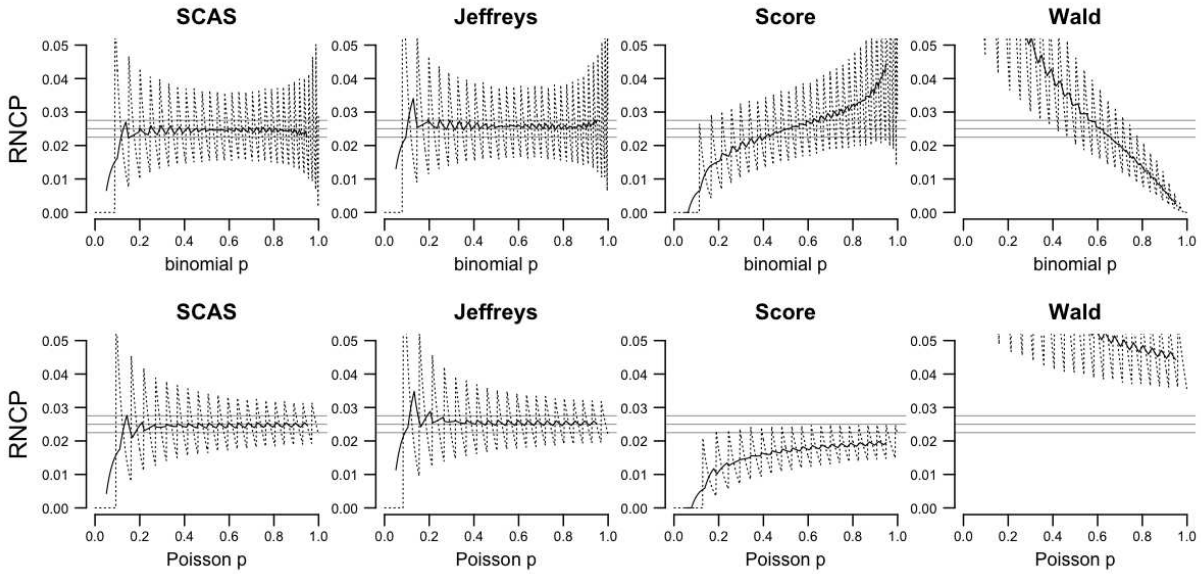

Figure S15. RNCP for single rate methods.

Table S2. Example 95% Confidence Intervals for a single rate

	5/56		0/29	
	Binomial	Poisson	Binomial	Poisson
SCAS	(0.034, 0.186)	(0.033, 0.197)	(0, 0.092)	(0, 0.097)
Jeffreys	(0.035, 0.185)	(0.034, 0.196)	(0, 0.082)	(0, 0.087)
Score	(0.039, 0.193)	(0.038, 0.209)	(0, 0.117)	(0, 0.132)
Wald	(0.015, 0.164)	(0.011, 0.168)	(0, 0)	(0, 0)

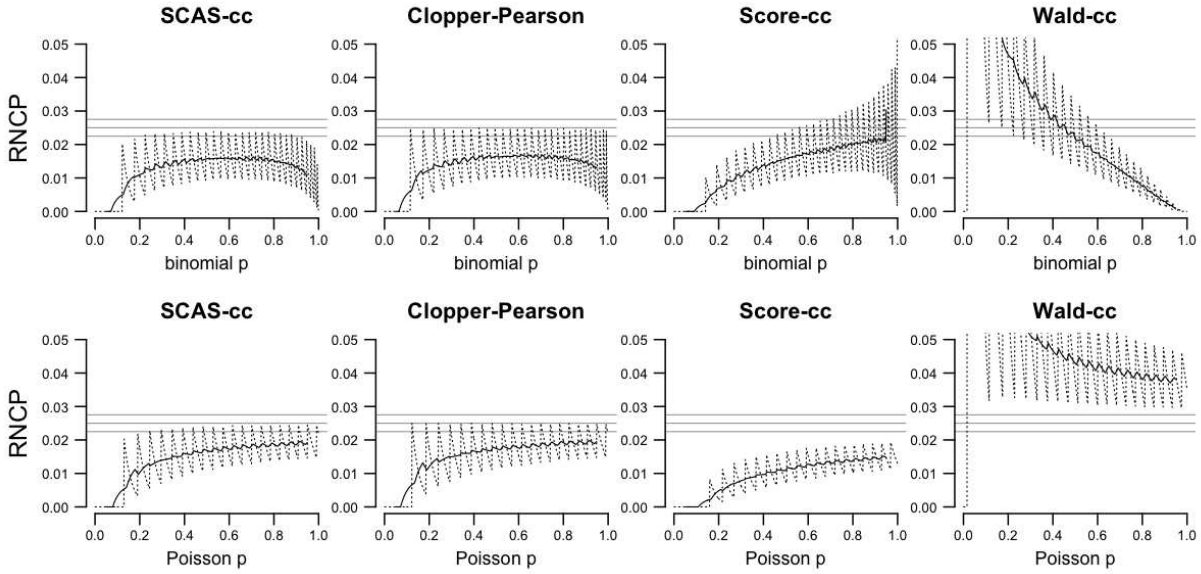

Figure S16. RNCP for continuity-corrected single rate methods.

APPENDIX S4
FURTHER DETAILS FOR STRATIFIED SCORE METHODS

S4.1 Stratum weights

S4.1.1 Inverse variance

Inverse variance (IV) weights based on a crude estimate of stratum variances are required for the random effects **DL** and **HKSJ** methods. Unless stratum sizes are very large, these weights are problematic, because estimating weights using observed variances is biased.[20] The results of the preliminary evaluation of stratified methods confirmed this - in particular, for the **AN** method the performance using IV weights was vastly inferior to 'Mantel-Haenszel' ('MH') weights. (This was observed independently of whether the conditions resulted in strata containing zero cell counts, necessitating the addition of 0.5 to the counts in affected strata, which can worsen coverage properties further.)

S4.1.2 Inverse variance of the score

As noted in Section 4, 'Inverse variance of the score' (IVS) weights may be defined in the context of score methods using the MLE of the stratum-specific variances of $S_j(\theta)$ at each value of θ, i.e. $w_j = \widetilde{V}_j^{-1}$. It is worth noting that for RR and OR, these weights are fundamentally different from those usually employed in a meta-analysis, which are conventionally based on the estimated variance of $ln(\widehat{\theta}_j)$. A full evaluation of the operating characteristics of these weights is beyond the scope of this paper, so it remains to be seen whether the use of weights other than IVS would improve or worsen coverage properties of the **SCAS** interval. Using IVS weights for the asymptotic score method as described here appears to remove the bias inherent in inverse variance weighting - coverage was found to be similar for **SCAS** using MH or IVS weights. This in turn enables the **TDAS** method to have more consistent coverage than **HKSJ** for a random effects analysis.

S4.1.3 Miettinen-Nurminen

Another alternative is to employ the iterative weights proposed by Miettinen and Nurminen ('MNi'). Their 'first approximation' (where $\widetilde{p}_1 = \widetilde{p}_2$) is the same as the MH

weights.[21] The formula defining the MNi weights can also be rearranged to take a similar form as the IVS weights presented here, differing only in the use of the weighted averages for $\widetilde{p}_1$ and $\widetilde{p}_2$ instead of the stratum-specific estimates. The conditions under which the MNi and IVS weights are similar may merit further investigation.

The non-iterative weights defined for OR by Miettinen and Nurminen appear to be effectively the same as $w_j = \widetilde{V}_j^{-1}$ in practice.

S4.1.4 Mantel-Haenszel

In the usual meta-analysis for RD,[1] the overall estimate is constructed as a weighted sum of the stratum-specific estimates, $\widehat{\theta} = \sum_j w_j \widehat{\theta}_j / W$, and the MH weights are $w_j = n_{1j} n_{2j} / N_j$. In the context of the score method for RD, this leads directly to the analogous weighted score function $S(\theta) = \sum_j w_j S_j(\theta)/W$.

For RR, the appropriate MH weights for the score method are identical to those for RD, unlike the usual MH weights. Usually, the MH estimator $\widehat{\theta}_{MH} = \left(\sum_j n_{2j} X_{1j}/N_j\right) / \left(\sum_j n_{1j} X_{2j}/N_j\right)$ is rearranged to give $\widehat{\theta} = \sum_j w'_j \widehat{\theta}_j / \sum_j w'_j$ with $w'_j = n_{1j} X_{2j}/N_j$ [22] (although this appears to require the assumption that the ratio n_{1j}/n_{2j} is constant across strata). For the score-based method, the MH estimator may instead be rearranged (with division by W in the numerator and denominator) as the weighted score function:

$$S(\theta) = \sum_j \frac{1}{W}\left[\frac{n_{1j} n_{2j}}{N_j}\frac{X_{1j}}{n_{1j}}\right] - \theta \sum_j \frac{1}{W}\left[\frac{n_{1j} n_{2j}}{N_j}\frac{X_{2j}}{n_{2j}}\right]$$
$$= \sum_j \frac{w_j \widehat{p}_{1j}}{W} - \theta \sum_j \frac{w_j \widehat{p}_{2j}}{W} = \sum_j w_j S_j(\theta)/W \quad \text{(S1)}$$

(Note that in practice, division by the constant $W = \sum_j w_j$ does not affect the calculations in Equation (4), because terms of W in the numerator and denominator cancel each other out.)

For OR, this conversion of the Mantel Haenszel estimator is not so simple, so there is no similar theoretical justification for the use of $w_j = n_{1j} n_{2j}/N_j$. The empirical performance of the **SCAS** method with these weights appears to be acceptable, although this is yet to be confirmed under a wider variety of stratum size conditions.

S4.2 Heterogeneity and random effects

S4.2.1 Test for 'quantitative' interaction

Heterogeneity of treatment effects (i.e. treatment-by-stratum interaction) can be described within the score test framework using the following chi-squared test statistic,

$$Q = \sum_j \frac{[S_j(\theta) - \sum_j (w_j/W) S_j(\theta)]^2}{\widetilde{V}_j} \sim \chi^2_{k-1} \quad \text{(S2)}$$

where k is the number of strata.

This statistic is defined in its full form (based on the usual form of the Cochran Q statistic[23]) for use in later sections. For a heterogeneity test (as suggested by Gart and Nam[24] from Tarone[25,26]) it is simplified by being evaluated (under the null hypothesis assumption of constant θ across strata) at the maximum likelihood estimate $\widehat{\theta}_{ML}$ for the common treatment effect, which is obtained by solving $\sum_j w_j S_j(\widehat{\theta}_{ML}) = 0$. Therefore the test statistic simplifies to:

$$\widehat{Q} = \sum_j \frac{[S_j(\theta)]^2}{\widetilde{V}_j}\bigg|_{\theta=\widehat{\theta}_{ML}} \quad \text{(S3)}$$

The proportion of variability due to heterogeneity may be quantified by using $\widehat{Q}$ in the usual formula[27] for I^2, that is $I^2 = max(0, (\widehat{Q} - (k-1))/\widehat{Q})$. The components of $\widehat{Q}$ may also be used to produce a Q-Q plot or a Galbraith plot (i.e. a plot of $S_j(\theta)/\widetilde{V}_j^{1/2}$ against $1/\widetilde{V}_j^{1/2}$), to identify outliers and explore the nature of any heterogeneity, or indeed to visually confirm the consistency of treatment effects (i.e. homogeneity). Note that this plot would not have the usual feature of Galbraith plots such that the slope describes the point estimate for θ, because it is based not on θ but $S(\theta)$, for which the expected value is always zero.

Note that $\widehat{Q}$ may be indeterminate for RR or OR if $\sum_j X_{1j} = 0$ or $\sum_j X_{2j} = 0$, because $\widehat{\theta}_{ML}$ cannot be determined. In such cases, a default value of 0 for $\widehat{Q}$ would seem sensible.

S4.2.2 Test for 'qualitative' interaction

A test for 'qualitative' interaction is also possible, for identifying crossover of treatment effects relative to the non-zero null hypothesis value for θ.

Usually, in the context of superiority testing, a 'qualitative' (or 'crossover') interaction indicates that the direction of treatment effects varies across strata. This concept may be adapted to describe the nature of heterogeneity in a stratified non-inferiority analysis, by considering the direction of the difference between the stratum treatment effects relative to the non-inferiority margin. That is, a qualitative interaction indicates that the comparative parameter falls below θ_0 for some strata and above θ_0 for others.

Following a significant result from the interaction test in Section S4.2.1, the nature of the type of interaction can be explored using a generalisation of the Gail and Simon approach[28] as follows, incorporating indicator functions $I()$

for the direction of each stratum's deviation from θ_0.

$$\widehat{Q}_c = min\Bigg\{ \sum_j \frac{S_j(\theta_0)^2 I(S_j(\theta_0) > 0)}{\widetilde{V}_j}, \sum_j \frac{S_j(\theta_0)^2 I(S_j(\theta_0) < 0)}{\widetilde{V}_j} \Bigg\}\Bigg|_{\theta=\widehat{\theta}_{ML}} \quad \text{(S4)}$$

With $\theta_0 = 0$, for RD it can be seen that this is essentially the same as the Gail and Simon test, except for the use of the maximum likelihood variance estimator in the denominators. Under the assumption that the $S_j(\theta)$ are normally distributed, the p-value for this test is calulated in the same way as for the Gail and Simon test, as:

$$p(Q_c) = \sum_{h=1}^{k-1} (1 - F_h(\widehat{Q}_c)) B(h; n = (k-1), p = 0.5)$$

where $F_h()$ is the cumulative chi-square distribution function with h degrees of freedom and $B()$ is the binomial probability mass function with parameters n and p.

S4.2.3 Incorporating stratum variability ('Random effects')

The 'Hartung-Knapp-Sidik-Jonkman' ('**HKSJ**') random effects confidence interval, with a derivation based on the t-distribution, has been found to be superior to the DerSimonian and Laird method.[29] By adapting the formulae defining the **HKSJ** method,[30,31,32] a new t-distribution asymptotic score method ('**TDAS**') may be devised as follows:

The variance of the stratum scores $S_j(\theta)$ can be estimated using the method of moments estimator, which, if inverse variance weights are used, is:

$$\widetilde{\tau}^2 = max\left\{ 0, \frac{Q - (k-1)}{W - (\sum_j w_j^2 / W)} \right\} \quad \text{(S5)}$$

where Q is defined as in Equation (S2). Then the IV weights are updated as $w_j^* = (\widetilde{V}_j + \widetilde{\tau}^2)^{-1}$, and $W^* = \sum_j w_j^*$.

Then, by analogy to Equation (3) of Sidik and Jonkman,[31] assuming that asymptotically:

$$Z(\theta) = \frac{\sum_j (w_j^*/W^*) S_j(\theta)}{1/(W^*)^{1/2}} \sim N(0,1),$$

and

$$Q(\theta) = \sum_j w_j^* \Big[S_j(\theta) - \sum_j (w_j^*/W^*) S_j(\theta) \Big]^2 \sim \chi^2_{k-1},$$

it follows that $t(\theta) = \dfrac{Z(\theta)}{[Q(\theta)/(k-1)]^{1/2}} \sim t_{(k-1)}$, and then the **TDAS** confidence interval is obtained by solving:

$$t(\theta) = \frac{\sum_j (w_j^*/W^*) S_j(\theta)}{\widetilde{V}_S^{1/2}} = \pm t_{k-1,1-\alpha/2} \quad \text{(S6)}$$

where

$$\widetilde{V}_S = \frac{\sum_j (w_j^*/W^*) \Big[S_j(\theta) - \sum_j (w_j^*/W^*) S_j(\theta) \Big]^2}{(k-1)} \quad \text{(S7)}$$

As before, throughout the iterative solution of the score function in Equation (S6), the variance function and $\widetilde{\tau}^2$ need to be recalculated for each given value of θ. Here a p-value for a null hypothesis value θ_0 for θ is obtained directly from $t(\theta_0)$. Interestingly, skewness correction does not appear to be necessary here.

If MH or other fixed stratum weights are chosen, then an approximate score function may be obtained by omitting the calculation of $\widetilde{\tau}^2$, and simply using the selected weights in Equations (S6) and (S7) instead of w_j^*. The effect of different weighting schemes may require further evaluation with a focus on more extreme conditions, including small event rates leading to many 'double-zero' strata.

Furthermore, it is possible that the t-distribution method may be improved further, by using robust estimators in Equation (S6), such as a sandwich variance estimator,[32] or trimmed mean with Winsorized variance estimator,[33] to reduce the impact of a small number of anomalous studies. This may be useful when calculating intervals for RR, where a stratum with very low event rates may be given excessive weight due to the resulting small estimated variance of $S(\theta)$. Naturally, a trimmed method would only be feasible with a reasonably large number of strata.

APPENDIX S5
EXTENDED GRAPHICAL EVALUATION FOR THE STRATIFIED METHODS

S5.1 Simulation study

Figures S18 to S21 show the simulated type I error rates for stratified confidence intervals for the parameters not shown in the main paper.

Using a modification of the model used by IntHout et al,[29] the simulations proceeded as follows, for each selected value of k, I^2 and allocation ratio n_{1j}/n_{2j}:

For RD, given the selected overall 'true' population values of p_1 and p_2, the true overall θ is $p_1 - p_2$. Stratum-specific θ_i were generated randomly from $N(\theta, \tau^2)$, where $\tau^2 = \epsilon^2 I^2/(1 - I^2)$, and $\epsilon^2 = k^{-1} \sum_j [p_1(1 - p_1)/n_{1j} + p_2(1 - p_2)/n_{2j}]$.

The nuisance parameter $\bar{p}_j$ was also allowed to vary between strata, by drawing randomly from a Beta$(100\bar{p}, 100(1 - \bar{p}))$ distribution, with $\bar{p} = (p_1 + p_2)/2$.

Stratum-specific true values of p_{1j} and p_{2j} were derived as $\bar{p}_j \pm \theta_j/2$. Event rates below 0.001 or above 0.999 were replaced with 0.001 and 0.999 respectively. Examples of simulated values of p_{1j} and p_{2j} for $k = 20$ and constant n_{ij} are shown in Figure S17 for the homogeneous and heterogeneous models ($I^2 = 0$ and $I^2 = 0.25$ respectively).

Event counts were then randomly generated for 10,000 simulations from binomial(n_{ij}, p_{ij}) distributions. Each simulated dataset was analysed using each of the selected methods using the metabin, metainc and scoreci functions in R. The proportion of replications with an upper confidence limit below the true θ was recorded as the estimated type I error rate.

For RR, a similar process was followed, but using a lognormal distribution for θ, i.e. $\lambda_j = ln(\theta_j)$ was drawn from a normal distribution with mean $ln(\theta)$ and variance $\tau^2 = \epsilon^2 I^2/(1 - I^2)$, where
$\epsilon^2 = k^{-1} \sum_j [(n_{1j}p_1)^{-1} - (n_{1j})^{-1} + (n_{2j}p_2)^{-1} - (n_{2j})^{-1}]$.
Then the stratum-specific p_{1j} and p_{2j} were generated as $exp[ln(\bar{p}_j) \pm \lambda_j/2]$.

For OR, a lognormal distribution was used with $\epsilon^2 = k^{-1} \sum_j [(n_{1j}p_1)^{-1} + (n_{1j}(1-p_1))^{-1} + (n_{2j}p_2)^{-1} + (n_{2j}(1-p_2))^{-1}]$, and the stratum-specific $p_{ij} = o_{ij}/(1 + o_{ij})$, where the group odds $o_{ij} = p_{ij}/(1 - p_{ij})$ were generated as $exp[ln(\bar{p}_j/(1 - \bar{p}_j)) \pm \lambda_j/2]$.

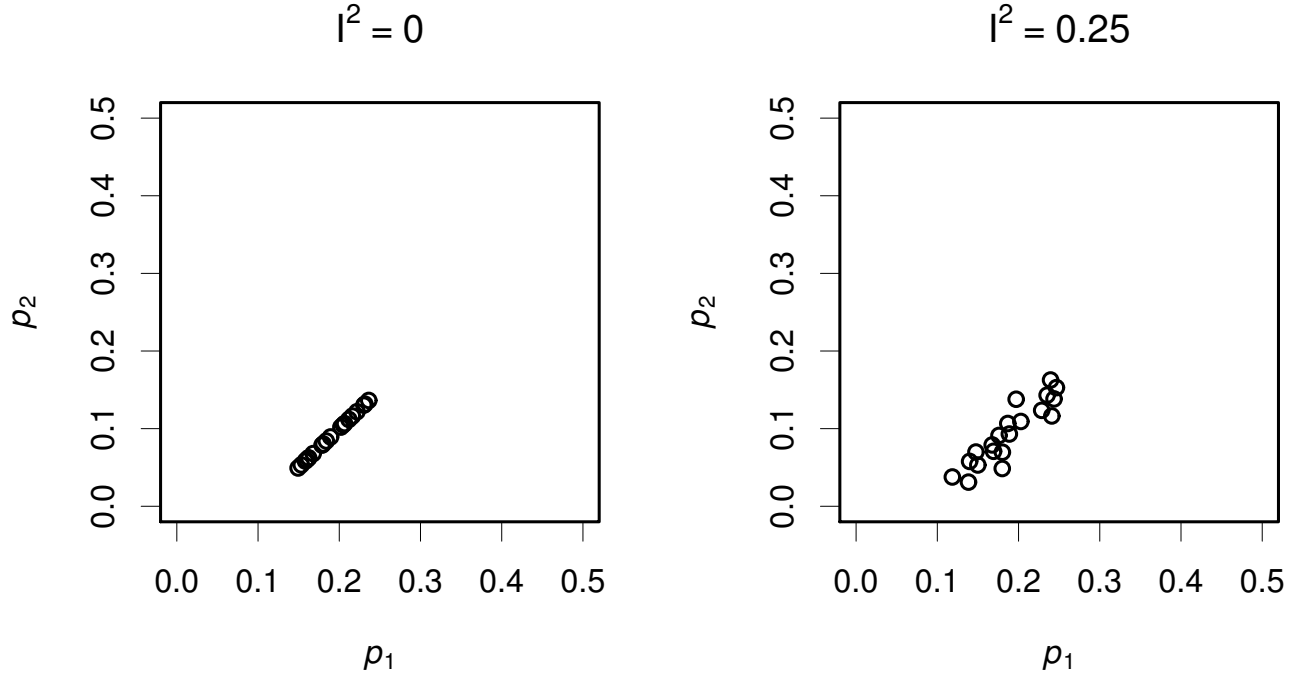

Figure S17. Simulated distributions of stratum rates. Left: homogeneous θ, Right: heterogeneous θ.

Poisson Rate Difference

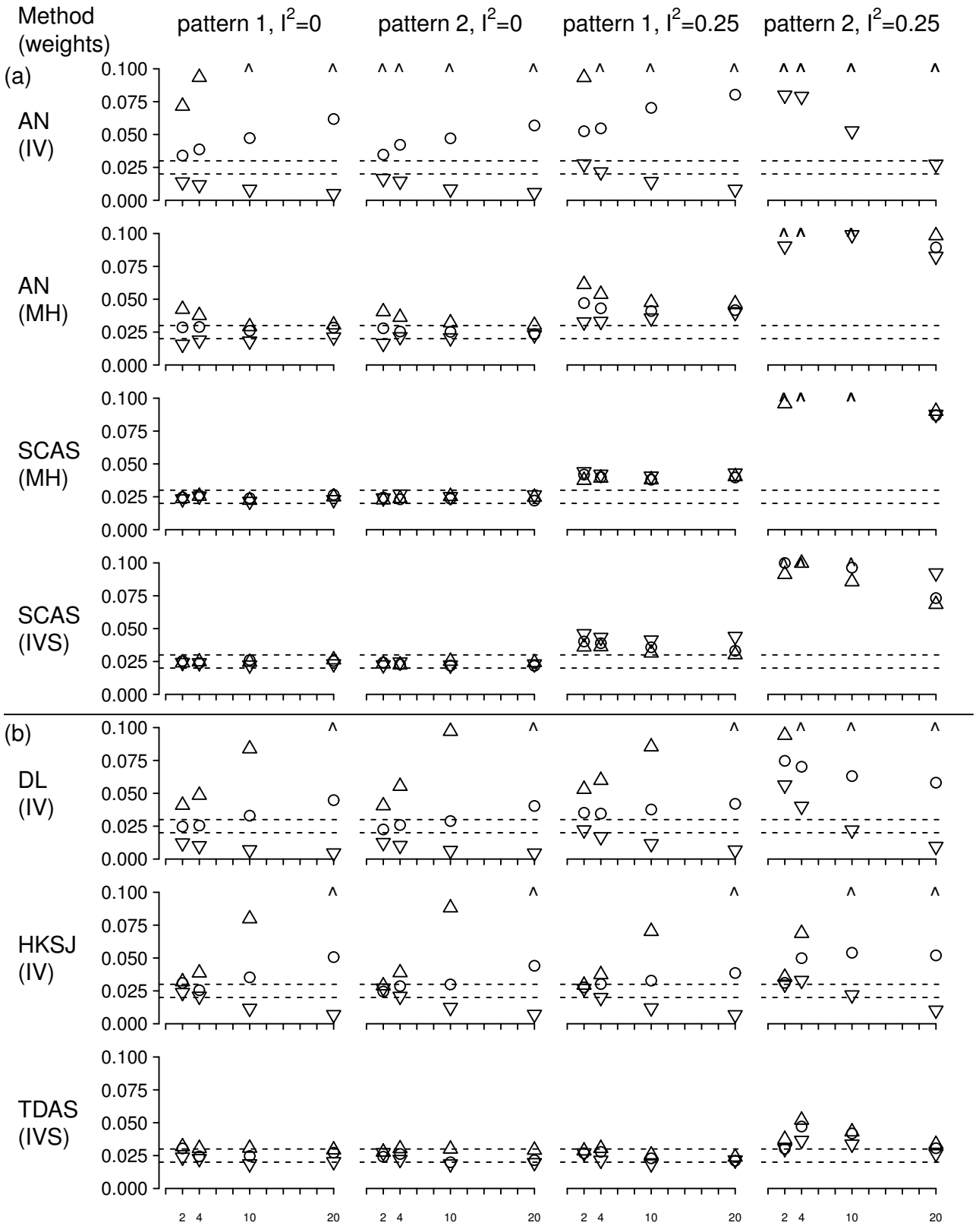

Figure S18. Simulated right-sided type I error rate for meta-analysis of Poisson RD with different number of strata and sample size allocation (△ 3:1, ○ 1:1, ▽ 1:3), under four different sets of conditions. Pattern 1: equal-sized strata; Pattern 2: one stratum 10× larger than other strata; I^2=0: homogeneous treatment effects across strata; I^2=0.25: modest heterogeneity. True overall event rates $p_1 = 0.2$, $p_2 = 0.1$. Reference lines: $\alpha/2 \pm 0.2\alpha/2$. ^ indicates RNCP >0.1. (a): Fixed effects methods, (b): Random effects methods

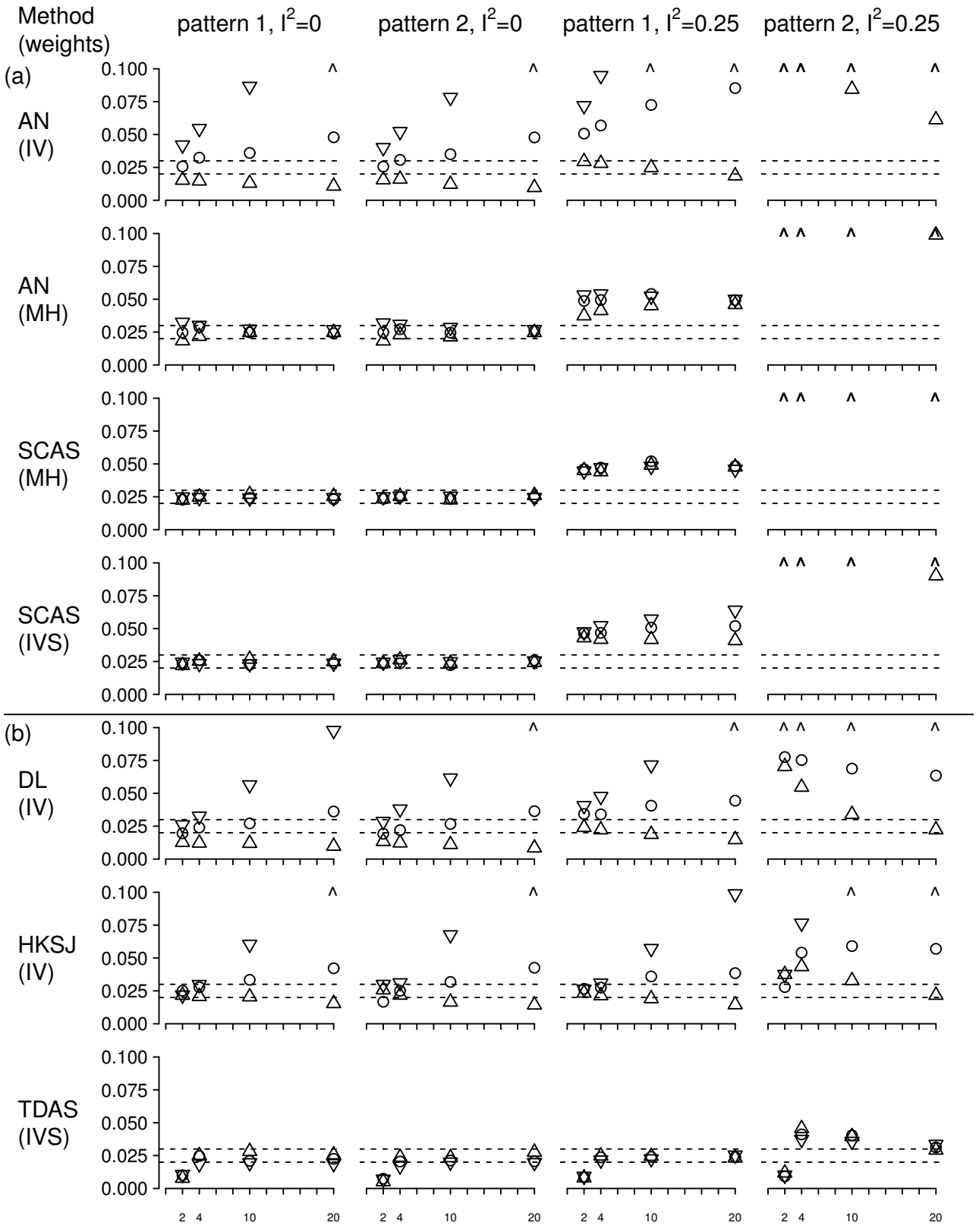

Figure S19. Simulated right-sided type I error rate for meta-analysis of binomial RR with different number of strata and sample size allocation (△ 3:1, ○ 1:1, ▽ 1:3), under four different sets of conditions. Pattern 1: equal-sized strata; Pattern 2: one stratum 10× larger than other strata; I^2=0: homogeneous treatment effects across strata; I^2=0.25: modest heterogeneity. True overall event rates $p_1 = 0.2$, $p_2 = 0.1$. Reference lines: $\alpha/2 \pm 0.2\alpha/2$. ^ indicates RNCP >0.1. (a): Fixed effects methods, (b): Random effects methods

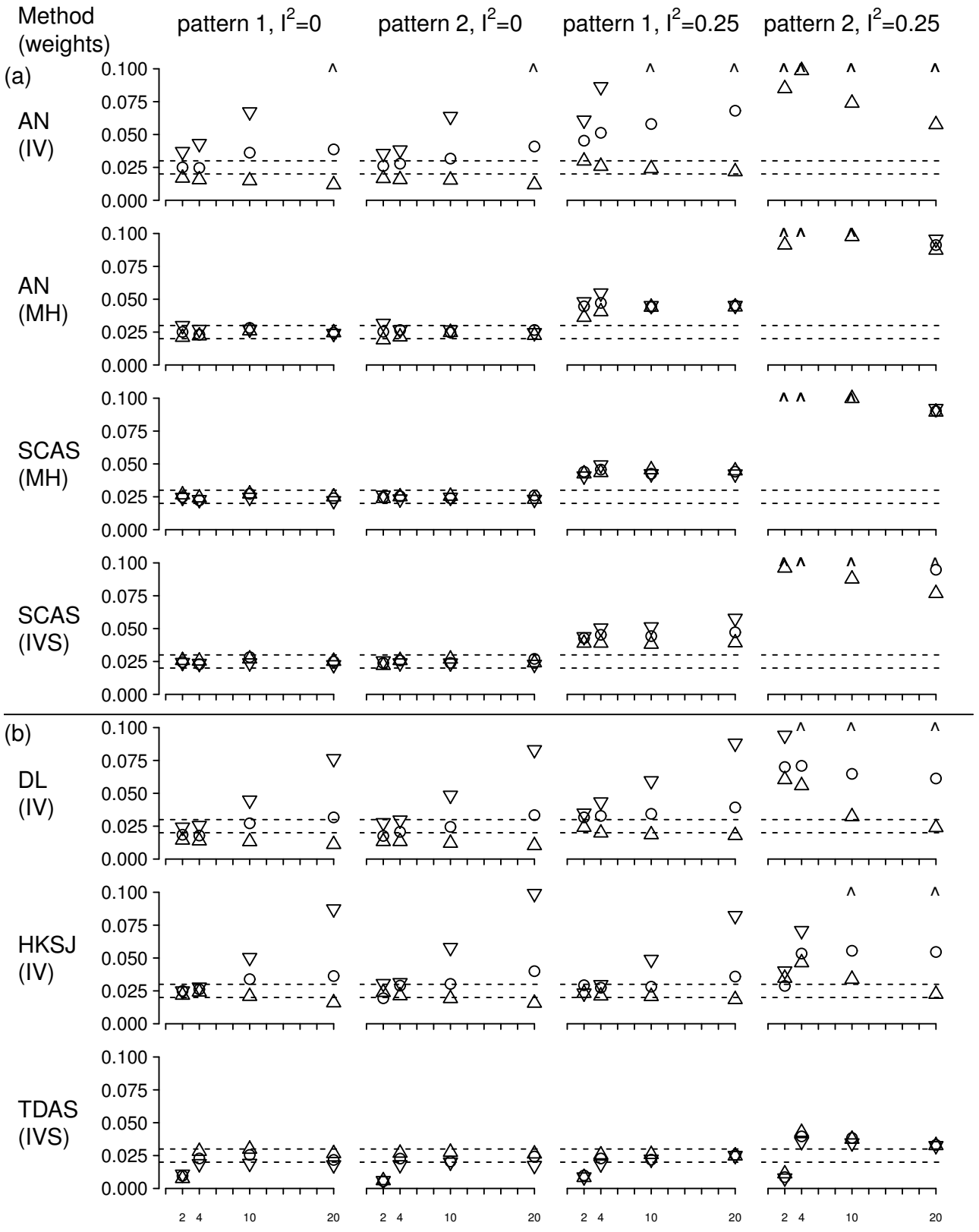

Figure S20. Simulated right-sided type I error rate for meta-analysis of Poisson RR with different number of strata and sample size allocation (△ 3:1, ○ 1:1, ▽ 1:3), under four different sets of conditions. Pattern 1: equal-sized strata; Pattern 2: one stratum 10× larger than other strata; I^2=0: homogeneous treatment effects across strata; I^2=0.25: modest heterogeneity. True overall event rates $p_1 = 0.2$, $p_2 = 0.1$. Reference lines: $\alpha/2 \pm 0.2\alpha/2$. ^ indicates RNCP >0.1. (a): Fixed effects methods, (b): Random effects methods

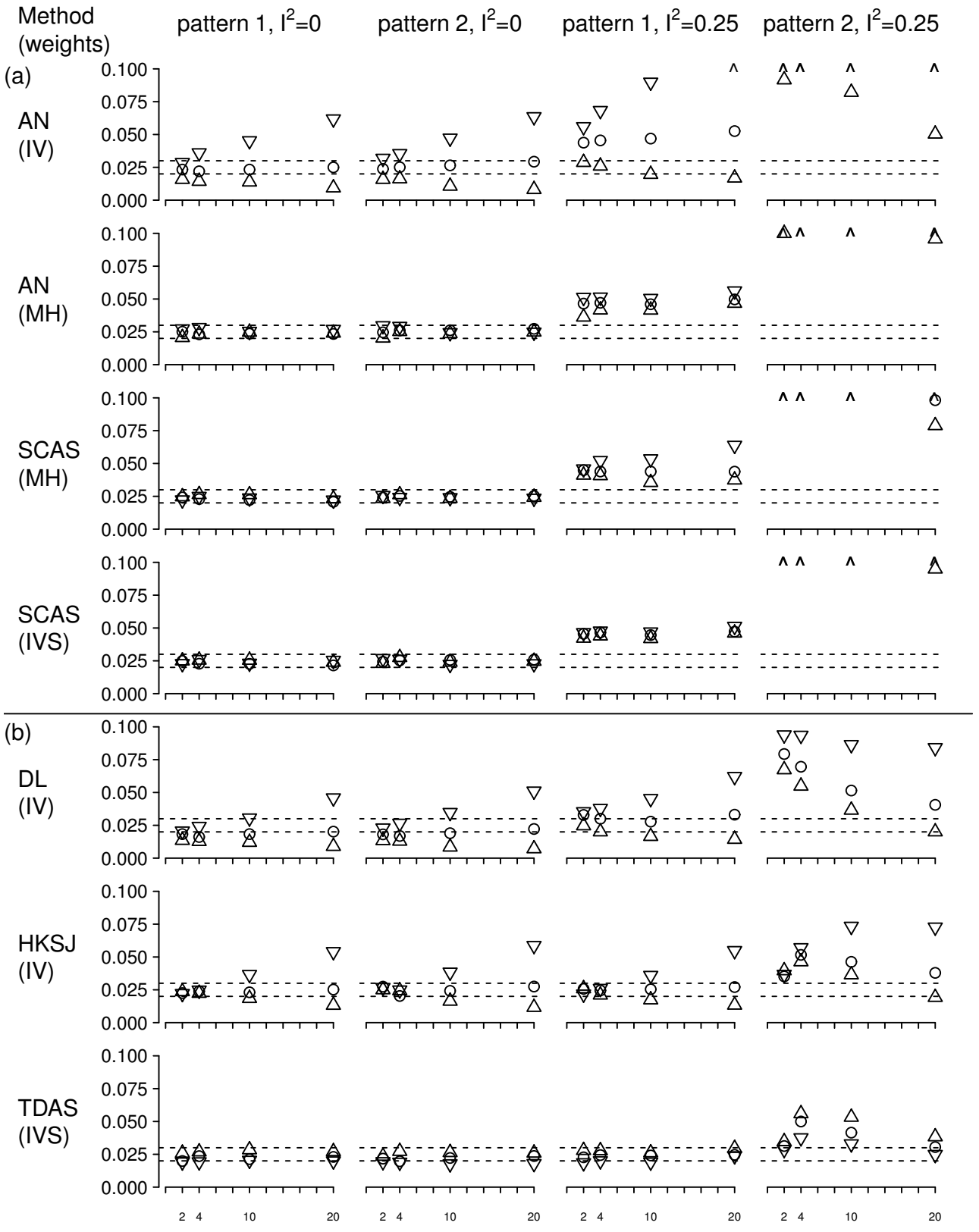

Figure S21. Simulated right-sided type I error rate for meta-analysis of OR with different number of strata and sample size allocation (△ 3:1, ○ 1:1, ▽ 1:3), under four different sets of conditions. Pattern 1: equal-sized strata; Pattern 2: one stratum 10× larger than other strata; I^2=0: homogeneous treatment effects across strata; I^2=0.25: modest heterogeneity. True overall event rates $p_1 = 0.2$, $p_2 = 0.1$. Reference lines: $\alpha/2 \pm 0.2\alpha/2$. ^ indicates RNCP >0.1. (a): Fixed effects methods, (b): Random effects methods

S5.2 Further discussion of random effects methods

It has been noted[34] that a severe reduction in power is observed for the **HKSJ** method with a small number of strata. This occurs because, for example, the confidence interval for the combination of two trials can be much wider than the intervals for each of the trials alone, due to there being very little information on the between-stratum variability. The same is true for the **TDAS** method, which can exhibit a more extreme form of this behaviour, often producing infinite width intervals, particularly for RR. This is somewhat disconcerting, but perhaps simply serves as a warning that there are insufficient strata to obtain a reliable estimate of the between-stratum variability. Clearly for the purposes of most meta-analyses, which often do contain very few studies, an infinite width confidence interval would be unsatisfactory. In such situations where the confidence interval is not being used for a formal hypothesis test, the **HKSJ** method can be recommended, provided treatment allocation is reasonably balanced.

For the assessment of power in the context of clinical trials with a non-inferiority objective, a simulation study could be carried out for any proposed analysis, incorporating the specific assumptions being made and the choice of weights. For example, if combining the results of two trials each with $N = 400$, giving 80% power for a 10% non-inferiority margin on RD assuming true event rates of $p_{1j} = p_{2j} = 0.15$ in both strata, then the combined analysis has only 25% power with either of the random effects methods.

If the same total sample size were divided amongst four strata, the power for the **TDAS** method would increase to around 75%. This raises the intriguing possibility that in the situation with very few strata, both power and type I error could be improved by splitting the data into multiple smaller strata. For instance, if the results of two multi-centre trials are to be combined, then stratifying by centre within trial might be a pragmatic solution. Technically, this approach would rely on an assumption that the variability in θ between centres within trials is similar to the variability between centres across trials, and clearly it would be unable to distinguish between these two sources of variability.

In contrast, the corresponding power for the **SCAS** method would be 97%, therefore careful consideration should be given to whether homogeneous treatment effects can be assumed, in which case **SCAS** would be the preferred method. However, this decision should be justified on the basis of similarity of the trials, and not made on the basis of a non-significant result in the test for heterogeneity.[35,36] On the other hand, it has also been argued[34] that a small number of trials should not be combined in a meta-analysis at all if there is evidence of any heterogeneity.

Conditions with very small numbers of events need further evaluation, but early indications (not shown) suggest that type I error rates for **TDAS** can be slightly too low here, while those for **HKSJ** become inflated, perhaps due to the 'sparse data correction' of 0.5 added to the cells of strata containing zero cell counts. Implementation of a similar (perhaps smaller) adjustment within the **TDAS** method may deserve some consideration.

REFERENCES

[1] Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* **1985**; 41(1):55–68, doi:10.2307/2530643.

[2] 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, doi:10.1002/sim.6383.

[3] Friedrich JO, Adhikari NKJ, 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, doi:10.1186/1471-2288-7-5.

[4] Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **1998**; 17(8):857–872, doi:10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.

[5] Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **2001**; 57(3):963–971. Available at: http://www.jstor.org/stable/3068439 (Accessed 16.03.2017).

[6] Laud PJ, Dane A. Confidence intervals for the difference between independent binomial proportions: comparison using a graphical approach and moving averages. *Pharmaceutical Statistics* **2014**; 13(5):294–308, doi:10.1002/pst.1631.

[7] Hauck WW, Anderson S. A comparison of large-sample confidence interval methods for the difference of two binomial probabilities. *The American Statistician* **1986**; 40(4):318–322, doi:10.1080/00031305.1986.10475426.

[8] Haldane JBS. The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics* **1956**; 20(4):309–311, doi:10.1111/j.1469-1809.1955.tb01285.x.

[9] Anscombe FJ. On estimating binomial response relations. *Biometrika* **1956**; 43(3-4):461, doi:10.1093/biomet/43.3-4.461.

[10] Gart JJ. Approximate confidence limits for the relative risk. *Journal of the Royal Statistical Society. Series B (Methodological)* **1962**; 24(2):454–463. Available at: http://www.jstor.org/stable/2984236 (Accessed 16.03.2017).

[11] Yates F. Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society* **1934**; 1(2):217–235. Available at: http://www.jstor.org/stable/2983604 (Accessed 16.03.2017).

[12] Cornfield J. A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Health*, University of California Press: Berkeley, Calif., 1956; 135–148. Available at: http://projecteuclid.org/euclid.bsmsp/1200502552 (Accessed 16.03.2017).

[13] Miettinen O, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* **1985**; 4(2):213–226, doi:10.1002/sim.4780040211.

[14] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**; 26(4):404, doi:10.1093/biomet/26.4.404.

[15] Garwood F. Fiducial limits for the poisson distribution. *Biometrika* **1936**; 28(3-4):437, doi:10.1093/biomet/28.3-4.437.

[16] Agresti A, Min Y. Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 x 2 contingency tables. *Biometrics* **2005**; 61(2):515–523, doi:10.2307/3695972.

[17] Nurminen M, Mutanen P. Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics* **1987**; 14(1):67–77, doi:10.2307/4616049.

[18] Barker L, Cadwell BL. An analysis of eight 95 per cent confidence intervals for a ratio of Poisson parameters when events are rare. *Statistics in Medicine* **2008**; 27(20):4030–4037, doi:10.1002/sim.3234.

[19] Cai TT. One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference* **2005**; 131(1):63 – 88, doi:10.1016/j.jspi.2004.01.005.

[20] Senn S. *Statistical Issues in Drug Development*. 2 edn., John Wiley & Sons, Ltd., 2007, p264.

[21] Lu K. *Cochran-Mantel-Haenszel Weighted Miettinen and Nurminen Method for Confidence Intervals of the Difference in Binomial Proportions from stratified 2x2 samples*. American Statistical Association (Alexandria, VA), 2008. Available at: http://markstat.net/en/images/stories/lu_asa_2008.pdf (Accessed 16.03.2017).

[22] Nurminen M. Asymptotic efficiency of general noniterative estimators of common relative risk. *Biometrika* **1981**; 68(2):525–530, doi:10.1093/biomet/68.2.525.

[23] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* **1986**; 7(3):177–188, doi:10.1016/0197-2456(86)90046-2.

[24] Gart JJ, Nam Jm. Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension to multiple tables. *Biometrics* **1990**; 46(3):637–643, doi:10.2307/2532084.

[25] Tarone RE. On heterogeneity tests based on efficient scores. *Biometrika* **1985**; 72(1):91–95, doi:10.1093/biomet/72.1.91.

[26] Tarone RE. Homogeneity score tests with nuisance parameters. *Communications in Statistics - Theory and Methods* **1988**; 17(5):1549–1556, doi:10.1080/03610928808829697.

[27] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **2002**; 21(11):1539–1558, doi:10.1002/sim.1186.

[28] Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **1985**; 41(2):361–372, doi:10.2307/2530862.

[29] 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 Methodol* **2014**; 14:25, doi:10.1186/1471-2288-14-25.

[30] Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* **2001**; 20(24):3875–3889, doi:10.1002/sim.1009.

[31] Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* **2002**; 21(21):3153–3159, doi:10.1002/sim.1262.

[32] Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. *Computational Statistics and Data Analysis* **2006**; 50(12):3681 – 3701, doi:10.1016/j.csda.2005.07.019.

[33] Emerson JD, Hoaglin DC, Mosteller F. Simple robust procedures for combining risk differences in sets of 2 x 2 tables. *Statistics in Medicine* **1996**; 15(14):1465–1488, doi:10.1002/sim.4780151402.

[34] Gonnermann A, Framke T, Großhennig A, Koch A. No solution yet for combining two independent studies in the presence of heterogeneity. *Statistics in Medicine* **2015**; 34(16):2476–2480, doi:10.1002/sim.6473.

[35] Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Ltd., 2008, p280-282.

[36] Whitehead A. *Meta-analysis of controlled clinical trials*. John Wiley & Sons, Ltd., 2002, p152.