

# A new and improved confidence interval for the Mantel-Haenszel risk difference

Bernhard Klingenberg \*

Writing the variance of the Mantel-Haenszel estimator under the null of homogeneity and inverting the corresponding test, we arrive at an improved confidence interval for the common risk difference in stratified  $2 \times 2$  tables. This interval outperforms a variety of other intervals currently recommended in the literature and implemented in software. We also discuss a score-type confidence interval that allows to incorporate strata/study weights. Both of these intervals work very well under many scenarios common in stratified trials or in a meta analysis, including situations with a mixture of both small and large strata sample sizes, unbalanced treatment allocation or rare events. The new interval has the advantage that it is available in closed form with a simple formula. In addition, it applies to matched pairs data. We illustrate the methodology with various stratified clinical trials and a meta analysis. R code to reproduce all analysis is provided in an appendix. Copyright © 2013 John Wiley & Sons, Ltd.

**Keywords:** Difference of proportion; Matched Pairs; Meta Analysis; Score Interval; Stratified  $2 \times 2$  tables;

## 1. Wald confidence interval for the common risk difference

When analyzing stratified  $2 \times 2$  tables, computing pooled estimators along the lines of Cochran [1] and Mantel and Haenszel [2] is an established procedure. The pooled estimator for the common difference of proportion  $\delta = \pi_{i1} - \pi_{i2}$  in  $i = 1, \dots, K$  stratified  $2 \times 2$  tables is given by [3]

$$\hat{\delta}_{MH} = \frac{\sum_{i=1}^K w_i \hat{\delta}_i}{\sum_{i=1}^K w_i} = \frac{\sum_{i=1}^K w_i (y_{i1}/n_{i1} - y_{i2}/n_{i2})}{\sum_{i=1}^K w_i} = \frac{\sum_{i=1}^K (n_{i2}y_{i1} - n_{i1}y_{i2})/n_{i+}}{\sum_{i=1}^K w_i}, \quad (1)$$

where  $w_i = n_{i1}n_{i2}/n_{i+}$  are so-called Cochran weights and  $n_{i+} = n_{i1} + n_{i2}$  is the total sample size in stratum  $i$ . Unless otherwise noted, we assume that the  $y_{ij}$ 's are independent binomial  $\text{Bin}(n_{ij}, \pi_{ij})$ ,  $j = 1, 2$ . To obtain a confidence interval for  $\delta$ , Greenland and Robins (GR) [3] plugged sample proportions into the expression for  $\text{Var}[(n_{i2}y_{i1} - n_{i1}y_{i2})/n_{i+}]$  to obtain an estimate for  $\text{Var}[\hat{\delta}_{MH}]$ .

Under homogeneity of the risk difference one can write  $\pi_{i1} = \delta + \pi_{i2}$  or  $\pi_{i2} = \pi_{i1} - \delta$  for all  $i$ . Substituting these into the variance for the numerator of (1) gives two different expressions, which, when averaged, yield

$$\text{Var}[(n_{i2}y_{i1} - n_{i1}y_{i2})/n_{i+}] = E[\delta P_i + Q_i], \quad (2)$$

where

$$P_i = [n_{i1}^2 y_{i2} - n_{i2}^2 y_{i1} + n_{i1} n_{i2} (n_{i2} - n_{i1})/2] / n_{i+}^2 \quad \text{and} \quad Q_i = [y_{i1} (n_{i2} - y_{i2}) + y_{i2} (n_{i1} - y_{i1})] / (2n_{i+}).$$

Replacing  $\delta$  by  $\hat{\delta}_{\text{MH}}$  in (2) and ignoring the expected value yields the variance estimator of  $\hat{\delta}_{\text{MH}}$  due to Sato [4]

$$\widehat{\text{Var}}[\hat{\delta}_{\text{MH}}] = (\hat{\delta}_{\text{MH}} P + Q) / W^2, \quad (3)$$

where  $P = \sum_i P_i$ ,  $Q = \sum_i Q_i$  and  $W = \sum_i w_i$ . While the GR variance estimator is only consistent when all strata sample sizes  $n_{i+}$ ,  $i = 1, \dots, K$  get large, the Sato variance estimator is also consistent in the so called ‘‘sparse’’ asymptotic case, where  $K$  grows but the strata sample sizes  $n_{i+}$  may remain small. Based on the Sato variance estimator, a Wald-type confidence interval (CI) for  $\delta$  has form

$$\hat{\delta}_{\text{MH}} \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\delta}_{\text{MH}}]},$$

where  $z_\gamma$  is the upper  $\gamma$  quantile of the standard normal distribution. This CI is presented in survey articles on stratified or meta analysis for  $2 \times 2$  tables (e.g., Agresti and Hartzel [5]), while the one based on the GR variance estimator is presented in popular epidemiology textbooks (e.g., Rothman [6]) and seems to be the default procedure implemented in software on meta analysis, both commercial and free. Note that the Sato interval is equivalent to the acceptance region of the test  $H_0 : \delta = \delta_0$  vs.  $H_a : \delta \neq \delta_0$  using  $T = (\hat{\delta}_{\text{MH}} - \delta_0) / \widehat{\text{Var}}[\hat{\delta}_{\text{MH}}]^{1/2}$  as a test statistic, which is asymptotically standard normal.

We expect improved performance when estimating the variance under the null  $H_0 : \delta = \delta_0$ . When  $K = 1$  (i.e., a single  $2 \times 2$  table), inverting a test statistic that uses the null variance leads to an interval that is not symmetric around the point estimate  $\hat{\delta}_{\text{MH}}$  and shifted relative to the Wald interval, resulting in a marked improvement in the coverage probability. In the next subsection we will observe a similar effect for the stratified case and present a new CI for  $\delta$  available in closed form. In Section 2, we look at the score test for  $H_0$  and discuss how one can incorporate strata weights for it. This yields a statistic and corresponding score-type CI (solved numerically) that is similar to one proposed by Miettinen and Nurminen (MN) [7] for the stratified case. In Section 3 we show via simulation that both the new CI and the MN interval, in contrast to the recommended intervals, perform very well under many scenarios, including settings under a typical meta analysis or under the assumption of rare events. In addition, we show that the new interval is a very simple and competitive interval for the difference of proportions in matched pairs. However, we also point out that all intervals discussed here perform poorly under heterogeneity of the risk difference. Section 4 illustrates the methods for three examples involving stratified data from different scenarios. In Section 5 we summarize the findings and give recommendations. R code to replicate our analysis is given in the Appendix.

### 1.1. A null variance estimator

Similar to the single  $2 \times 2$  table case ( $K = 1$ ), we try to improve the performance of the GR and Sato Wald-type intervals above by estimating the variance under the null  $H_0 : \delta = \delta_0$ . Replacing  $\delta$  by  $\delta_0$  in (3), we obtain  $\widehat{\text{Var}}_{\delta_0}[\hat{\delta}_{\text{MH}}] = (\delta_0 P + Q) / W^2$  as the null variance estimator. Note that, by construction through (2),  $\widehat{\text{Var}}_{\delta_0}[\hat{\delta}_{\text{MH}}]$  is an unbiased estimator of the variance of  $\hat{\delta}_{\text{MH}}$  under the null. Then,  $T_0 = (\hat{\delta}_{\text{MH}} - \delta_0) / \widehat{\text{Var}}_{\delta_0}[\hat{\delta}_{\text{MH}}]^{1/2}$  is an alternative test statistic to  $T$  for  $H_0$ .

The QQ-plots in Figure 1 compare the exact null distribution (based on 1 Mio. simulations) of  $T$  and  $T_0$  to the standard normal one, under scenarios that we will revisit in the simulation studies in Section 3. We see that for many scenarios the null distributions of  $T$  and  $T_0$  are very close to standard normal. However, for some cases (Scenarios 4, 7 and 8, and to some extent 3) the null distribution of  $T$  is stochastically larger than standard normal in the upper tail, i.e.,  $T$  is liberal. When inverting, this leads to intervals that are too short and generally have coverage that is below the nominal level. On the other hand, for those same scenarios, the null distribution of  $T_0$  is stochastically smaller than the standard normal, resulting in intervals that are a bit conservative. However, the liberal performance of  $T$  is the greater concern, as we will see in Section 3. Not shown in Figure 1 is the null distribution of the GR test statistic, which is not competitive, see later.

To check the behavior of  $T$  and  $T_0$  under the alternative, where both statistics are asymptotically normal, we simulated data assuming  $\delta = 0.2$  but computed the test statistics under  $H_0 : \delta = \delta_0 = 0$ . Then,  $\hat{\delta}_{MH}$ , which is unbiased for  $\delta$  differs from  $\delta_0$  and we would expect some separation between  $T$  and  $T_0$  based on their denominators using  $\hat{\delta}_{MH}$  and  $\delta_0$ , respectively. For the scenarios with a small overall sample size, we found that  $T_0$  is slightly stochastically larger than  $T$ , resulting in more power. However, when the sample size is large (such as in Scenario 5 or 6) the two distributions are indistinguishable.

### 1.2. A new confidence interval for the Mantel-Haenszel risk difference

Inverting  $T_0$ , i.e., solving  $T_0 = \pm z_{\alpha/2}$  for  $\delta_0$  leads to a quadratic equation. Solving it yields the following closed-form solution for the upper and lower bound of the CI for  $\delta$ , where the midpoint and margin of error (ME) for the interval are given by

$$\text{new CI} : \hat{\delta}_{\text{Mid}} \pm \text{ME}, \quad \hat{\delta}_{\text{Mid}} = \hat{\delta}_{\text{MH}} + 0.5z_{\alpha/2}^2(P/W^2), \quad \text{ME} = \sqrt{\hat{\delta}_{\text{Mid}}^2 - \hat{\delta}_{\text{MH}}^2 + z_{\alpha/2}^2(Q/W^2)}.$$

This interval, as opposed to the Wald-type intervals above, is not symmetric about  $\hat{\delta}_{MH}$ , which is advantageous when the distribution of  $\hat{\delta}_{MH}$  is skewed. As we will see, this interval has excellent performance characteristics and yet is easy to compute. Under balanced sample sizes,  $\hat{\delta}_{\text{Mid}}$  is slightly biased as an estimator for  $\delta$  by a factor of  $(1 - w)$ , where  $w = 0.5z_{\alpha/2}^2/N_1$  and  $N_1 = \sum_i n_{i1}$ , but its variance is smaller by a factor of  $(1 - w)^2$  compared to  $\hat{\delta}_{MH}$ , typically resulting in an overall smaller mean squared error. In fact,  $\text{MSE}[\hat{\delta}_{\text{Mid}}] = \text{MSE}[\hat{\delta}_{\text{MH}}](1 - w)^2 + (w\delta)^2$ . For large  $N_1$ ,  $w \rightarrow 0$  and the MSE's are virtually identical, but Figure 2 shows some scenarios with  $N_1 = K \times n_{11} = 3 \times 25 = 75$  (or  $N_1 = 5 \times 25 = 125$ ) where they do differ. Overall,  $\hat{\delta}_{\text{Mid}}$  seems to be a better choice for the midpoint of a (symmetric) interval.

Sato [8] used similar arguments to above for constructing confidence intervals for a common odds ratio, risk ratio or rate ratio. For the difference between two homogeneous rates  $\lambda_{i1} - \lambda_{i2} = \lambda$  when each stratum consists of two independent Poisson observations  $Y_{ij} \sim \text{Pois}(n_{ij}\lambda_{ij})$ ,  $j = 1, 2$ , our statistic  $T_0$  is equivalent to one first derived in Sato [9].

Returning to the binomial case, Sato [10] proposed another interval for the Mantel-Haenszel risk difference using a counterfactual approach. If in strata  $i$  the  $n_{i1}$  subjects in group 1 would have been assigned to group 2 instead, we would expect to see  $n_{i1}\pi_{i2} = n_{i1}(\pi_{i1} - \delta) = n_{i1}\pi_{i1} - n_{i1}\delta$  successes. Estimating  $n_{i1}\pi_{i1}$  by  $y_{1i}$ , the actually observed number of successes, gives rise to the counterfactual (because we assume all subjects had been assigned to group 2)  $2 \times 2$  table with number of successes equal to  $y_{i1} - n_{i1}\delta$  (out of  $n_{i1}$ ) in group 1 and  $y_{i2}$  (out of  $n_{i2}$ ) in group 2. Using these expressions for the cell counts, one can write the well known Cochran-Mantel-Haenszel test statistic as a function of  $\delta$  to test  $H_0 : \delta = \delta_0$ . Inverting it yields a closed form solution for a CI for  $\delta$ , see [10]. We investigated two versions (with and without a continuity correction) of these closed form intervals. The one using the continuity correction was overly conservative throughout the parameter space. The one without the continuity correction performed very well when the true  $\delta = 0$ , but

suffered from a very large variability in the true coverage probability around the nominal level for values of  $\delta$  such as 0.1 or 0.2, sometimes resulting in coverages severely below the nominal level. For these reasons, we do not consider either of these two intervals further, given the available alternatives.

A completely different option for constructing a confidence interval is to estimate the variance of  $\hat{\delta}_{MH}$  directly under  $H_0$  without going through a relationship such as (2). This is discussed in the next section. Several other Wald-type intervals for  $\delta$  were reviewed in [11], many of which are based on heuristic arguments, such as estimating the standard deviation always at  $\delta = 0$ . None of these are competitive with the intervals proposed here.

## 2. The score interval for $\delta$ : Unweighted and weighted case

The score test statistic for  $H_0 : \delta = \delta_0$  for the case of a common  $\delta$  in multiple  $2 \times 2$  tables is given in Gart and Nam [12]. It equals

$$S = \frac{\left(\sum_{i=1}^K (y_{i1} - n_{i1}\tilde{\pi}_{i1}) / [\tilde{\pi}_{i1}(1 - \tilde{\pi}_{i1})]\right)^2}{\sum_{i=1}^K [\tilde{\pi}_{i1}(1 - \tilde{\pi}_{i1})/n_{i1} + \tilde{\pi}_{i2}(1 - \tilde{\pi}_{i2})/n_{i2}]^{-1}}, \quad (4)$$

where  $\tilde{\pi}_{i1}$  is the restricted MLE of  $\pi_{i1}$  under  $H_0$  and  $\tilde{\pi}_{i2} = \tilde{\pi}_{i1} - \delta_0$ .  $S$  is asymptotically Chi-squared with  $df = 1$  when the strata sample sizes  $n_{ij}$  become large. The CI for  $\delta$  is obtained by solving  $S = \chi_\alpha^2$  for  $\delta_0$ , where  $\chi_\alpha^2$  is the upper quantile of a Chi-square distribution with one degree of freedom. There is no closed-form solution, but the bounds can easily be computed numerically, e.g., by interval halving. Gart and Nam [12] also present a skewness-corrected version which might improve the performance, especially when  $n_{i1} \neq n_{i2}$ .

### 2.1. A score interval incorporating weights

Often, it is desired to weigh the contribution of a stratum (i.e., a center or a study) by a weight  $w_i$ , reflecting the amount of information provided by the stratum, precisely as is done in the Mantel-Haenszel estimator  $\hat{\delta}_{MH}$  in (1). This is especially popular in a meta analysis, where one weights the results from different studies before pooling them. The regular score and its skewness-corrected approach do not allow to incorporate explicit weights  $w_i$  and it is unclear just how the MLE for  $\delta$  weights the strata. (The MLE for  $\delta$  does not have a closed form solution, see the Appendix, but as a referee pointed out it is asymptotically equivalent to the inverse-variance weighted least squares estimator.) In this section we develop a score-type statistic that allows to incorporate weights  $w_i$  explicitly.

Let  $\hat{\pi}_{ij} = y_{ij}/n_{ij}$  denote the sample proportions. Rewriting  $\hat{\delta}_{MH}$  as  $\sum_{i=1}^K \sum_{j=1}^2 c_{ij}\hat{\pi}_{ij}$  with  $c_{i1} = w_i^*$  and  $c_{i2} = -w_i^*$  (where in our case  $w_i^* = w_i / \sum_i w_i$  are the standardized Cochran weights), we see that  $\hat{\delta}_{MH}$  is a linear combination of sample proportions. Let  $\delta_w = \sum_i \sum_j c_{ij}\pi_{ij}$  be the corresponding linear combination of success probabilities. For the score confidence interval for  $\delta_w$  (Andres et al. [13]) one inverts the score test of  $H_0 : \delta_w = \delta_{w0}$ ,

$$\frac{(\sum_i \sum_j c_{ij}\hat{\pi}_{ij} - \delta_{w0})^2}{\sum_i \sum_j c_{ij}^2 \tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})/n_{ij}},$$

where  $\tilde{\pi}_{ij}$  are the restricted (under  $H_0$ ) MLEs of  $\pi_{ij}$  subject to the constraint  $\sum_i \sum_j c_{ij}\pi_{ij} = \delta_{w0}$ .

Note that under homogeneity  $\delta_w = \sum_i w_i^*(\pi_{i1} - \pi_{i2}) = \delta \sum_i w_i^*$  reduces to  $\delta$ . Then, the above general approach motivates an alternative to the score interval of Gart and Nam that allows incorporation of weights: Invert the test  $H_0 : \delta = \delta_0$  using test statistic

$$S_w = \frac{(\sum_i \sum_j c_{ij}\hat{\pi}_{ij} - \delta_0)^2}{\sum_i \sum_j c_{ij}^2 \tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})/n_{ij}}, \quad (5)$$

where  $\tilde{\pi}_{ij}$  are the restricted MLEs of  $\pi_{ij}$  under homogeneity which are available in closed form (see Appendix). The denominator of  $S_w$  is simply the variance of the numerator estimated under the null of homogeneity. For  $K = 1$ ,  $S_w$  reduces to the regular score statistic for the difference of proportion [14] while for the special case of  $\delta_0 = 0$  and any  $K$ , (i.e., homogeneity or conditional independence in each of the  $K$   $2 \times 2$  tables)  $S_w$  is the Cochran statistic given in [1]. To obtain the lower and upper bound for the interval we solve  $S_w = \chi_\alpha^2$  for  $\delta_0$  numerically, e.g., through interval halving.

## 2.2. Miettinen-Nurminen statistic for common $\delta$

Note that when  $\delta_0 = 0$ ,  $\tilde{\pi}_{i1} = \tilde{\pi}_{i2} = (y_{i1} + y_{i2})/n_{i+}$  and  $E \left[ \sum_j \tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})/n_{ij} \right] = [(n_{i+} - 1)/n_{i+}] \text{Var}_{\delta=0}[\hat{\pi}_{i1} - \hat{\pi}_{i2}]$ , showing that the variance estimator in  $S_w$  is biased. This suggests the alternative statistic

$$S_{MN} = \frac{(\sum_i \sum_j c_{ij} \hat{\pi}_{ij} - \delta_0)^2}{\sum_i (\frac{n_{i+}}{n_{i+} - 1}) \sum_j c_{ij}^2 \tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})/n_{ij}} \quad (6)$$

which uses the bias-corrected (albeit only under the null  $H_0 : \delta = 0$ ) variance estimator of the numerator  $\sum_j c_{ij} \hat{\pi}_{ij}$  in the denominator.  $S_{MN}$  is essentially the same statistic that Miettinen and Nurminen [7] (see also [15]) proposed for the stratified  $2 \times 2$  case, although they used different weights. Also, as noted by [7], when  $\delta_0 = 0$ ,  $S_{MN}$  is equal to the Mantel-Haenszel statistic [2] for testing conditional independence. For the  $K = 1$  case, Newcombe and Nurminen [16] found that incorporating this bias correction is very effective in bringing the corresponding interval obtained by inverting  $S_{MN}$  close to the nominal level and that it performs better than the score interval [14], for all  $\delta$ . For large  $n_{i+}$ , the intervals based on inverting  $S_{MN}$  or  $S_w$  will be virtually identical. However, since  $S_{MN} < S_w$ , inverting the former leads to wider intervals when the strata sample sizes are small, the most extreme case being that of matched pairs, with  $n_{ij} = 1$  for all  $i$  and  $j$ . Figure 1 shows the exact distribution of  $S_{MN}$  alongside those for  $T$  and  $T_0$ . Under Scenarios 1 through 6, its distribution is very close to standard normal. For Scenarios 7 and 8 (rare events and matched pairs), it seems to be conservative, while for Scenario 9 (heterogeneity), it is liberal.

## 3. Simulation studies to evaluate performance

Sections 1 and 2 introduced various CIs for the common difference of proportion. In this section via evaluate their coverage probability, average length and power via extensive simulation under many different scenarios. Figures 3 and 4 show boxplots of the coverage probability under balanced ( $n_{i1} = n_{i2} = 25$  for all  $i = 1, \dots, K$ ) and unbalanced ( $n_{i2} = 15, n_{i1} = 3n_{i2} = 45$  for all  $i = 1, \dots, K$ ) strata sample sizes, when the number of strata  $K = 3, 5$  or  $10$  and the true difference of proportion  $\delta = 0, 0.1$  or  $0.2$ . For each  $(K, \delta)$  combination, we generated 500 different sets of true strata probabilities  $\pi_{i2}, i = 1, \dots, K$  for the control group from the uniform distribution on  $[0, 1 - \delta]$  and set  $\pi_{i1} = \pi_{i2} + \delta$ . Then, with each of the 500 sets of strata probabilities we generated 7600 datasets (i.e.,  $\text{binomial}(n_{ij}, \pi_{ij})$  responses), computed the CI for each dataset and checked if  $\delta$  was contained in it. This estimates the true coverage probability to within a margin of error of  $\pm 0.5\%$ . The boxplots in Figures 3 and 4 show the distribution of these 500 estimated coverage probabilities for the different methods.

It is evident from these plots that the new CI introduced in Section 1.1 and the Miettinen-Nurminen (MN) type interval proposed in Section 2.2 outperform the rest under all considered scenarios. Their median coverage is very close to the nominal level of 95%, with little spread around that value. In contrast, the 75th percentile of the coverage for the recommended GR or Sato intervals stay below 95% in almost all cases. While the Sato interval gets closer to the nominal value of 95% with increasing strata sample sizes, the coverage of the GR intervals stays significantly below 95%

(simulation results not shown). The score-type interval (5) seems to be consistently inferior to the MN-type interval. Not shown in Figures 3 and 4 are the performances of the Gart and Nam interval and its skewness corrected version. For both of these, the spread of the coverage was very large and the 25th percentile of the coverage was close to only 91%. These two intervals did not perform well in any of the simulation studies we ran and are dropped from further considerations.

Although the new interval looks to be a bit more conservative (see also Figure 1), it is actually as powerful as the MN-type CI. For the balanced case with  $n_{ij} = 25$  and  $\delta = 0.1$ , the average power to find  $\delta > 0$ , estimated by the percent of times (out of 500 simulations) that the lower bound is strictly larger than zero equaled 30%, 46% and 75% for  $K = 3, 5$  and 10 for both procedures. (The standard deviation of the estimated power was 6% throughout.) Similarly, the average power to find e.g.,  $\delta > 0.05$  or 0.1 when  $\delta = 0.2$  was almost the same (e.g., 95% and 73%, respectively, when  $K = 10$ ). The two intervals also have a very similar average length. Finally, for all intervals presented in Figures 3 and 4, the lower and upper non-coverage, i.e., the probability that the lower (upper) bound is above (below)  $\delta$  was fairly symmetric at around 2.5%. For the MN interval, it was almost perfectly 2.5% for all considered scenarios, while for the new interval, the lower non-coverage was 2.2% and the upper non-coverage was 2.7% for some cases.

### 3.1. Large number of strata with both large and small sample sizes

We further evaluated the different intervals under a scenario with a mix of strata with large and small sample sizes and a fairly large number of strata (we chose  $K = 20, 30$  and 60). To simulate datasets with these features, we first generated sample sizes  $(n_{11}, \dots, n_{1K})$  following a multinomial distribution with overall sample size  $N_1 = \sum_j n_{1j}$  in the treatment group and probability vector  $(\tau_1, \dots, \tau_K)$  following a Dirichlet distribution with parameters all equal to 1. (For a rationale for these simulation settings in terms of an underlying Poisson process mimicking recruitment of patients to centers see [17]). This results in a sample size distribution where a typical dataset would have many strata with small to medium and a few strata with large sample size. Figure 5 shows the sample size distribution in the treatment arm under three such settings. We first consider a scenario of very sparse sets of  $K = 20$   $2 \times 2$  tables, with a total sample size of only  $N = 150$  in the treated and control group, resulting in an average strata size of just over 7 (similar to the first example in Section 4). Figure 6a shows the coverage probability under balanced sample sizes ( $n_{i1} = n_{i2}$ ) when  $\delta = 0, 0.1$  and 0.2. We also considered one balanced (Figure 6b,  $n_{i1} = n_{i2}, N_1 = \sum_{i=1}^K n_{i1} = 1000, K = 30, 60$ ) and one unbalanced (Figure 6c,  $n_{i1} = 3n_{i2}, N_1 = \sum_{i=1}^K n_{i1} = 1300, K = 30, 60$ ) case with both large and small sample sizes per cluster, where the unbalanced case mimics the example in Section 4. Under all scenarios in Figure 6, the new interval and the MN-type interval perform very well, while this is not the case for the other intervals, especially in the settings with more sparse data.

### 3.2. Rare Events

Pooling results from different centers/studies is especially valuable when the event rate is very rare (i.e., both  $\pi_{i1}$  and  $\pi_{i2}$  are very small) and there are no or only very few events per stratum. A couple of papers [18, 19, 20] investigated the performance of various confidence intervals under such settings and mentioned the unsatisfactory behavior for the GR interval. Figure 7a shows the behavior of the CIs discussed in this article (the score and MN-type intervals performed virtually identical, so we only show results for the MN-type interval) when  $\pi_{i2}$  is selected randomly from the uniform distribution on  $[0, 0.005]$  and  $\delta$  is small (e.g., 0.004). For the simulations, we selected the same sample sizes as in a dataset given in [19] which investigated rare adverse events of a pharmaceutical product in  $K = 17$  centers. Overall, 19 of the 5126 patients under treatment had an event (0.37%) compared to 5 out of 1941 patients under control (0.26%). The allocation to treatment or control was unbalanced (roughly 3:1) and in 8 of the 17 centers no adverse events were observed. The

simulation results show that even under such extreme settings, the new interval performs reasonable, while the standard GR and the Sato interval have serious deficiencies in terms of coverage below the nominal level. The MN-type interval is a bit conservative, leading to a loss in power compared to the new interval (and to the GR and Sato interval when these hold the nominal level). Here, we have also included the performance of the interval one gets by collapsing (which may be justified when all the event rates in the control group are similar, which is the case here as all event rates are small) the  $K \times 2 \times 2$  tables into a single table showing the overall success and failure counts. This is often done when dealing with rare events and one then computes the regular MN-interval (or another interval) for the risk difference in the single  $2 \times 2$  table. This procedure also seems to perform acceptable in terms of coverage, but it is less powerful. For instance, the power to declare  $\delta > 0$  for the collapsed, new and MN-type interval is (57%, 64%, 56%) when  $\delta = 0.004$  and (86%, 89%, 86%) when  $\delta = 0.006$ . There seems to be no reason to use the collapsed interval.

Note that the Sato and new interval return a valid interval as long as there is at least one success (or at least one failure) among all  $K$  strata. For strata where both treatment groups show no success ( $y_{i1} = y_{i2} = 0$ ) and the sample size is balanced we get  $P_i = Q_i = 0$ , but they still contribute to the point estimate and margin of error, moving the confidence interval closer to zero. In standard software for meta analysis which computes the GR interval it is customary to add 0.5 to the success count in each group when  $y_{i1} = y_{i2} = 0$ . Another strategy is to drop centers with no successes from the analysis. For the dataset mentioned above this would mean throwing away the information provided by almost half of all centers included in the meta analysis, although they *are* informative about the risk difference. Further, we don't feel adding an arbitrary constant is necessary, as clearly the simulation studies show that intervals (in particular the new interval) perform well without such an adjustment. If there are no successes in any of the  $K$  strata (a rather extreme case), the Sato and the new interval return  $[0, 0]$ , while the MN-type interval still produces a sensible interval around 0.

### 3.3. Matched Pairs

Another extreme scenario is that of matched pairs, where a stratum consists of a single observation in each group, i.e.,  $n_{i1} = n_{i2} = 1$ . Then, we can collapse the data into a single  $2 \times 2$  table with cells showing the counts (out of the  $K$  matched pairs) for the 4 different response types  $(y_{i1}, y_{i2}) = (1, 1), (1, 0), (0, 1)$  and  $(0, 0)$ . The hypothesis of  $H_0 : \delta = \delta_0$  (i.e., a common difference of proportion of  $\delta_0$  for each matched pair) means that successes in group 1 are by  $\delta_0$  higher than in group 2, for each pair. Referring to the collapsed table, this means that the proportion of  $(1, 0)$  and  $(1, 1)$  outcomes (all pairs with a success in group 1) is by  $\delta_0$  higher than the proportion of  $(0, 1)$  and  $(1, 1)$  outcomes (all pairs with a success in group 2). Equivalently, the proportion of  $(1, 0)$  outcomes is by  $\delta_0$  larger than the proportion of  $(0, 1)$  outcomes. (When  $\delta_0 = 0$ , this is the usual symmetry or marginal homogeneity hypothesis.)

Under the matched pairs setting  $\hat{\delta}_{MH}$  in (1) reduces to  $(b - c)/K$ , where  $b$  is the number of pairs with responses  $(y_{i1}, y_{i2}) = (1, 0)$  and  $c$  the number of pairs with  $(y_{i1}, y_{i2}) = (0, 1)$ . Since  $P = (c - b)/4$  and  $Q = (b + c)/4$ , the variance estimate (2) reduces to  $\widehat{\text{Var}}[\hat{\delta}_{MH}] = [(b + c) - (b - c)^2/K] / K^2$  and the Sato interval is simply the regular Wald interval for the difference of proportion in the paired case. The new interval in Section 1.1 is then an interesting alternative with a simple closed-form formula that should perform better, especially when the number of matched pairs is small. Tango [21] derived the score interval for the difference of proportions in the paired case, and [22] recently showed that it has a closed form solution, although it is very complicated. This could be a reason to prefer the new interval.

We simulated the performance of the different methods with  $K = 25, 50$  or  $100$  matched pairs,  $\pi_{i2}$  uniform(0,1) and  $\delta = 0, 0.05, 0.1$  and  $0.2$ . Coverage probabilities (estimated as before by simulating 7600 datasets, each with  $K$  matched pairs and  $p_{i2}$  drawn from the uniform distribution and repeated 500 times to form each boxplot) are shown in Figure 7b. As expected the Sato (i.e. Wald) interval performs inadequately, with coverage probability well below the nominal level, even

for a sample size of 100 pairs. By contrast, the new interval performs very well and is on par with Tango's score interval, which in [23] was shown to perform the best among several competitors. (For  $K = 25$ , the new interval seems to do better for small  $\delta$ , but worse for larger  $\delta$ ). The MN-type interval seems to be very conservative, although this doesn't show in terms of the average power to declare  $\delta > 0$ . For instance, for  $K = 100$ ,  $\delta = 0.2$ , the average coverage probability for the new, Tango and MN interval are (95.1%, 95.1%, 98.6%) with average power in terms of the lower bound being larger than zero of (90.3%, 89.6%, 89.6%). However, in terms of the lower bound being larger than 0.05, the average power is given by (67.2%, 67.2%, 0%), explained by the average length, which is (0.238, 0.240, 0.291) for the three intervals. Collapsing the  $K$  tables to a single table showing the success and failure counts for each group leads to unnecessarily wide intervals and a substantial loss of power. This procedure as well as the GR and score interval (not shown in Figure 7b) are not recommended.

### 3.4. Performance under heterogeneity

One assumption of all methods presented in this paper is the homogeneity of the risk difference. To investigate the performance under heterogeneity, we simulated datasets when  $\pi_{i1}$  and  $\pi_{i2}$  are modeled through a generalized linear mixed model with identity link,  $\pi_{i1} = \alpha + (\beta + b_i)/2 + u_i$ ,  $\pi_{i2} = \alpha - (\beta + b_i)/2 + u_i$ , where  $u_i$  and  $b_i$  are independent random effects with a  $N(0, \sigma_u)$  and  $N(0, \sigma_b)$  distribution, respectively (Agresti and Hartzel, [5]). Then, the difference of proportion  $\delta_i = \beta + b_i$  varies like a normal random variable with standard deviation  $\sigma_b$ . Note that while simulating responses under such a model is not problematic, fitting a GLMM with identity link to a given dataset is often not possible as the formulation with a normal random effect implies values outside the permissible range for  $\pi_{ij}$ .

Figure 8 shows coverage probabilities for various methods when simulating datasets with  $\alpha = 0.4$ ,  $\sigma_u = 0.05$ ,  $\beta = 0.1$  and  $\sigma_b = 0.01, 0.02, 0.03$  and  $0.04$ . When  $\sigma_b = 0.01(0.04)$ , 95% of the times the  $\delta_i$ 's are within  $\pm 0.02$  ( $\pm 0.08$ ) of  $\beta = 0.1$ . We defined the coverage probability as the proportion of times the interval covers the expected value of  $\delta_i$ , which is  $\beta$ . Under little heterogeneity ( $\sigma_b = 0.01$ ) the coverage probabilities of the new interval and the MH-type interval are still acceptable, being very close to nominal, albeit slightly below. However, under considerable heterogeneity ( $\sigma_b = 0.04$ ), the coverage drops significantly below the nominal level and the spread also increases. Under such a scenario, the intervals proposed here should not be used.

## 4. Three Examples

Lipsitz et al. [24] mention a randomized trial carried out in  $K = 21$  institutions comparing two chemotherapy treatments with respect to survival (lived/died by the end of the study) in patients with multiple myeloma. The data (available online, see Appendix) show small sample sizes, with an average of 7.4 patients per institution. The MH estimate of a common difference of proportion equals 0.057 and the new CI proposed in Section 1.1 is given by  $[-0.102, 0.211]$  ( $\hat{\delta}_{\text{Mid}} = 0.055$ ,  $\text{ME} = 0.157$ ). It indicates no significant difference between the two therapies. The Sato interval equals  $0.057 \pm 1.96 \times 0.080 = 0.057 \pm 0.157 = [-0.099, 0.214]$ , but its overall performance was shown to be inferior under such settings (see Figure 4a). The more computationally complex MN interval equals  $[-0.010, 0.206]$ .

In a recent vaccine trial, a flu vaccine was compared to placebo in  $K = 56$  different centers where a success was defined as not developing the flu over a period of 3 months after vaccination. The sample sizes in the treatment group varied from 17 to 65 per center and, due to a 3:1 allocation ratio between treatment and placebo, only 6 to 21 in the placebo group. Overall,  $N_1 = 1346$  patients were randomized to the vaccine and  $N_2 = 460$  to the placebo group. Table



1 shows results for some of the centers. Based on the information in all 56 centers, the MH estimator for the common risk difference between the vaccinated group and the placebo group equals 0.030 and the new 95% confidence interval for it is given by  $[-0.007, 0.065]$  ( $\hat{\delta}_{\text{Mid}} = 0.029$ ,  $\text{ME} = 0.036$ ). The Sato interval equals  $0.030 \pm 0.036 = [-0.006, 0.066]$  and the MN interval is  $[-0.008, 0.064]$ . Based on the simulation results in Figure 4c, all three intervals are trustworthy to have an actual coverage rate of 95% and reflect the tendency for a higher success rate with the vaccine, but no statistical significance can be reached. The GR interval, which was too liberal in the simulations (median coverage of only 93.5%) equals  $[-0.004, 0.064]$ , and the Gart and Nam interval (with skewness correction because of the unbalanced sample size) is given by  $[-0.012, 0.073]$ , but it also cannot be recommended based on the simulation results.

For the meta analysis of rare adverse events in a drug safety study mentioned in Section 3.1.1 (data available online, see Appendix), we obtain  $\hat{\delta}_{\text{MH}} = 0.10\%$  with the new 95% CI given by  $0.07\% \pm 0.073\% = [-0.21\%, 0.35\%]$ . By comparison, the GR and Sato interval equal  $[-0.18\%, 0.38\%]$  (symmetric around 0.10%), but these are not to be trusted based on the simulation results. The MN-type interval equals  $[-0.27\%, 0.38\%]$  and, based on the simulation results, might be a bit conservative.

## 5. Summary and Recommendations

We have proposed a new interval and discussed a Miettinen-Nurminen (MN) type interval for the risk difference in stratified  $2 \times 2$  tables in the homogeneous case. The new interval is very easy to compute (closed-form expression) and hence suitable to be presented in introductory lectures and books. Its properties are as good or better than the more complicated MN-type interval and it also performs excellent under the matched pairs case. Both of these intervals outperform currently recommended methods in the literature and implemented in software (e.g., the GR interval) in terms of the coverage probability and work very well for sparse and/or unbalanced data with small or large number of strata. The MH risk difference is an important and popular parameter in a meta analysis, and we have shown that under typical settings the new interval (as well as the Sato and MN-type interval) are preferable to the GR interval. Pooling studies in a meta analysis is especially valuable when the event of interest is rare and the new interval has a very good coverage performance in this situation. In addition, it is not necessary to exclude studies from the meta analysis that show no successes, which would bias results for the risk difference, nor is it necessary to add arbitrary constants to cell counts to make them positive.

Under heterogeneity, when the difference of proportion varies among strata, the new and MN-type CI perform satisfactorily as long as the variation is relatively small. However, if considerable heterogeneity exists, none of the intervals presented here works well. Research is currently underway to evaluate methods under such circumstances. For instance, one possibility is to compute, for each  $2 \times 2$  table separately, the confidence interval for  $\delta_i$  and then form a weighted average of the midpoints and margin of errors to obtain one for the expected value of  $\delta_i$ . An alternative is to use the methods of [25] or [26] to combine lower and upper bounds of such intervals. Also, in this article we only discussed two-sided intervals, but one-sided bounds are of equal importance. All methods discuss here allow to construct one-sided bounds. Finally, we defined all intervals in terms of Cochran weights, but many other alternative weighting schemes exists and it will be interesting to see how these might effect the performance under various scenarios.

## 6. Appendix

### 6.1. R code and datasets

The following can be directly copied and pasted into R. After scouring the necessary functions (first line) and loading the dataset (second line), we show how to compute the new, MN-type, Sato and GR CI:

```
> source(file="http://sites.williams.edu/bklingen/files/2013/06/stratMHRD.r")
> ## Myeloma dataset in Lipsitz et al. (1998):
> myeloma <- read.table(file="http://sites.williams.edu/bklingen/files/2013/06/myel.txt", header=TRUE)
> head(myeloma)
  y1 n1 y2 n2
1  3  4  1  3
2  3  4  8 11
3  2  2  2  3
4  2  2  2  2
5  2  2  0  3
6  1  3  2  3
> strat.MHRD(myeloma) #new interval
$delta.MH
[1] 0.05716832
$delta.Mid
[1] 0.05466977
$pseudo.se
[1] 0.04076493
$CI
      [,1]      [,2]
[1,] -0.101927 0.2112666
$confllev
[1] 0.95
$var.estimator
[1] "Sato0"

> strat.MHRD.MN(myeloma) #Miettinen-Nurminen type interval
$delta.MH
[1] 0.05716832
$CI
[1] -0.09968344 0.20645934
$confllev
[1] 0.95

> strat.MHRD(myeloma, method="Sato") #Sato interval
$delta.MH
[1] 0.05716832
$se.delta.MH
[1] 0.07988762
$CI
[1] -0.09940854 0.21374519
$confllev
[1] 0.95
$var.estimator
[1] "Sato"

> strat.MHRD(myeloma, method="GR") #Greenland-Robins interval
$delta.MH
[1] 0.05716832
$se.delta.MH
[1] 0.06319183
$CI
[1] -0.0666854 0.1810220
$confllev
[1] 0.95
$var.estimator
[1] "Greenland-Robins"

> ## Adverse Event dataset in Bradburn et al. 2007:
> adevent <- read.table(file="http://sites.williams.edu/bklingen/files/2013/06/adevents.txt", header=TRUE)
> head(adevent)
```

```

  y1  n1  y2  n2
1  0 432  0 142
2  1 375  0 125
3  0  80  0  40
4  0 248  1  84
5  0  50  0  24
6  2 251  0  85
> strat.MHRD(myeloma) #new interval
$delta.MH
[1] 0.05716832
$delta.Mid
[1] 0.05466977
$pseudo.se
[1] 0.04076493
$CI
      [,1]      [,2]
[1,] -0.101927 0.2112666
$conflv
[1] 0.95
$var.estimator
[1] "Sato0"

```

## 6.2. Restricted MLEs

Let

$$l(\delta, \boldsymbol{\pi}_1) = \sum_{i=1}^K l_i(\delta, \pi_{i1}) = \sum_{i=1}^K y_{i1} \log(\pi_{i1}) + (n_{i1} - y_{i1}) \log(1 - \pi_{i1}) + y_{i2} \log(\pi_{i1} - \delta) + (n_{i2} - y_{i2}) \log(1 - \pi_{i1} + \delta) \quad (7)$$

denote the log-likelihood for the  $K \times 2$  tables, with  $\delta$  the common difference of proportion and  $\boldsymbol{\pi}_1 = (\pi_{11}, \dots, \pi_{K1})$  as nuisance parameters. Under  $H_0 : \delta = \delta_0$ , we know from the  $K = 1$  setting [7, 27] that maximizing  $l_i(\delta, \pi_{i1})$  under  $\pi_{i1} - \pi_{i2} = \delta_0$  yields a closed form solution for  $\pi_{i1}$ . Denote it by  $\tilde{\pi}_{i1}$  and let  $\tilde{\pi}_{i2} = \tilde{\pi}_{i1} - \delta_0$ . These are the restricted MLEs appearing in the score statistic (4) and in (5) and (6).

To obtain the MLE for  $\delta$ , note that taking partial derivatives of (7) w.r.t. the  $\pi_{i1}$ 's and setting them equal to zero yields  $\pi_{i1} = \hat{\pi}_{i1}, i = 1, \dots, K$ . Plugging these into the partial derivative w.r.t.  $\delta$  and solving yields the equation

$$\sum_{i=1}^K y_{i1} / (\hat{\pi}_{i2} + \delta) = \sum_{i=1}^K (n_{i1} - y_{i1}) / (1 - \hat{\pi}_{i2} - \delta),$$

the solution of which is the MLE for  $\delta$ . Note that no closed formula for the MLE can be given and the equation has to be solved numerically.

## References

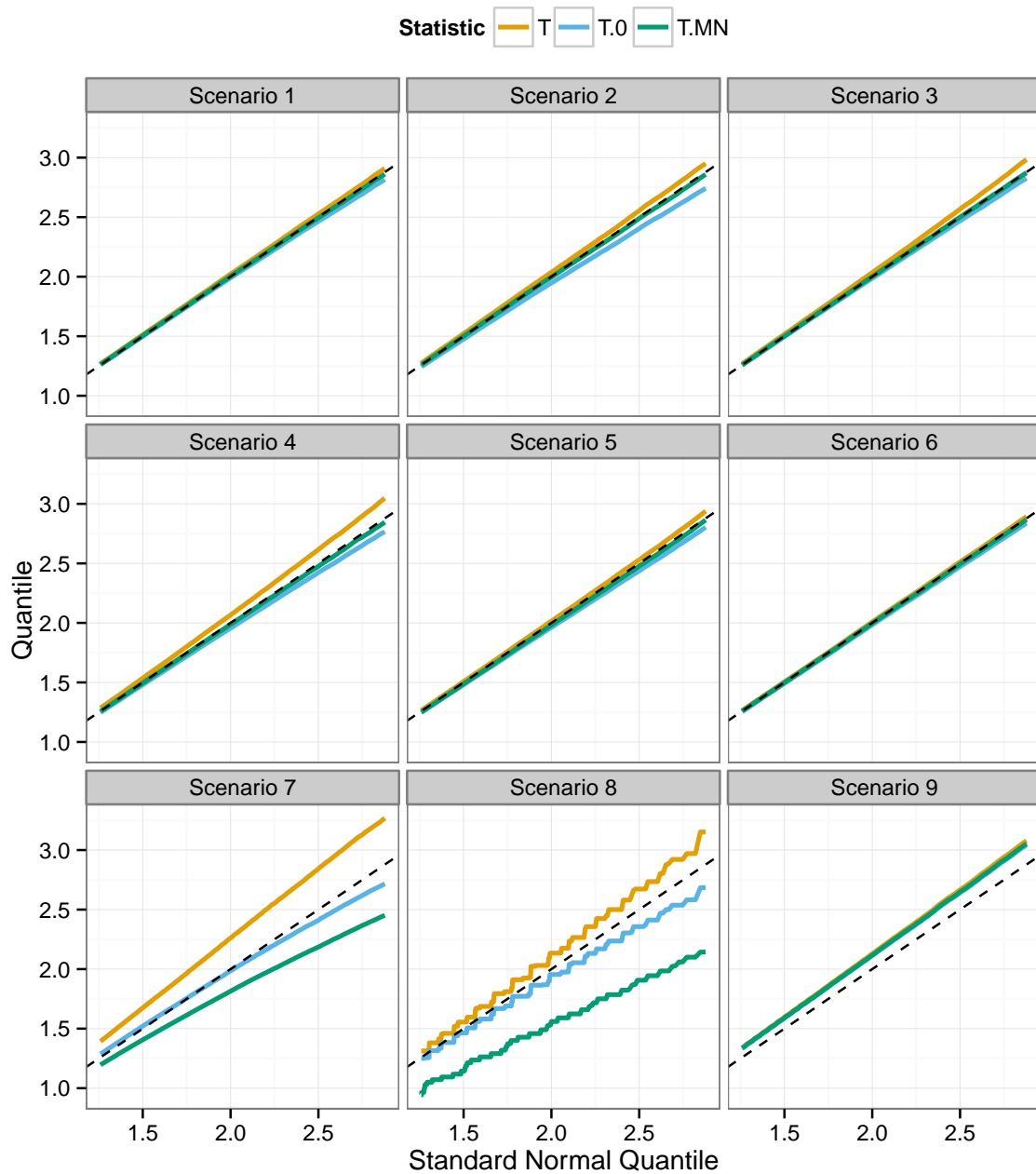
1. Cochran W. Some methods for strengthening the common  $\chi^2$  tests. *Biometrics* 1954; **10**:417–451.
2. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* 1959; **22**:719–748.
3. Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 1985; **41**:55–68.
4. Sato T. On the Variance Estimator for the Mantel-Haenszel Risk Difference (letter). *Biometrics* 1989; **45**:1323–1324.
5. Agresti A, Hartzel J. Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine* 2000; **19**:1115–1139.
6. Rothman K. 2002. *Epidemiology: an introduction*. Oxford University Press.
7. Miettinen O, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* 1985; **4**:213–226.
8. Sato T. Confidence limits for the common odds ratio based on the asymptotic distribution of the Mantel-Haenszel estimator. *Biometrics* 1990; **46**:71–80.
9. Sato T. Confidence intervals for effect parameters common in cancer epidemiology. *Environmental Health Perspectives* 1990; **87**:95–101.
10. Sato T. A further look at the Cochran-Mantel-Haenszel risk difference. *Controlled Clinical Trials* 1995; **16**:359–361.

11. Sanchez-Meca J, Marin-Martinez F. Meta-analysis of 2x2 tables: Estimating a common risk difference. *Educational & Psychological Measurement* 2001; **61**:249–276.
12. Gart J, Nam J. Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension to multiple tables. *Biometrics* 1990; **46**:637–643.
13. Andres M, Hernandez A, Tejedor H. Inferences about a linear combination of proportions. *Statistical Methods in Medical Research* 2011; **20**:369–387. (Errata in **21**:427).
14. Mee R. Confidence bounds for the difference between two probabilities (letter). *Biometrics* 1984; **40**:1175–1176.
15. Lu K. Cochran-Mantel-Haenszel weighted Miettinen & Nurminen method for confidence intervals of the difference in binomial proportions from stratified 2x2 samples. *JSM Proceedings* 2008; Denver, CO: American Statistical Association.
16. Newcombe R, Nurminen M. In defence of score intervals for proportions and their differences. *Communication in Statistics - Theory and Methods* 2011; **40**: 1271–1282.
17. Lin Z. An issue of statistical analysis in controlled multi-center studies: How shall we weight the centers? *Statistics in Medicine* 1999; **18**:365–373.
18. Tian L, Cai T, Pfeffer M, Piankov N, Cremieux P-Y, Wei LJ. Exact and efficient inference procedure for meta analysis its application to the analysis of independent  $2 \times 2$  tables with all available data but without artificial continuity correction. *Biostatistics* 2009; **10**:275–281.
19. Bradburn M, Deeks J, Berlin J, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine* 2007;**26**:53–77.
20. Tianxi C, Parast L, Ryan L. Meta-analysis for rare events. *Statistics in Medicine* 2010;**29**:2078–2089.
21. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired sample design. *Statistics in Medicine* 1998;**17**(8):891–908.
22. Yang Z, Sun X, Hardin J. A non-iterative implementation of Tangos score confidence interval for a paired difference of proportions. *Statistics in Medicine* 2013; **32**: 1336–1342.
23. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 1998;**17**(22):2635–2650.
24. Lipsitz S, Dear K, Laird N, Molenberghs G. Tests for homogeneity of the risk difference when data are sparse. *Biometrics* 1998;**54**:148–160.
25. Price RM, Bonett DG. An improved confidence interval for a linear function of binomial proportions. *Computational Statistics & Data Analysis* 2004;**45**:449–456.
26. Zou G. Y., Huang W. and Zhang X. A Note on Confidence Interval Estimation for a Linear Function of Binomial Proportions. *Computational Statistics & Data Analysis* 2009; **53**:1080–1085.
27. Nurminen M. Confidence intervals for the ratio and difference of two binomial proportions. *Biometrics* 1986; **42**:675–676.

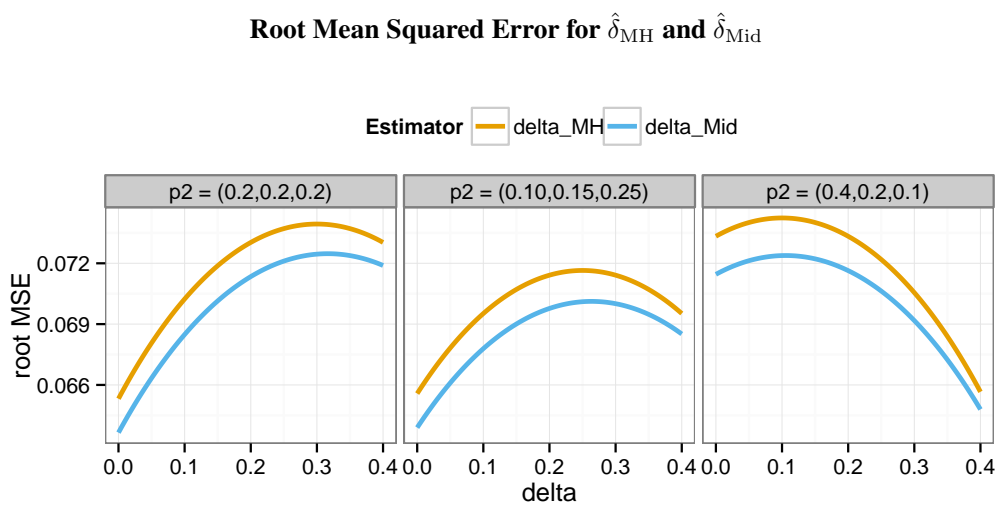
**Table 1.** Data collected in  $K = 56$  centers comparing the success of a flu vaccine. Shown are the sample proportions of patients not developing the flu over a three-month period after vaccination, for the treated and control group and their differences for a few selected centers. The column labeled  $w_i^*$  refers to the standardized Cochran weight for each center, and the centers are ordered according to this weight.

Center	$y_{i1}$	$n_{i1}$	$y_{i2}$	$n_{i2}$	$\hat{\pi}_{i1} - \hat{\pi}_{i2}$	$w_i^*$	Center	$y_{i1}$	$n_{i1}$	$y_{i2}$	$n_{i2}$	$\hat{\pi}_{i1} - \hat{\pi}_{i2}$	$w_i^*$
1	6	58	2	16	-0.02	0.04	38	4	18	0	6	0.22	0.01
10	4	27	1	9	0.04	0.02	46	5	17	1	6	0.13	0.01
16	4	25	0	9	0.16	0.02	47	3	16	2	7	-0.10	0.01
22	1	21	1	8	-0.08	0.02	56	2	18	1	5	-0.09	0.01

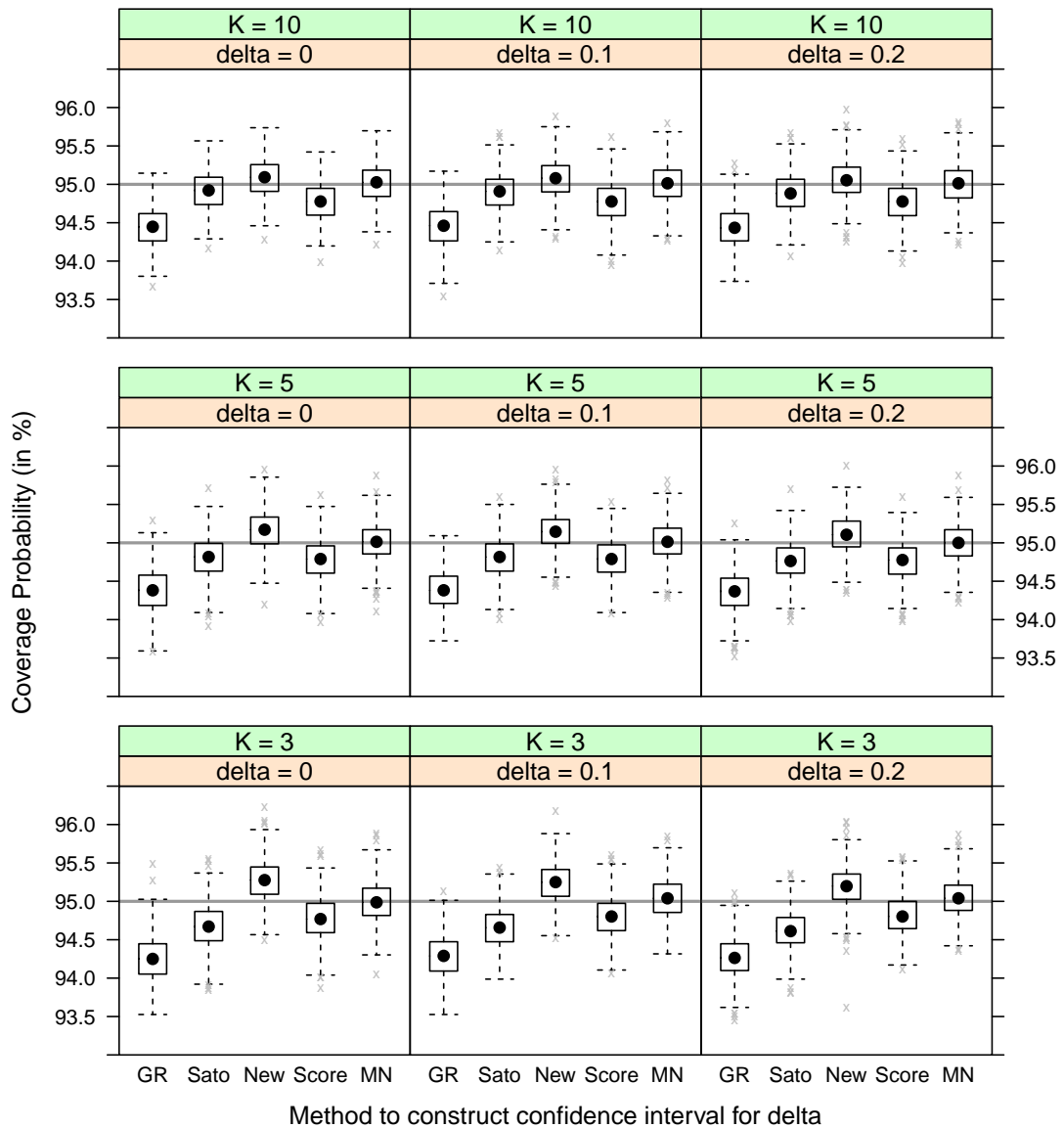
## Null distribution of $T$ , $T_0$ and $T_{MN}$



**Figure 1.** QQ-plot for the distribution of  $T$  (yellow),  $T_0$  (blue) and  $S_{MN}$  (green) when testing  $H_0 : \delta = \delta_0$  under different scenarios. The black dashed line corresponds to quantiles from the standard normal distribution. The null distribution was simulated under  $\delta = 0$  and  $\delta = 0.2$  under balanced (Scenarios 1 & 2 with  $n_1 = n_2 = 30$ ,  $K = 5$ ) and unbalanced (Scenarios 3 & 4:  $n_1 = 3n_2 = 45$ ,  $K = 5$ ) situations. Scenario 5 corresponds to having small sample sizes in each stratum ( $\delta = 0$ ,  $K = 21$ , overall sample size in group 1  $N_1 = 150$ , balanced), while Scenario 6 ( $\delta = 0$ ,  $K = 30$ ,  $N_1 = 1000$ , balanced) correspond to having many strata and with a mixture of both large and small sample sizes as is typically the case in a meta analysis. Scenario 7 simulates the situation of rare events ( $\pi_{i2} \sim U(0, 0.005)$ ,  $\delta = 0.002$ ,  $K = 17$ ,  $N_1 = 5125$ , unbalanced), while Scenario 8 corresponds to observing matched pairs ( $\delta = 0$ ,  $n_1 = n_2 = 1$ ,  $K = 50$ ). Finally, Scenario 8 simulated the data under heterogeneity of the risk difference.

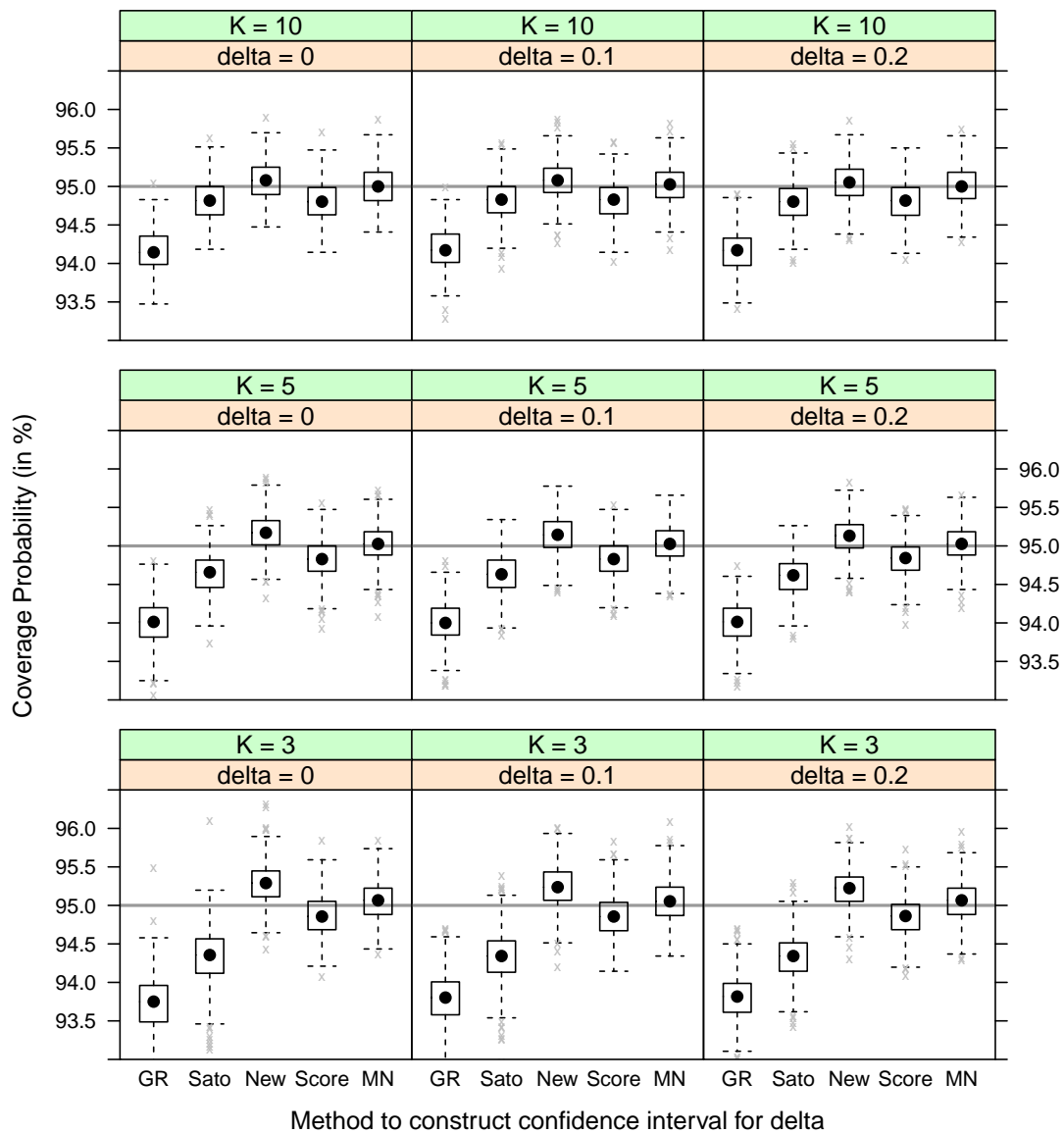


**Figure 2.** Comparison of the root MSE of  $\hat{\delta}_{MH}$  and  $\hat{\delta}_{Mid}$  when estimating  $\delta$  under true strata probabilities  $\pi_{12} = \pi_{22} = \pi_{32} = 0.2$  (first panel),  $\pi_{12} = 0.10$ ,  $\pi_{22} = 0.15$ ,  $\pi_{32} = 0.25$  (second panel) and  $\pi_{12} = 0.4$ ,  $\pi_{22} = 0.2$ ,  $\pi_{32} = 0.1$  (third panel) in group 2. The sample size was set equal to  $n_{i1} = n_{i2} = 25$ ,  $i = 1, 2, 3$  in the three strata.

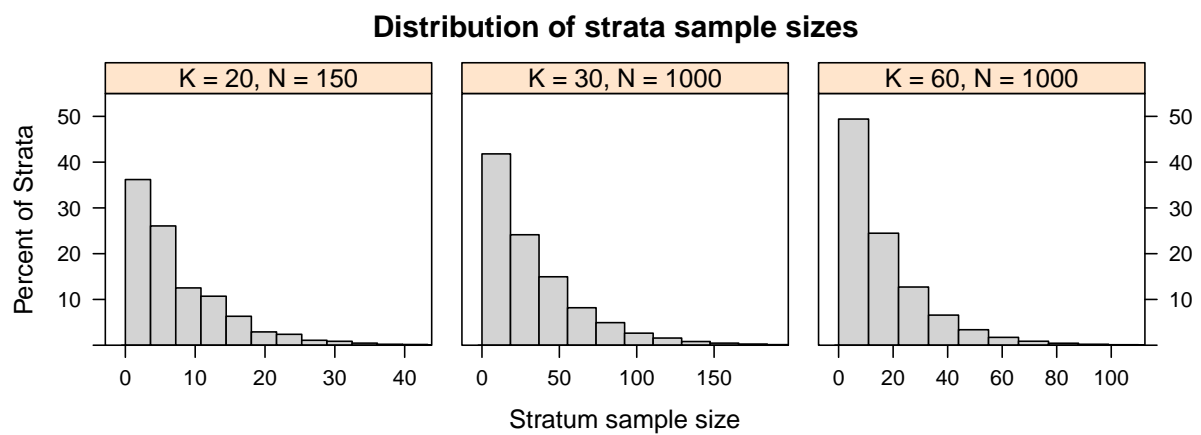


**Figure 3.** Coverage probability under various settings for the number  $K$  of strata and common risk difference  $\delta$  under a balanced sample size of  $n_{1i} = n_{2i} = 25$ . The interval estimators investigated (see Sections 1 and 2) are: Greenland-Robins (GR), Sato, the new approach proposed in Section 1 (“New”), the score interval for a linear combination of proportions (“Score”), and the score interval similar to the one proposed by Miettinen and Nurminen (“MN”).

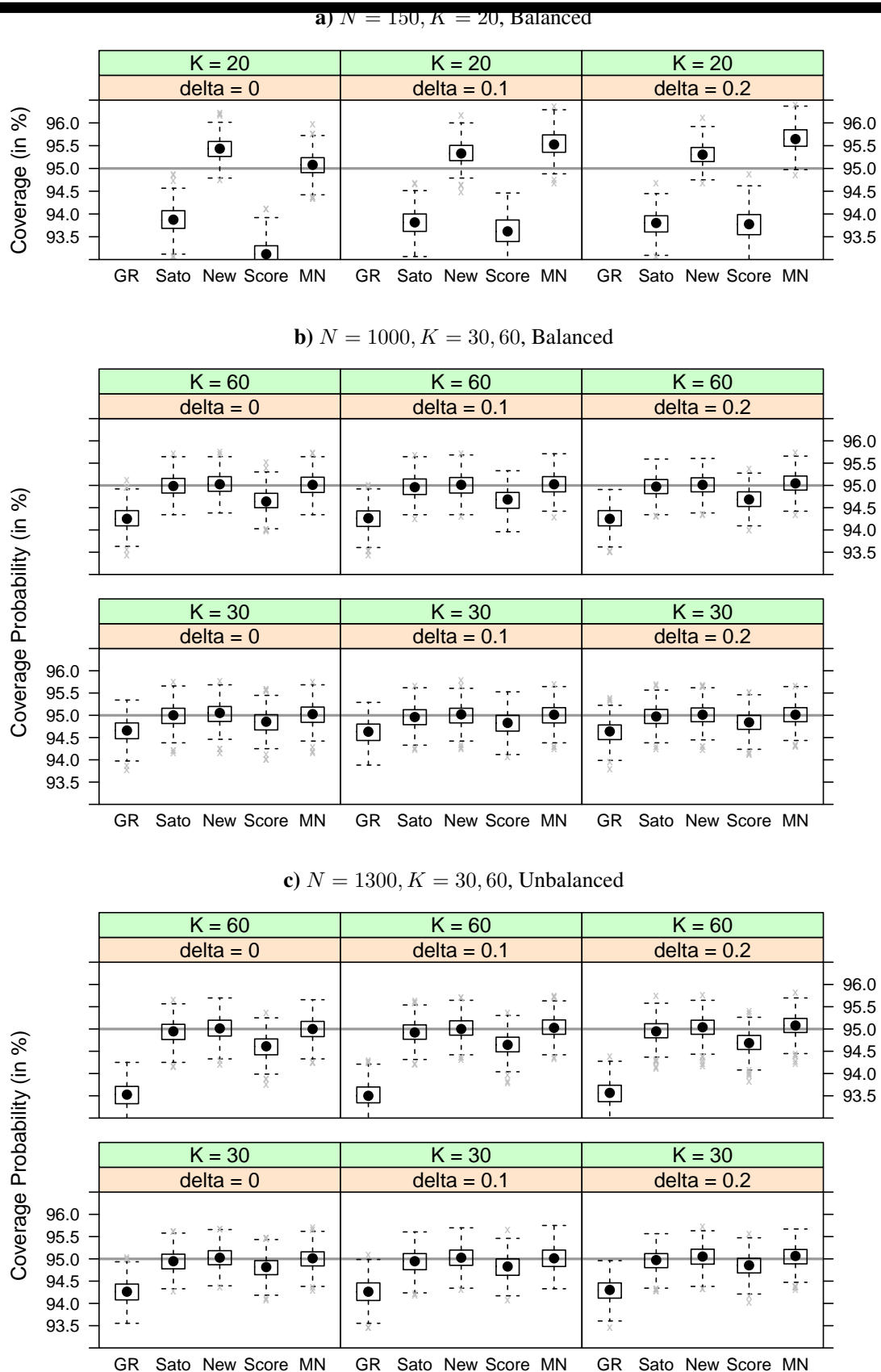




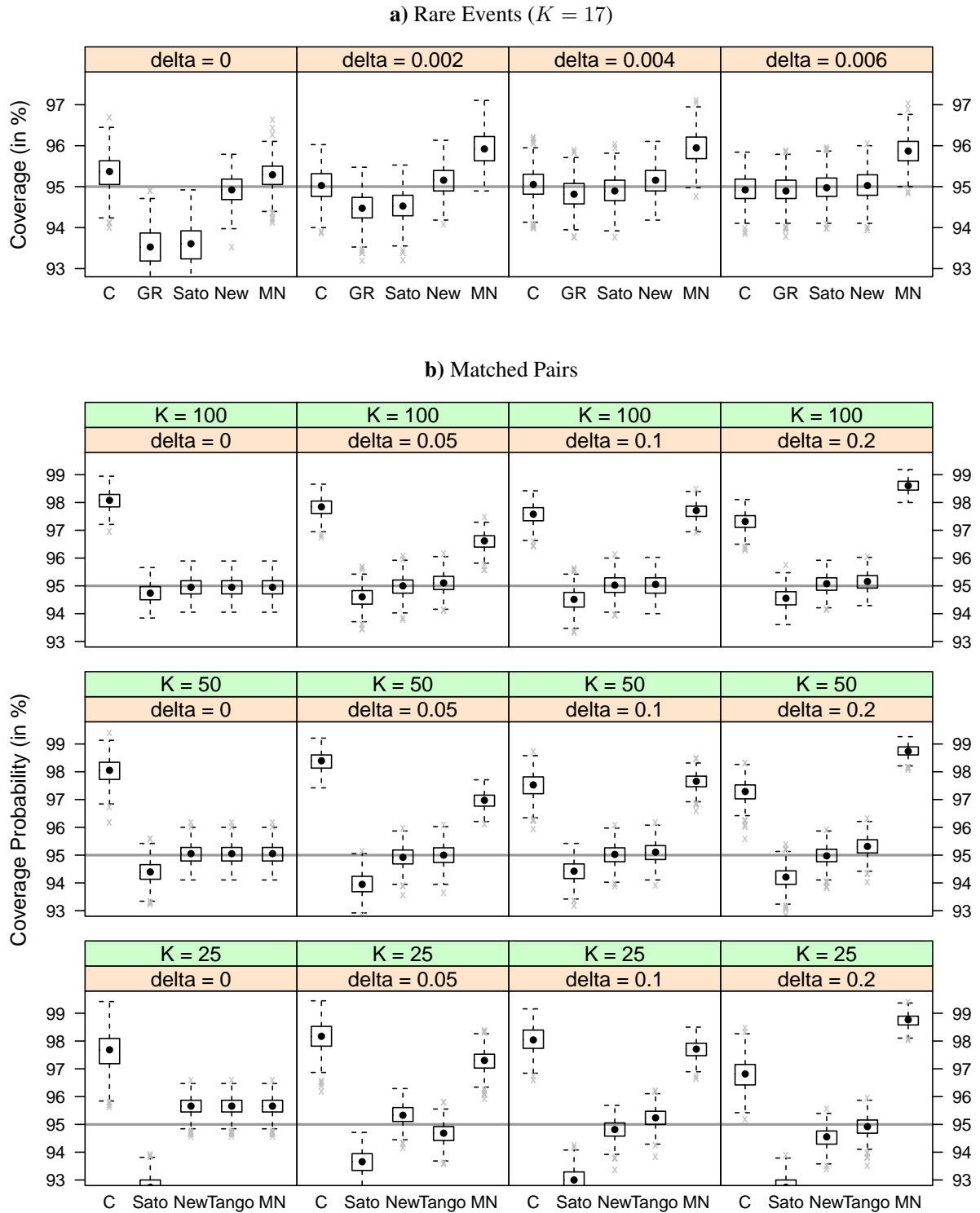
**Figure 4.** Coverage probability under various settings for the number  $K$  of strata and common risk difference  $\delta$  under an unbalanced sample size of  $n_{2i} = 15$ ,  $n_{1i} = 3n_{2i} = 45$ . The interval estimators are the same as in Figure 1.



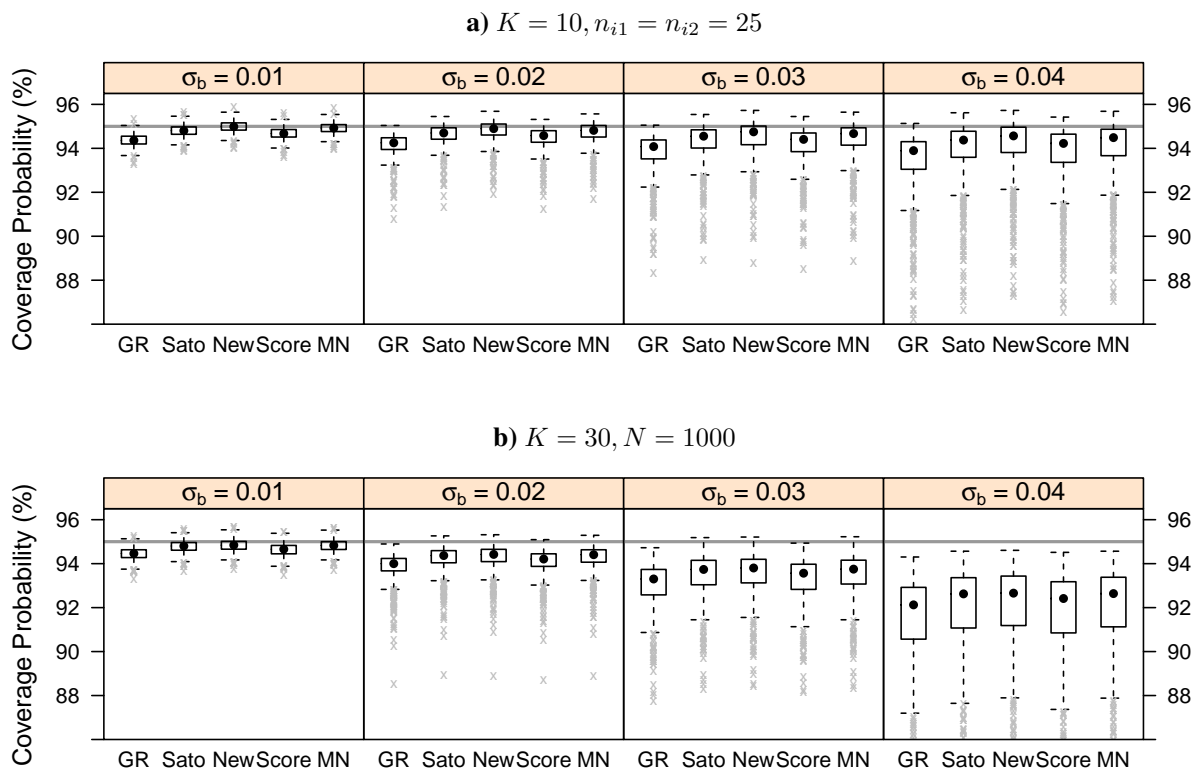
**Figure 5.** Distribution of strata sample sizes. The first panel shows the distribution when the number of strata  $K = 20$  and the total sample size in each group  $N = 150$  (sparse data situation), while the second and third panel show the distribution when  $K = 30$  or  $K = 60$  and  $N = 1000$ .



**Figure 6.** Coverage probability under strata with both large and small total sample sizes  $n_{i+}$  and common risk difference  $\delta$  under a) a very sparse setting with  $K = 20$  and a total of only  $N_1 = \sum_{i=1}^K n_{i1} = 150$  observations in each treatment group (balanced), b) balanced sample sizes  $n_{i1} = n_{i2}$  with  $N_1 = 1000$  and  $K = 30$  or  $60$  and c) unbalanced sample sizes in the two treatment groups ( $n_{i1} = 3n_{i2}$ ,  $N_1 = \sum_{i=1}^K n_{i1} = 1300$ ,  $K = 30$  or  $60$ ). The interval estimators are the same as in Figure 1 and 2.



**Figure 7.** Extreme Cases: Coverage probability under a) rare event setting with  $\pi_{i2} \sim \text{Uniform}[0, 0.0005], i = 1, \dots, K = 17$  and  $\delta$  ranging from 0 to 0.0006 and b) matched pairs with  $n_{i1} = n_{i2} = 1$  and  $\pi_{i2} \sim \text{Uniform}[0, 1 - \delta], i = 1, \dots, K$ . The interval estimators are the same as in previous Figures, while “C” denotes the procedure that collapses the tables into a single  $2 \times 2$  table and “Tango” stands for the Tango score interval for the difference of proportions in matched pairs.



**Figure 8.** Coverage probability under heterogeneity of the risk difference when  $\delta_i \sim N(\delta, \sigma_b)$  with  $\delta = 0.1$  and  $\sigma_b = 0.01, 0.02, 0.03$  or  $0.04$  for a) number of strata  $K = 10$  with balanced sample size  $n_{i1} = n_{i2} = 25$  in each stratum or b) number of strata  $K = 30$  with overall sample size  $N = \sum_{i=1}^K n_{i1} = \sum_{i=1}^K n_{i2} = 1000$  and varying strata sample sizes (see Figure 3). The interval estimators are the same as in Figures 1, 2 and 4.