Biometrics 64, 1–20
December 2008
DOI: 10.1111/j.1541-0420.2005.00454.x

Simultaneous Confidence Intervals for Ranks With Application to Ranking Institutions

Diaa Al Mohamad*, Jelle J. Goeman, and Erik W. van Zwet

Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.

**email:* diaa.almohamad@gmail.com

Summary: When a ranking of institutions such as medical centers or universities is based on a numerical measure of performance provided with a standard error, confidence intervals (CIs) should be calculated to assess the uncertainty of these ranks. We present a novel method based on Tukey's honest significant difference (HSD) test to construct simultaneous CIs for the true ranks. When all the true performances are equal, the probability of coverage of our method attains the nominal level. In case the true performance measures have no exact ties, our method is conservative. For this situation, we propose a rescaling method to the nominal level which results in shorter CIs while keeping control of the simultaneous coverage. We also show that a similar rescaling can be applied to correct a recently proposed Monte-Carlo based method which is anticonservative. After rescaling, the two methods perform very similarly. However, the rescaling of the Monte-Carlo based method is computationally much more demanding and becomes infeasible when the number of institutions is larger than 30 to 50. We discuss another recently proposed method similar to ours based on simultaneous CIs for the true performance. We show that our method provides uniformly shorter CIs for the same confidence level. We illustrate the superiority of our new methods with a data analysis for travel time to work in the U.S. and on rankings of 64 hospitals in the Netherlands.

Key words: League tables; Monte-Carlo; multiple comparisons; rankability; Tukey's HSD.

This paper has been submitted for consideration for publication in *Biometrics*

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/biom.13419

1. Introduction

Estimation of ranks is an important statistical problem which appears in many applications in healthcare, education and social services (Goldstein and Spiegelhalter, 1996). Institutions such as hospitals and universities are compared through league tables based on some numerical performance measure. Since such measures are accompanied by a standard error, the ranks are uncertain and CIs for the ranks are thus crucial (Marshall and Spiegelhalter, 1998; Goldstein and Spiegelhalter, 1996). Since ranking of an institution involves comparison with all the other institutions, the uncertainty related to ranks may be very high.

In applications, ranks are rarely accompanied with CIs and, if so, these are generally pointwise. Pointwise CIs are useful when we are interested in a single, a priori named, institution. However, if the institution of interest is chosen after seeing the data, simultaneity of CIs is crucial. Moreover, simultaneity is also necessary to quantify the uncertainty about which institutions are ranked best, second best, etc. In this paper, we present a method to produce simultaneous CIs with correct coverage of the true ranks at a prespecified joint level $1 - \alpha$.

The literature includes several methods for pointwise CIs for the ranks. We mention the parametric bootstrap method of Goldstein and Spiegelhalter (1996) which is widely used, see Marshall and Spiegelhalter (1998); Gerzoff and Williamson (2001); Feudtner et al. (2011) among others. It was pointed out, however, by Hall and Miller (2009) and Xie, Singh, and Zhang (2009) that the bootstrap pointwise CIs fail to cover the true ranks in the presence of ties or near ties among the compared institutions. Other methods were proposed based on testing pairwise differences between institutions, see Lemmers, Kremer, and Borm (2007); Holm (2012); Bie (2013). Lemmers et al. (2007) tested pairwise differences among Dutch hospitals by calculating Z-scores for their performance measures, but they did not correct for multiple testing and thus their CIs for ranks are pointwise and not simultaneous. Holm

(2012) and Bie (2013) also calculated a Z-score, but applied Holm's sequential algorithm to correct for multiple comparisons on the institution level, that is for each institution they correct for comparisons with other institutions.

In a recent report at the U.S. Bureau of the Census, Klein et al. (2018) showed how to construct simultaneous CIs for ranks based on CIs for the means. Zhang et al. (2014) introduced a method that produces simultaneous CIs for ranks using a Monte-Carlo approach. The method was used in recent papers such as Moss, Liu, and Zhu (2018). The method of Zhang et al. (2014) can be seen as a generalization of the method proposed by Goldstein and Spiegelhalter (1996) and can be considered as a parametric bootstrap method. However, it is not clear why this method should have a simultaneous coverage of at least $1-\alpha$. Since it depends on the method of Goldstein and Spiegelhalter (1996), we argue that it inherits that method's lack of correct coverage as shown by Hall and Miller (2009) and Xie et al. (2009). We show through extensive simulations that the method of Zhang et al. (2014) is indeed highly anticonservative when the performance measures are close to each other, and it attains the nominal level only when the performance measures are quite far from each other.

We present a novel method which uses Tukey's honest significant difference (HSD) test (Tukey, 1953). We show that Tukey's HSD can be used to produce simultaneous CIs for ranks with simultaneous coverage of at least $1-\alpha$ and exactly $1-\alpha$ if all true performances are equal. Our method bear similarities to the methodology presented by Klein et al. (2018). We show in paragraph 3 that our method is more powerful than the methods proposed in Klein et al. (2018).

Next, we focus our attention to an important and practical situation where we can assume that the institutions performance measures are all different. In this situation, we may have near ties but no exact ties in the true performance. In this situation our method becomes

conservative. We show that it is then possible to adjust the confidence level so that we reduce this conservativeness. We show similarly that it is possible to repair the method of Zhang et al. (2014) in order to regain control of the confidence level. After rescaling, the two methods produce similar results, but the rescaling method for Zhang et al. (2014) becomes computationally infeasible as the number of institutions exceeds 30 to 50.

The paper is organized as follows. In Section 2, we explain the context of this paper, the notations and the objective. In Section 3, we review Tukey's HSD and show that it can be used to provide simultaneous CIs for the ranks. In Section 3, we review the Monte-Carlo method of Zhang et al. (2014). In Section 4, we show how to rescale the confidence level of our method and the method of Zhang et al. (2014). Section 5 is devoted to simulation studies comparing our method with the method of Zhang et al. (2014) with and without rescaling the coverage. Finally, an example of travel time to work from Klein et al. (2018) and a new example on ranking Dutch hospitals are also used to compare our methods with the ones available from the literature. Software for the methods presented in this paper is available in R package `ICRanks`, downloadable from CRAN. The supplementary information provides further details, proofs and code.

2. Context and Objective

Let $\mu_1, \cdots, \mu_n$ be real valued numbers which represent the true performance of the n institutions we want to rank, for example the mortality rates of hospitals in our example in paragraph 5.3. For each institution i, we have an observed performance y_i, which is an estimator of μ_i. We assume that each institution's estimator is based on many independent subjects (e.g. students, patients) within the institution, so that it becomes reasonable to assume that y_i is normally distributed with known standard error σ_i. Our starting point, therefore, is a sample $y = (y_1, \cdots, y_n)$ of n independent observed performances, each drawn

from a Gaussian distribution

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{for } i \in \{1, \cdots, n\}, \tag{1}$$

where the standard deviations $\sigma_1, \cdots, \sigma_n$ are known but the means $\mu_1, \cdots, \mu_n$ are unknown.

Denote by $r_1, \cdots, r_n$ the true ranks of the means of institution $i = 1, \ldots, n$. These ranks are our target of inference, and our objective is to build simultaneous CIs for them. Let us first define the ranks $r_1, \cdots, r_n$, allowing for the possibility of ties.

DEFINITION 1 (set-ranks): We define the lower-rank l_i and the upper-rank u_i of mean μ_i by

$$l_i = 1 + \sum_{j \neq i} \mathbb{1}_{\mu_j < \mu_i}, \quad u_i = n - \sum_{j \neq i} \mathbb{1}_{\mu_j \geqslant \mu_i}.$$

We finally define the set-rank of μ_i as the set of natural numbers $r_i = \{l_i, l_i + 1, \cdots, u_i\}$ denoted here $[l_i, u_i]$.

When there are ties between the means, we suppose that each of the tied means possesses a set of ranks $r_i = [l_i, u_i]$. For example, suppose that we only have 3 means μ_1, μ_2 and μ_3 such that $\mu_1 = \mu_2 < \mu_3$. Then, the set-rank of μ_1 is $[1, 2]$ which includes both ranks 1 and 2, and the set-rank of μ_2 is also $[1, 2]$, whereas the rank of μ_3 is $[3, 3]$ which is simply rank 3. The rationale of the definition of the set-ranks is that in case of ties, the ranking is arbitrary, and a small perturbation of the true performance may produce any rank in the set of ranks. We call the ranks induced from the observed sample y the empirical ranks. These ranks might be different from the true ranks of the means, and since the sample is assumed to have a continuous distribution, the empirical ranks are all singletons.

In the second part of this paper we will avoid the set-ranks by assuming the following.

Assumption 1 The means have no ties.

Under this assumption, we get $r_i = l_i = u_i$. Thus, the set-ranks coincide with the usual

ranking definition; the ranks are calculated for each mean by counting down how many means are below it.

We aim on the basis of the sample y to construct simultaneous CIs for the set-ranks of the means. In other words, for each i we search for a $[L_i, U_i]$ such that:

$$\mathbb{P}\left([l_i, u_i] \subseteq [L_i, U_i], \forall i \in \{1, \cdots, n\}\right) \geqslant 1 - \alpha \tag{2}$$

for a pre-specified confidence level $1 - \alpha$. It is worth noting that the CIs here are sets in $\mathbb{N}$, the set of natural numbers.

Two different types of statement can be obtained from the simultaneous CIs (2). First, for each institution the possible ranks that it might take (which is our main objective). Second, for each rank (first best, second best, etc.) the list of institutions that might attain this specific rank. In other words, we have confidence sets for the best mean(s), second best mean(s), etc. These confidence sets have also a joint confidence level of at least $1-\alpha$. Indeed, in order to find the means that can be the best, it suffices to check for the means whose rank CI starts at 1. In the same way, we can look at the means whose rank CI includes rank 2 to obtain a confidence set of the means ranked second best and so on.

3. Simultaneous CIs for Ranks: The General Case

We assume here that the means $\mu_1, \cdots, \mu_n$ might have ties. We start with our main new method using Tukey's HSD and show the link with the work of Klein et al. (2018). Then, we compare with the method of Zhang et al. (2014).

Tukey's pairwise comparison procedure (Tukey, 1953) is an easy way to compare means of observations with (assumed) Gaussian distributions especially in ANOVA models. The procedure is usually used for providing simultaneous confidence statements about the differences between the means and controls the FWER at level α. We show how it can be used to construct simultaneous confidence intervals for ranks.

We consider the general case with possibly unequal σ_i's. Tukey's HSD tests all null hypotheses $H_{i,j} : \mu_i - \mu_j = 0$ at level α using the rejection region

$$\left\{ \frac{|y_i - y_j|}{\sqrt{\sigma_i^2 + \sigma_j^2}} > q_{1-\alpha} \right\} \tag{3}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the distribution of the Studentized range

$$\max_{i,j=1,\cdots,n} \frac{|\tilde{Y}_i - \tilde{Y}_j|}{\sqrt{\sigma_i^2 + \sigma_j^2}}, \tag{4}$$

and $\tilde{Y}_1, \cdots, \tilde{Y}_n$ are independent and $\tilde{Y}_i \sim \mathcal{N}(0, \sigma_i^2)$.

The following theorem states the main result of our paper. We show how to construct simultaneous CIs for the ranks using Tukey's HSD, allowing the possibility of ties.

THEOREM 1: *Let*

$$\begin{aligned} L_i =& 1 + \#\left\{ j : \; y_i - y_j - \sqrt{\sigma_i^2 + \sigma_j^2} q_{1-\alpha} > 0 \right\} \\ U_i =& n - \#\left\{ j : \; y_i - y_j + \sqrt{\sigma_i^2 + \sigma_j^2} q_{1-\alpha} < 0 \right\}. \end{aligned}$$

The intervals $[L_i, U_i]$ *for* $i = 1, \cdots, n$ *are* $(1-\alpha)$*-joint CIs for the ranks of means* $\mu_1, \cdots, \mu_n$.

Klein et al. (2018) also proposed to construct simultaneous CIs for ranks using Theorem 1 but started from simultaneous CIs for the means and not the pairwise differences. It is more efficient to look at pairwise differences directly as we do in Theorem 1. The optimality of Tukey's HSD (Hochberg and Tamhane, 1987, p. 81) allows our procedure to be more powerful. We prove this for the equal σ case in the following proposition.

PROPOSITION 3.1: Let $[y_i - \sigma z_\alpha, y_i + \sigma z_\alpha]$ be simultaneous CIs for μ_i for $i = 1, \cdots, n$ with joint confidence level $1 - \alpha$, then the simultaneous CIs for the ranks using the method of Klein et al. (2018) based on those intervals are uniformly longer than the CIs $[L_i, U_i]$ defined in Theorem 1.

When the standard deviations are not equal, we do not have a proof of the analogous result. Still, Tukey's HSD is a preferred procedure for simultaneous CIs for the differences

and still has some optimality properties when the standard deviations are not equal, see for example Hochberg and Tamhane (1987, p. 81) and Rafter, Abell, and Braselton (2002). We will see in Section 5.2 how our Tukey-based CIs for ranks are uniformly shorter than those obtained using the method of Klein et al. (2018).

Theorem 1 states that the Tukey-based method controls the $\alpha-$level, but allows that it may be conservative in general. We show in the following result that when all the means are equal, it is exact.

PROPOSITION 3.2: Under the full null, that is when $\mu_1 = \cdots = \mu_n$, the simultaneous coverage of the CIs $[L_i, U_i]$ for $i = 1 \cdots, n$ produced by Tukey's HSD is exactly $1 - \alpha$.

This result demonstrates that, if it cannot be excluded that the true means are equal, there is no room for improvement of the procedure by increasing the α-level. In other scenarios, such as under Assumption 1, we may improve upon the resulting confidence intervals by rescaling α, as we shall see in the next section.

Zhang et al. (2014) were the first to discuss and propose a method for simultaneous CIs for ranks. While our method based on Tukey's HSD has a proven simultaneous coverage proven in Theorem 1, the method of Zhang et al. (2014) fails to achieve nominal coverage. This is mainly because the method of Zhang et al. (2014) is based on the method of Marshall and Spiegelhalter (1998), see Web Appendix A. In the settings where there are ties or near ties (Xie et al., 2009; Hall and Miller, 2009) showed that confidence intervals for ranks based on bootstrap, such as the method of Marshall and Spiegelhalter (1998), have coverage less than the nominal level. In the next sections, we will investigate this in more detail. We will show that the simultaneous coverage of the resulting CIs reaches the nominal level $1-\alpha$ only when the differences among the means are large enough. The true coverage depends not only on the range of values of the means, but also on n and on the way the means are dispersed in their range. Otherwise, the method is anti-conservative and the simultaneous coverage could

drop to very low levels. We see this phenomenon in simulations (see Section 5) both with and without Assumption 1. To remedy this problem, we will propose a method to readjust the simultaneous coverage in paragraph 4.

4. Simultaneous CIs for Ranks When Ties Are not Allowed

It might be reasonable in the context of ranking to assume Assumption 1, hence the means $\mu_1, \cdots, \mu_n$ are all different, and that there are in reality no ties. Under Assumption 1, the intervals (set-ranks) that should be covered include only one element, so covering them is relatively easy for Tukey's method. We show in this section that power may be gained by adjusting the $\alpha-$level of that method. We first start by treating the case when the standard deviations are the same, that is $\sigma_1 = \cdots = \sigma_n = \sigma$. We move then to the general case of different standard deviations. Finally, we argue that a similar approach may be used to repair the anticonservativeness of the method of Zhang et al. (2014).

When there are no ties, the configuration of Proposition 3.2 is excluded, and the Tukey-based approach becomes more conservative. Moreover, as the differences among the means become greater, we found empirically that the joint coverage probability increases. We illustrate this in the left part of Figure (1) by considering vectors of means of the form $\varepsilon\mu$ where $\varepsilon \in (-1, 1)$ and $\mu = (1, \cdots, 10)^t$ with dimension 10. The common standard deviation is set to 1. When $\varepsilon = 0$, we assume arbitrary ranks for the means, say $1, \cdots, 10$, in order to conform with the assumption that there are no ties. In other words, the case $\varepsilon = 0$ corresponds to a situation where the means are unequal but arbitrary small, say of order 10^{-100}, that we cannot distinguish from 0 by machine precision.

In Figure (1), the coverage probability of the Tukey-based method reaches a minimum when $\varepsilon = 0$. The worst case which corresponds to the minimum simultaneous coverage at level $1 - \alpha$ happens when all the means are arbitrarily small while not having ties. The

following fact was verified empirically. For any $\mu \neq 0$

$$\mathbb{P}_{\mu}\left(\forall i, r_i \in [L_i, U_i]\right) \geqslant \mathbb{P}_{\mu=0}\left(\forall i, i \in [L_i, U_i]\right). \tag{5}$$

Note that when $\mu = 0$, we attribute the ranks $1, \cdots, n$ to the coordinates of μ. This worst case configuration is known in hypothesis testing for example when we test if the vector of means has an ascending order (Robertson and Wegman, 1978). The type I error is then highest when all the means are equal. In the case of the Kramer-Tukey procedure (Tukey, 1953; Kramer, 1956), Hayter (1984) showed that it is conservative and has a worst case configuration when all the standard deviations are the same. In these procedures, the worst case configuration corresponds to a type I error exactly equal to α. In the Tukey-based method, the worst case configuration corresponds to a type I error equal to α based on Proposition 3.2, but much less than α under Assumption 1. The gap between the true and the nominal level will be exploited in order to gain more power.

[Figure 1 about here.]

We propose to rescale our method so that the Tukey-based method delivers a simultaneous coverage of at least $1-\alpha$ but in a less conservative way (a scaling down). Due to (5), the problem reduces to rescaling the worst case. We look for $\tilde{\alpha} \in (\alpha, 1)$ such that

$$\mathbb{P}_{\mu=0}\left(\forall i, i \in [L_i(\tilde{\alpha}), U_i(\tilde{\alpha})]\right) = 1-\alpha. \tag{6}$$

We have now, for any $\mu \neq 0$,

$$\mathbb{P}_{\mu}\left(\forall i, r_i \in [L_i(\tilde{\alpha}), U_i(\tilde{\alpha})]\right) \geqslant \mathbb{P}_{\mu=0}\left(\forall i, i \in [L_i(\tilde{\alpha}), U_i(\tilde{\alpha})]\right) = 1-\alpha.$$

If we do so, the simultaneous coverage of the Tukey-based method will be equal to the nominal level $1-\alpha$ near zero and higher than $1-\alpha$ elsewhere as illustrated in the right part of Figure (1). Since the coverage probability increases as α decreases, then using $\tilde{\alpha}$ the joint coverage at any μ is lower than the joint coverage using α. Moreover, equation (6) has a unique solution in the interval $(\alpha, 1)$.

Solving equation (6) can be performed using any mathematical program, for example using function `uniroot` available in the statistical program R. Table (1) shows the rescaled significance level $\tilde{\alpha}$ necessary to reach an actual coverage of $80\%, 90\%$ and 95% when the number of means increases from 10 to 100. Although the rescaled significance level moves towards 1 for the Tukey-based method, the resulting CIs retain simultaneous coverage of at least $1-\alpha$. Therefore, they will be shorter than the ones we obtain using α.

[Table 1 about here.]

Note that as n increases, the rescaled level increases because the actual simultaneous coverage increases as illustrated through simulations in Web Appendix D. When the standard deviations are not the same, the worst case configuration still happens when the means are arbitrarily close to each other, but the order of the σ's has an influence on it. We propose to reorder the standard deviations in the following manner

$$\sigma_1 \leqslant \cdots \leqslant \sigma_{n/2}, \sigma_{n/2} \geqslant \cdots \geqslant \sigma_n, \tag{7}$$

The idea is that the middle (empirically) ranked means tend to have large CIs for their ranks whereas the lowest and highest (empirically) ranked means tend to have small CIs. In the Web Appendix A, we show through simulations that our chosen configuration, while not the worst case, is indeed very close to the worst case.

We observed a similar empirical result to (5) for the method of Zhang et al. (2014), see Figure (1). However, the problem with this method is that it is, in contrast to the Tukey-based method, very anticonservative. Therefore, we will readjust the significance level in order to regain control of the joint confidence level at $1-\alpha$. Therefore, let $[L_1^Z, U_1^Z], \cdots, [L_n^Z, U_n^Z]$ be the simultaneous CIs produced by the method of Zhang et al. (2014) at level $1-\alpha$. We need to find an $\tilde{\alpha}$ such that

$$\mathbb{P}_{\mu=0}\left(\forall i, i \in [L_i^Z(\tilde{\alpha}), U_i^Z(\tilde{\alpha})]\right) - (1-\alpha) = 0,$$

which can also be done using function `uniroot` from R. We then get the rescaled coverage

in the right part of figure (1) and the corresponding $\tilde{\alpha}$ in table (1). The table shows that in order to use the method of Zhang et al. (2014) and make sure not to be anticonservative, we need to use very small values of the significance level. However, as $\tilde{\alpha}$ becomes smaller we need to increase the number of Monte-Carlo samples K required to estimate the joint distribution of the ranks as mentioned by Zhang et al. (2014). For example, when $n = 50$ we need at least $K = 10^6$ $n-$samples, and thus rescaling the method of Zhang et al. (2014) becomes quickly infeasible for higher number of means so that the resulting CIs are not ensured to have the desired coverage of $1 - \alpha$.

5. Simulation Study and Real Data Analysis

In this section we provide several examples (real and simulated) to demonstrate the CIs produced using our approaches from Sections 3 and 4. We also compare the coverage and the efficiency of the CIs produced by the method proposed by Zhang et al. (2014) and Klein et al. (2018) to the ones obtained by our method in different scenarios. The efficiency is calculated as $1 - \hat{R}_n$ where $\hat{R}_n$ is called the (estimated) rankability which is equal to the average lengths of the CIs

$$\hat{R}_n(\alpha) = 1 - \frac{1}{n(n-1)} \sum_{i=1}^{n} (U_i - L_i). \tag{8}$$

Web Appendix C provides further discussion on this measure and the idea behind it. In Web Appendix A, we also provide further simulations that shows that even if the distribution of the data is not Gaussian, our methods are still robust.

The analysis is done using the statistical program R, and the code of the functions is available in the R package `ICRanks` which can be downloaded from the CRAN repository. The R code for the method of Zhang et al. (2014) is provided in the supplementary materials (see also Web Appendix E). We note that the running time for our approaches never exceeded 3 seconds using a standard laptop.

5.1 *The case of a common standard deviation*

The simulation setup is the following. We aim to estimate the average simultaneous coverage of the Monte-Carlo method of Zhang et al. (2014) and the Tukey-based method. To do so, we generate the means μ_i's independently from the Gaussian distribution $\mathcal{N}(0, \tau^2)$ for $\tau = 0.5, 1, 2$, so that Assumption 1 holds with probability 1. For each value of τ, we generate 1000 $n-$samples of means $\mu = (\mu_1, \cdots, \mu_n)$ for $n = 10, 30$ and 50. Then, a Gaussian vector y is generated from the multivariate Gaussian distribution $\mathcal{N}(\mu, I_n)$. The simultaneous coverage based on these samples is estimated. The rescaled values of α for both methods have already been calculated in Table (1). We provide a table of the estimated coverage before and after rescaling the significance level so that the actual coverage at the worst case becomes $1 - \alpha$ for $\alpha = 0.1$. We calculate also the average $1 - \hat{R}_n(\alpha)$ where $\hat{R}_n(\alpha)$ is the rankability measure (8). The results are in Table 2.

[Table 2 about here.]

We conclude from the table the following points.

(1) The method of Zhang et al. (2014) clearly provides shorter CIs for ranks. However, this comes at the cost of an unacceptably low simultaneous coverage.

(2) The simultaneous coverage of the method of Zhang et al. (2014) increases as the range of means increases at a fixed n. On the other hand, it decreases as n increases when the range of the means is held fixed.

(3) In average, the Tukey-based method seems to produce shorter CIs than the method of Zhang et al. (2014) when they are both rescaled, but this difference is not statistically significant.

(4) Reducing the conservativeness of the Tukey-based method is always possible since the rescaled α is in the interval $(\alpha, 1)$.

(5) Repairing the anticonservativeness of the method of Zhang et al. (2014) is only possible in practice for $n \leqslant 50$ since the rescaled α becomes too close to 0 in that case.

5.2 *Travel time to work case study: A case of different standard deviations*

We use a dataset collected by the American Community Survey for the average time to travel to work in 51 states in the U.S. The data is available from Klein et al. (2018) and they apply their new method to obtain simultaneous CIs for the ranks of the 51 states. We use it to illustrate our method based on Tukey's HSD and to compare it with the method proposed in Klein et al. (2018) using the Sidàk correction $1-(1-\alpha)^{1/n}$ that gives the best result in their report. The full comparison is in the Web Appendix D, and we only cite the results of four states. The CIs obtained by our Tukey-based procedure are shorter than the ones obtained by the method of Klein et al. (2018) for 28 states. The ones obtained by the rescaled Tukey procedure are better for 45 states.

5.3 *Ranking hospitals in the Netherlands: A case of different standard deviations*

We studied a dataset for Dutch hospitals concerning abdominal aneurysms surgery. The study included 9489 patients operated at 64 hospitals in the Netherlands at dates mostly between the years 2012 and 2016. The number of patients per hospital ranged from 3 to 358 with an average of 150 patients per hospital. The dataset included the following variables

- the hospital ID where the patient was treated;
- the date of surgery;
- the context of surgery: Elective, Urgent, Emergency;
- the surgical procedure: "Endovascular", "Endovascular converted" and "Open". "Endovascular" means the patient had a minimal invasive procedure through the femoral artery in the groin. "Endovascular converted" means the surgeons first tried a minimal invasive

procedure through the femoral artery in the groin, but then realized they had to do an open surgery;

- a complication within 30 days (yes or no);
- the mortality within 30 days (yes or no);
- VpPOSSUM: a numerical score that summarizes the pre-operative state of the patient.

In order to conform to the normality assumption in our model, we excluded hospitals with small number of patients. The hospital effect is then estimated based on enough samples to assume an approximate Gaussian distribution. This left us with 61 hospitals and each one of them had at least 54 patients. We compared these hospitals according to two kinds of measures, a performance measure which is the complication within 30 days and a process measure which is the surgical procedure. Differences among hospitals based on the complication within 30 days are not usually expected to be huge. However, hospitals may be quite different in the choice of the type of surgery because the "Endovascular" type is rather new to surgeons so that the choice of surgery will depend on what the surgeon is most practiced at. Therefore, we can expect to obtain huge differences among the hospitals.

We corrected for case-mix effect with a fixed effect logistic regression model using the `VpPOSSUM` variable and with fixed effects for the hospitals. To make a fair comparison between the hospitals, it is important to adjust for the pre-operative state of the patients, that is the so-called "case mix". We use the clinical measure V(p)-POSSUM which is specifically developed for patients undergoing abdominal aortic aneurysm surgery, see Prytherch et al. (2001). We do stress that our ranking serves as an illustration of the method, and in any real application one should study the case-mix adjustment in more detail. Here, we model the probability $p_{i,j}$ that patient j from hospital i gets an open surgery by

$$\log\left(\frac{p_{i,j}}{1-p_{i,j}}\right) = \alpha_i \mathrm{id}_i + \beta_{i,j}\mathrm{VpPOSSUM}_{i,j}.$$

The measures are then the estimates of the hospital fixed effect α_i accompanied by its standard error. The resulting scores based on the complication seems to have a few differences especially between the two extremities; see the left part of Figure (2). We also fit a random effect mixed-model to the hospital effect using function `rma` from package `metafor` (Viechtbauer, 2010) and estimated the variance of the random effects using the Sidiki-Jonkman method and tested for heterogeneity among the hospitals. We found a p-value of 0.09 which was inline with the result of the forest plot that the overall differences do not seem substantial.

By correcting for case-mix effect similarly to the complication we obtained the right part of Figure (2). One of the hospitals had no patients operated with an open surgery, therefore, we added to all the hospitals a row of data with a virtual patient who had an open surgery and with a value of `VpPOSSUM` equal to the average in the corresponding hospital. The resulting scores showed clearly more differences than the ones obtained using the complication with smaller standard deviation. When we tested for heterogeneity among the hospitals, we got a p-value lower than 0.001 which also supports the large differences seen in the forest plot (Figure (2)).

We calculated simultaneous CIs for the ranks of these hospitals. We started with the complication. We calculated the rescaled significance level at the worst case configuration, that is we considered a null vector of means and the vector of standard deviations, obtained from the data, ordered according to the worst configuration we found in paragraph 4, namely configuration (7). We used the first 10 hospitals, 30 hospitals and finally all the hospitals and looked at how the rescaled significance level changed.

In order to apply the Monte-Carlo method of Zhang et al. (2014) and make sure that the confidence level was at least 90%, we needed to use a significance level of 0.0003 when taking the first 10 hospitals. We then needed a rescaled significance level below 5×10^{-6} for

more than 30 hospitals. Thus, it was not possible to rescale the simultaneous coverage of the method of Zhang et al. (2014) on the full dataset to 90%. Therefore, we only show the results for the Tukey-based method. For the latter, the rescaled significance level is 0.583 on the full dataset.

The simultaneous CIs for the ranks of the hospitals based on the complication at the joint level 90% are illustrated in the Web Appendix D. The CIs cover almost the whole range of ranks, and there are barely any differences among the hospitals according to the complication. The rankability is 0.098 for the Tukey-based method without rescaling, and is 0.186 after rescaling. Simultaneous CIs for the ranks of the hospitals based on the type of surgery at joint level 90% are illustrated in Figure (3) with a rankability of 0.240 for Tukey's HSD. Rescaling the significance level clearly improves the results of the Tukey-based CIs. The rescaled significance level is $\tilde{\alpha} = 0.607$. The new rankability is 0.358. Here again, we could not apply the method of Zhang et al. (2014) because the number of hospitals was too large. The simultaneous CIs for the ranks using the method of Klein et al. (2018) are larger than the ones obtained using our Tukey-based procedures. The full list of resulting CIs are provided in the Web Appendix D.

[Figure 2 about here.]

[Figure 3 about here.]

6. Discussion

We presented a novel method to produce simultaneous CIs for ranks based on Tukey's HSD and proposed a practical improvement under Assumption 1 that there are no ties among the means. We showed through simulations that the simultaneous confidence level of the method of Zhang et al. (2014) goes below the nominal level unless the means are very far from each other. We proposed a solution to fix this problem by rescaling the confidence

level. Surprisingly, after rescaling, the results of the Tukey-based method and the method of Zhang et al. (2014) seem to be almost the same and the differences are not significant. We also compared our method to a similar method proposed by Klein et al. (2018) and showed that our method provides uniformly shorter CIs for ranks.

By providing valid methods for simultaneous CIs for ranks, practitioners may look at all the institutions together instead of only looking at a specific one, and they obtain valid CIs chosen on the basis of the data such as the one with the worst empirical rank. Simultaneity also provides a way to state which of the institutions could (or could not) be ranked first best, second best, and so one.

The data analysis of two real data sets show that although the league tables tend to show a ranking, this ranking is not really reflected in the CIs for the ranks showing again that the rank of an institution is one of the most difficult quantities to estimate (Spiegelhalter, 2005).

Our methods are shown in Web Appendix A to be robust against deviations from the normality assumption. Deviations from the independence assumption is not studied here, but further research could benefit from the paper of Seco et al. (2001) that studies this in Tukey's HSD. Another limitation of our approach is that the proof of the rescaled methods is based only on simulations. Therefore, a more solid proof is needed.

For a different objective, using Dunnet's test, it is possible to look for the rank of only one prespecified institution that we are interested in, see also Finner and Strassburger (2007).

Acknowledgements

This research is funded by the NWO VIDI grant 639.072.412.

Data Availability

The data that support the findings of paragraph 5.2 in this paper are openly available in Klein et al. (2018) through the link `https://www.census.gov/content/dam/Census/library/working-pap`

The data consist of the average travel time to work and its standard error for the states in the United States of America. The data that support paragraph 5.3 consist of information about the patients who had abdominal aneurysms surgery in Dutch hospitals. This data is confidential and cannot be shared.

References

Bie, T. (2013). Confidence intervals for ranks: Theory and applications in binomial data. Master's thesis, Uppsala University, Sweden. Master thesis under the supervision of R. Larsson.

Feudtner, C., Berry, J. G., Parry, G., Hain, P., Morse, R. B., Slonim, A. D., Shah, S. S., and Hall, M. (2011). Statistical uncertainty of mortality rates and rankings for children's hospitals. *Pediatrics* **128,** e966–e972.

Finner, H. and Strassburger, K. (2007). Step-up related simultaneous confidence intervals for mcc and mcb. *Biometrical Journal* **49,** 40–51.

Gerzoff, R. B. and Williamson, G. D. (2001). Who's number one? the impact of variability on rankings based on public health indicators. *Public Health Reports (1974-)* **116,** 158–164.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **159,** 385–443.

Hall, P. and Miller, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Statist.* **37,** 3929–3959.

Hayter, A. J. (1984). A proof of the conjecture that the tukey-kramer multiple comparisons procedure is conservative. *Ann. Statist.* **12,** 61–75.

Hochberg, Y. and Tamhane, A. (1987). *Multiple comparison procedures.* Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Holm, S. (2012). Confidence intervals for ranks. *Department of Mathematical Statistics* Unpublished manucript.

Klein, M., Wright, T., and Wieczorek, J. (2018). A simple joint confidence region for a ranking of k populations. *Research Report Series, U.S. Bureau of the Census* pages 1–18.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* **12,** 307–310.

Lemmers, O., Kremer, J. A., and Borm, G. F. (2007). Incorporating natural variation into IVF clinic league tables. *Human Reproduction* **22,** 1359–1362.

Marshall, E. C. and Spiegelhalter, D. J. (1998). Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ : British Medical Journal* **316,** 1701–1705.

Moss, J. L., Liu, B., and Zhu, L. (2018). State prevalence and ranks of adolescent substance use. *Preventing chronic disease* **15,**.

Prytherch, D., Sutton, G., and Boyle, J. (2001). Portsmouth possum models for abdominal aortic aneurysm surgery. *Br. J. Surg.* **88,** 958–963.

Rafter, J. A., Abell, M. L., and Braselton, J. P. (2002). Multiple comparison methods for means. *SIAM Review* **44,** 259–278.

Robertson, T. and Wegman, E. J. (1978). Likelihood ratio tests for order restrictions in exponential families. *Ann. Statist.* **6,** 485–505.

Seco, G. V., Menéndez de la Fuente, I. A., and Escudero, J. R. (2001). Pairwise multiple comparisons under violation of the independence assumption. *Quality and Quantity* **35,** 61–76.

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine* **24,** 1185–1202.

Tukey, J. W. (1953). The problem of multiple comparisons. *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948 - 1983* pages 1–300. Unpublished manuscript.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36,** 1–48.

Xie, M., Singh, K., and Zhang, C.-H. (2009). Confidence intervals for population ranks in the presence of ties and near ties. *Journal of the American Statistical Association* **104,** 775–788.

Zhang, S., Luo, J., Zhu, L., Stinchcomb, D. G., Campbell, D., Carter, G., Gilkeson, S., and Feuer, E. J. (2014). Confidence intervals for ranks of age-adjusted rates across states or counties. *Statistics in Medicine* **33,** 1853–1866.

Supporting Information

Web Appendices A-E referenced in Sections 1, 3-5 are available with this paper at the Biometrics website on Wiley Online Library. The code of our proposed methods are available from the R package `ICRanks` available on CRAN on the following link `https://cran.r-project.org/web/packages/ICRanks/`. More illustrative code of using the discussed methods in the paper including ours are available at the Biometrics website on Wiley Online Library.

Received October 2019. *Revised February* 2019. *Accepted* .

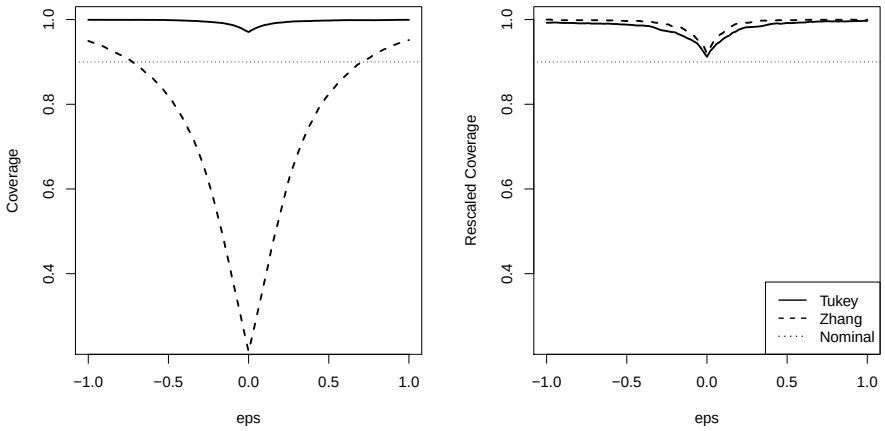

Figure 1. The simultaneous coverage of the Tukey-based method at vectors of true means of the form $\varepsilon\mu$ with $\varepsilon \in (-1, 1)$. The nominal level is $1-\alpha = 0.9$. The left figure corresponds to the actual coverage $\beta_\mu(\alpha)$ at joint confidence level $1-\alpha$ whereas the right one corresponds to the actual coverage $\beta_\mu(\tilde{\alpha})$ after rescaling the worst case to the nominal level. The coverage curve for the method of Zhang et al. (2014) is also illustrated before and after rescaling (paragraph 4).

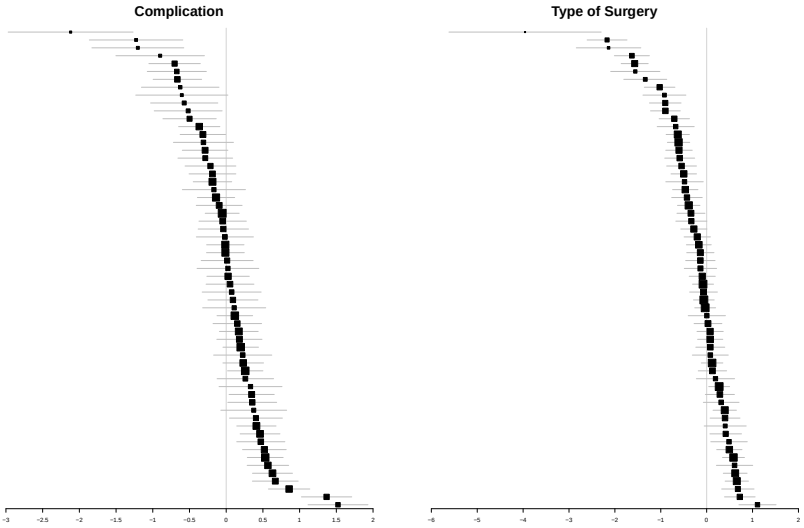

Figure 2. Forest plots for the hospitals effect after case-mix correction based on the complication or the surgical procedure.

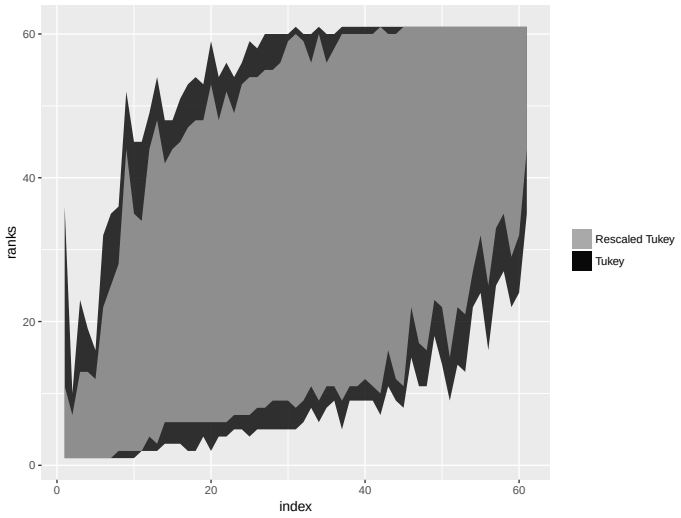

Figure 3. Simultaneous CIs for the ranks of 61 hospitals in the Netherlands. Data is corrected for case-mix effect.

Table 1
Values of $\tilde{\alpha}$ necessary to rescale the coverage at the worst case back to $1-\alpha$.

	Rescaled coverage					
	95%		90%		80%	
	Tukey	Zhang	Tukey	Zhang	Tukey	Zhang
$n = 10$	0.158	6.5×10^{-4}	0.285	0.0015	0.467	0.006
$n = 30$	0.303	9.8×10^{-6}	0.491	4.6×10^{-5}	0.693	4×10^{-5}
$n = 50$	0.418	$< 5 \times 10^{-6}$	0.574	7×10^{-6}	0.778	3.1×10^{-5}
$n = 100$	0.545	$< 5 \times 10^{-6}$	0.725	$< 5 \times 10^{-6}$	0.893	5×10^{-6}

Table 2
Coverage probability and efficiency when $\tau \in \{0.5, 1, 2\}$ and $\alpha = 0.1$ before and after rescaling the worst case.

	Coverage				$1 - \hat{R}_n(\alpha)$			
	not rescaled		rescaled		not rescaled		rescaled	
	Tukey	Zhang	Tukey	Zhang	Tukey	Zhang	Tukey	Zhang
				$\tau = 0.5$				
$n = 10$	0.998	0.468	0.961	0.976	0.990	0.789	0.971	0.977
$n = 30$	1.000	0.027	0.978	0.987	0.998	0.740	0.990	0.991
$n = 50$	0.997	0.000	0.976	0.984	0.999	0.726	0.994	0.995
				$\tau = 1$				
$n = 10$	0.996	0.603	0.972	0.994	0.959	0.698	0.916	0.935
$n = 30$	1.000	0.088	0.993	0.996	0.987	0.651	0.957	0.967
$n = 50$	0.999	0.016	0.996	0.998	0.992	0.640	0.970	0.976
				$\tau = 2$				
$n = 10$	0.997	0.814	0.989	0.997	0.811	0.529	0.734	0.788
$n = 30$	0.998	0.262	0.988	0.996	0.888	0.479	0.802	0.844
$n = 50$	1.000	0.065	0.997	1.000	0.911	0.468	0.831	0.867