Comparison of bibliometric measures for assessing relative importance of researchers

Rok Blagus[1] · **Brane L. Leskošek**[1] · **Janez Stare**[1]

Abstract Quantitative evaluation of citation data to support funding decisions has become widespread. For this purpose there exist many measures (indices) and while their properties were well studied there is little comprehensive experimental comparison of the ranking lists obtained when using different methods. A further problem of the existing studies is that lack of available data about net citations prevents researchers from studying the effect of measuring scientific impact by using net citations (all citations minus self-citations). In this paper we use simulated data to study factors that could potentially influence the degree of agreement between the rankings obtained when using different indices with the emphasis given to the comparison of the number of net citations per author to other more established indices. We observe that the researchers publishing papers with a large number of co-authors are systematically ranked higher when using h-index or total citations (TC) instead of the number of citations per author (TCA), that the researchers who publish a small proportion of papers which receive many citations while the rest of their papers receive only few citations are systematically ranked higher when using TCA or TC instead of h-index, and that the authors who have lower proportion of self-citations are ranked higher when considering indices which include the number of net citations in comparison with indices considering only the total citation count. Results are verified and illustrated also by analyzing a large dataset from the field of medical science in Slovenia for the period 1986–2007.

✉ Rok Blagus
  rok.blagus@mf.uni-lj.si

  Brane L. Leskošek
  brane.leskosek@mf.uni-lj.si

  Janez Stare
  janez.stare@mf.uni-lj.si

[1] Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia

Introduction

Quantitative (bibliometric, scientometric) evaluation of publication and citation data is
now used in almost all countries around the globe, especially at universities and research
institutions, typically to support promotional or funding decisions. The simplest but useful
method for measuring researcher's impact is the citation count (Garfield 1972; Nerur et al.
2005), and among other, more complex, citation-based indices the Hirsch's (2005) *h*-index
has become the most popular. Variants of *h*-index include *g*-index (Egghe 2006), *c*-index
(Bras-Amorós et al. 2011) and *s*-index (Silagadze 2010). Properties of these measures were
well studied (Panaretos and Malesios 2009), however a comprehensive experimental
comparison of different methods on large datasets is scarce.

Bornmann et al. (2008) compared nine different variants of the *h*-index using data from
biomedicine and concluded that the indices that they studied can be divided in two groups.
The first group, including, among others, the *h* and *g*-indices, describes the most productive
core of a scientist's output and gives the number of papers in that core. The second group,
including the *m*-index (Bornmann et al. 2008) depicts the impact of the papers in that core.
Later this finding was verified by using molecular life sciences data (Bornmann et al. 2009).

Bornmann et al. (2011) performed a multilevel meta-analysis of studies reporting
correlations between the *h*-index and its variants and concluded that there is redundancy
between most of the *h*-index variants and the *h*-index, with few exceptions which included,
for example, the *m*-index.

One problem of the existing studies is that they mainly consider the total citation counts
when calculating various bibliometric measures. Total citation count is a sum of self-
citations and net citations where only the former truly express the scientific impact of the
researcher. However, lack of available data about net citations prevents researchers from
studying the effect of measuring scientific impact by using net citations. Schreiber (2007)
studied the effect of self-citations on the *h*-index using 13 physicists and showed that the *h*-
index can reduce from 10 to 46 % when excluding self-citations, however, this work only
considered *h*-index in the field of physics. Roediger (2006) also argued about self-citations
and concluded that they are not problematic when considering researchers with a large
number of citations, however, they can make a big difference for the scientists with a low
citation count.

Another issue with the current research, at least in our opinion, is that the citation count
is not measured in the number of citations per author. This means that if a paper has $k > 1$
authors, and the paper has n citations, each author is rewarded with all of them, instead of
only n / k. We strongly believe that this is wrong. An author, who published a paper on his/
her own and received n citations, should not be equaled to another, who coauthored a paper
with the same number of citations, but there were, say, 10 authors on the paper.

This issue for the *h*-index has already been studied (Bornmann and Daniel 2007; Hirsch
2007; Imperial and Rodrguez-Navarro 2007) and several corrections of the *h*-index that
account for the co-authorship were proposed. For example, Batista et al. (2006) proposed
to divide the *h*-index with the mean number of co-authors for the *h*-defining set of pub-
lications yielding h_I-index (Batista et al. 2006) and Schreiber (2008b) proposed to use

fractionalized counting of the papers according to the number of authors, i.e. the effective rank of publications is determined proportional to the number of authors, which yields an effective number which is utilized to define the h_m-index as that effective number of papers that have been cited h_m or more times. There are also modifications of the h-index that take into account also the author's rank in the byline (Wan et al. 2007; Tol 2011), however since the author's rank with the exception of the first and the last author is in biomedicine generally irrelevant these are not studied here.

One goal of our paper is then to asses the importance of using net citations per author instead of the total citation count when ranking the researchers. The second goal is to study the agreement between the rankings of researchers obtained by various bibliometric measures and to identify the factors influencing the level of agreement.

We study the above by using simulations, described in the next section, and by analyzing a real dataset. This dataset includes bibliometric measures for 1882 researchers from the field of medical science for the period 1986–2007. These measures are: the number of citations, the number of citations per author, the number of net citations, the number of net citations per author and various versions of the h-index. The main strength of our database, besides the large sample size and a long follow-up period, is that the net citations were measured very accurately and that we were able to very accurately calculate the number of citations per author. This was done automatically by analyzing the Science Citation Index database (and later Web of Science) using a program developed by Hristovski et al. (1996). Such data gathering became impossible when internet access to databases became much more restrictive, which is why our data do not include years after 2007.

In our study, a self-citation is every citation coming from a paper where at least one of the coauthors was also a coauthor of the cited paper. Net citations are then all citations minus the self-citations. Every net citation is then divided by the number of authors of the cited paper to get net citations per author. The starting point for such calculations is always a publication, and we've been building a national bibliographic database since 1975.

Methods

Three factors that could potentially influence the degree of agreement were considered,

1. Number of co-authors,
2. Distribution of citations across author's publications,
3. Proportion of self-citations.

We study these factors using simulations and a real data set. Obviously, one can expect that the first factor will show differences between indices which calculate citations per author and those which don't, the second will affect comparisons of citation counts with the h-index and its variants, and the third will rank higher those authors who have lower proportions of self-citations. Nevertheless, it is still interesting to see just what these effects are. And while real data sets are useful for illustration, it is better to study the behavior of the indices using simulations, since in simulations one controls the assumptions and the effects of the factors being studied.

Simulation study

In the simulation study we used Pareto type I distribution (Pareto 1897) to simulate the number of publications, the number of citations and the number of co-authors. Pareto type I

distribution has a strong positive skew and was shown to be appropriate for the simulation of publication and citation data and has been used extensively in the bibliometric research (Egghe 1987, 1991, 1998; Saam and Reiter 1999; Egghe 2005; Glanzel 2006; Tol 2011). The simulation setup for each factor was as follows.

1. Number of co-authors. Here we investigate the strength of agreement in a setting where there are two groups of authors, one group publishing papers with more coauthors than the other. For each of the 100 authors ($n = 100$) the total number of papers was simulated from the Pareto type I distribution (denoted as $P(\cdot)$), with the scale parameter (σ) set to 10 and shape parameter (α) set to $\alpha = 1.25$, resulting in 50 expected publications per researcher ($P(10, 1.25)$). The scale parameter which determines the minimum of the distribution was set to 10, so that the number of publications was not a possible factor influencing the degree of agreement, and the shape to 1.25 to mimic the results from our real data set. After obtaining the total number of publications for each author, the number of citations received by each paper was simulated. To simulate the number of citations we used the Pareto distribution with $\sigma = 1.8$ and $\alpha = 2.5$ ($P(1.8, 2.5)$) resulting in 3 expected citations per publication; we decided to use $\alpha > 2$, as the first and the second moment of the distribution are in this case finite. Note that since the minimum number of publications was set to 10 for each author, the expected number of citations is also the expected value of the h-index. For each paper we also simulated the number of co-authors. For this purpose the sample size was divided in two groups of equal size ($n_1 = n_2 = n/2$). The number of co-authors for the first group was simulated from $P(1.8, 2.5)$, corresponding to the expected value of 3 co-authors for each publication, while for the second group $P(a, 2.5)$ was used, where different values of a were considered ($a = 1.8, 3, 9, 15, 30, 60$, which corresponds to the expected number of co-authors in the second group equal to 3, 5, 15, 25, 50 and 100). Note that larger values of a mean larger heterogeneity between the two groups of authors.

2. Distribution of citations across author's publications. In this simulation setup, there were again two groups of authors, in one group the distribution of the number of citations was uniform across all publications, i.e. the expected number of citations was the same for all papers, while in the second group some papers received a large number of citations and the other papers were cited only a few times. For this purpose the total number of papers was simulated from $P(10, 1.25)$ and the number of co-authors was simulated from $P(1.8, 2.5)$ for all authors ($n = 100$). The entire sample was then divided in two groups ($n_1 = n_2 = n/2$). The total number of citations for each paper for the authors in the first group was simulated from $P(1.2, 2.5)$, while for the second group of authors a proportion of papers (π) were included in the "low" citation group where the number of citations was simulated from the $P(1.2, 2.5)$ and the rest of the papers ($1 - \pi$) were included in the "high" citation group, where the number of citations was simulated from $P(b, 2.5)$, where different values of b were considered ($b = 1.2, 3, 12, 30, 60$, resulting in the expected number of citations in the "high" citation group of papers equal to 2, 5, 20, 50 and 100). Different values of $1 - \pi$ were also considered ($1 - \pi = 0.05, 0.1, 0.2, 0.5, 0.9$).

3. Proportion of self-citations. Here the two groups of authors differed in the proportion of self-citations. The total number of papers was simulated from $P(10, 1.25)$, the number of co-authors was simulated from $P(1.8, 2.5)$ and the total number of citations for each paper was simulated from $P(6, 2.5)$, resulting in 10 expected citations per publication for all samples ($n = 100$). Then the sample was divided in two groups. For

the first group, a Bernoulli distribution with $p = 0.95$ was used to simulate if the citation received for each paper was a net citation or a self-citation; the expected proportion of net citations for each paper was thus 95 %. For the second group a Bernoulli distribution with $p = 0.05, 0.10, 0.20, 0.50, 0.75, 0.90$ was used to simulate if the citation was a net citation; thus the expected percentage of net citations that were received by each paper was equal to $p \cdot 100$ % for this group of authors.

In each setting all authors were ranked according to their total citation count, the number of citations per author, h-index, h_I-index, h_m-index (factors 1 and 2) as well as the number of net citations, the number of net citations per author and h_n-index, i.e. h-index considering only net citations (factor 3) and the strength of agreement was evaluated with Spearman's rank correlation coefficients (ρ). Each step of the simulation was repeated 10,000 times and average correlation coefficients and standard deviation (SD) obtained over 10,000 simulation runs are reported.

Data collection

The process of data collection and methodology applied in the empirical study was based on the well established national bibliometric database system Biomedicina Slovenica (BMS; http://ibmi.mf.uni-lj.si/en/services/biomedicina-slovenica), specialized for biomedicine and run by Institute for Biostatistics and Medical Informatics (IBMI). BMS contains all scientific papers published (or co-authored) by Slovenian authors since 1957. The citations for each paper/journal were collected from the internet version of Web of Science (WoS) database run by Thomson Scientific.

Our system behind BMS corrects for errors in the WoS like author names, citation linking, address information and similar. For this study the citation data were exported from BMS and organized in separate SQL database (MySQL running on Linux and supported by Perl scripts for data processing and parsing) where they were merged with additional metadata (e.g. author's institutions, research groups, etc.) received from publicly available national databases like Slovenian Research Agency's (ARRS—http://www.arrs.gov.si/en/dobrodoslica.asp) database of researchers, research groups and research organizations. From final merged database the indices were calculated and exported for processing in R.

In this study we used 146,102 papers published from 1957 until 2007 and cited from 1986 until 2007. The database contained the bibliometric measures for 1882 researchers divided among 56 research groups and 93 research organizations.

Indices

Below is a short description of the bibliometric indices used in our analysis.

– Total citation count (TC)—number of all citations for a researcher in publications where the researcher was (co)author.
– Total citation count per author (TCA)—number of all citations for a researcher in publications where the researcher was (co)author weighted by the number of co-authors.
– Net citation count (NC)—number of all citations excluding self-citations for a researcher in publications where the researcher was (co)author. Self-citation is defined as a citation where the citing and the cited papers shared at least one author.

– Net citation count per author (NCA)—number of all citations excluding self-citations for a researcher in publications where the researcher was (co)author and weighted by the number of co-authors.
– h-index—originally defined metric proposed by Hirsch, the h-index (Hirsch 2005), is defined as follows: a researcher has index h if h of researcher's N publications have at least h citations each and the other $(N - h)$ publications have less or equal to h citations each. A h-index of 7 means that the author has published 7 publications each having at least 7 citations. The value of h can only increase with time as more papers are published and cited.
– h_n-index—h-index that takes into consideration only net citations (Schreiber 2007; Bartneck and Kokkelmans 2011).
– h_{na}-index—h-index that takes into consideration only net citations in the same way as h_n-index. Additionally each citation was weighted (divided) by the number of authors of cited papers. This index gives more credit to the researchers publishing with less co-authors. A similar modification of the h-index was considered by Schreiber (2008a), but in this definition of the h-index, all citations were considered.
– h_f-index, h_{fn}-index and h_{fna}-index where defined in the same way as previous indexes h-index, h_n-index and h_{na}-index, respectively with the additional condition that for the observed researcher only publications where a researcher is the first author were included in index calculation. These indices are more rewarding for the first authors of the paper. Counting only those papers where researcher is the first author dates back to 1973 and the concept of first author counting (Cole and Cole 1973) but has not been applied to the h-index.

Additionally, in the simulation study we considered also the following modifications of the h-index that try to adjust for co-authorship,

– h_I index proposed by Batista et al. (2006), where h-index is divided by by the mean number of researchers in the h-core, i.e., in the h-defining set of publications;
– h_m index proposed by Schreiber (2008b) which is determined in analogy to the h-index, but counting the papers fractionally according to the number of authors, i.e. the effective rank of the publications is determined proportional to the number of authors, which yields an effective number which is utilized to define the h_m-index as that effective number of papers that have been cited h_m or more times.

Data analysis

The data are presented as median and interquartile range (IQR). When reporting proportions, exact binomial based confidence intervals are calculated. For all indices other than the h-type indices sum over the study period was calculated for each index and the sum was considered for ranking purposes. For h-type indices the largest value achieved in the entire study period was used when ranking the researchers. Two methods were used to estimate the level of agreement between the rankings obtained by different bibliometric measures; Spearman's rank correlation coefficients (ρ) with corresponding 95 % confidence intervals (CI) and a p value for testing the hypothesis of no association and Bland–Altman plots were constructed. Bland–Altman plots give a visual representation of the agreement between two measurements and were proposed as high correlation does not automatically imply good agreement (Bland and Altman 2010). When constructing Bland–Altman plots for our real data set relative ranks, calculated as $(R_i - 0.5)/n$, where R_i is absolute rank of

the individual i and n is the sample size, were used. All statistical analyses were performed with R language for statistical computing [(R version 3.0.1, R Core Team (2013)].

Results

Simulation study

Number of co-authors

The effect of the number of co-authors is shown in Table 1.

The results in the first column (three expected co-authors per paper) refer to the situation where there is no heterogeneity between the two groups. In this setting the best agreement was observed between the total citation count (TC) and number of citations per author (TCA) followed by the agreement between TC and h-index and TCA and h-index, while the worst agreement was observed between TC and h_I-index. The agreement between the variants of the h-index was a bit weaker (the strongest agreement was observed between h-index and h_m-index and the worst between h-index and h_I-index). When one

Table 1 Mean (SD) Spearman's rank correlation coefficients for different number of expected co-authors per paper

	Expected number of co-authors per paper					
	3	5	15	25	50	100
TCA and TC	0.9642	0.8334	0.5844	0.5363	0.5051	0.492
	(0.0104)	(0.0339)	(0.0729)	(0.0808)	(0.0844)	(0.0867)
TCA and h-index	0.8316	0.7319	0.5142	0.4676	0.4366	0.4233
	(0.0379)	(0.053)	(0.0792)	(0.0846)	(0.087)	(0.0871)
TC and h-index	0.8537	0.8533	0.8521	0.8522	0.853	0.8526
	(0.0328)	(0.0334)	(0.0334)	(0.0336)	(0.0336)	(0.0335)
TC and h_I-index	0.5284	0.4116	0.2833	0.2804	0.2807	0.2805
	(0.0793)	(0.0859)	(0.0944)	(0.0955)	(0.0943)	(0.0951)
TC and h_m-index	0.8132	0.7412	0.5694	0.5071	0.4636	0.4443
	(0.0425)	(0.0535)	(0.0708)	(0.08)	(0.0849)	(0.088)
TCA and h_I-index	0.6239	0.7344	0.8717	0.8919	0.9033	0.9074
	(0.0685)	(0.0503)	(0.0245)	(0.0203)	(0.018)	(0.0173)
TCA and h_m-index	0.8459	0.8619	0.9569	0.9649	0.9692	0.9708
	(0.0362)	(0.032)	(0.0108)	(0.009)	(0.008)	(0.0077)
h-Index and h_m-index	0.8019	0.7094	0.5292	0.4691	0.4252	0.4054
	(0.0432)	(0.0576)	(0.0751)	(0.082)	(0.0856)	(0.0867)
h-Index and h_I-index	0.6089	0.4727	0.3278	0.3242	0.3236	0.3237
	(0.0711)	(0.0812)	(0.0931)	(0.0941)	(0.0928)	(0.0924)
h_m-Index and h_I-index	0.6673	0.6763	0.8504	0.8815	0.8985	0.9047
	(0.0631)	(0.0576)	(0.0274)	(0.0211)	(0.018)	(0.0174)

group of authors published papers with more co-authors as the other group the strength of agreement between TCA and TC, TCA and h-index, TC and h_l-index, TC and h_m-index, h-index and h_l-index and h-index and h_m-index decreased substantially, while the agreement between TCA and h_l-index, TCA and h_m-index as well as h_l-index and h_m-index even slightly increased; this had no effect on the agreement between TC and h-index. This can be explained by noting that the information on the number of co-authors is not included in the calculation of TC and h-index, hence different distribution of this variable for each group of authors does not affect the strength of agreement.

Bland–Altman plot for a typical simulated dataset (a data set for which Spearman's rank correlation coefficient is equal to the mean correlation coefficient across the simulations) for the situation with 25 expected co-authors per paper is shown in Fig. 1. Since we are mainly interested in the comparison of TCA with other indices, only plots comparing the agreement between TCA and other indices are shown. Bland–Altman plot shows the association between the difference in the ranking obtained by two indices and the average
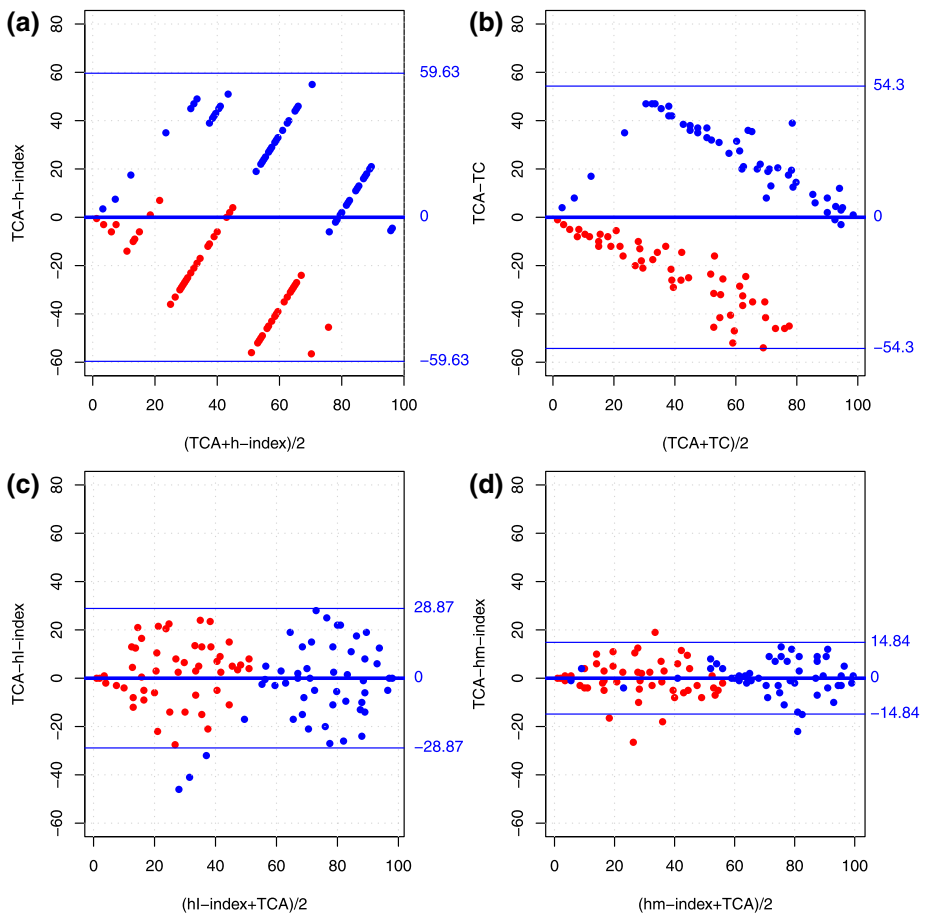

Fig. 1 Bland–Altman plots for a typical simulated data set. *Blue points* represent the "high" co-authorship group (expected value is 25 co-authors) and *red points* represent the "low" co-authorship group (expected value is 3 co-authors). (Color figure online)

rank. Ideally, the points would be close to the horizontal line and there should not be any pattern visible in the plot. We used different colors to denote the researchers publishing papers with a large number of co-authors (blue points) and those that publish papers with a small number of co-authors (red points). We can see in the first two panels in Fig. 1 that by looking from the direction of the y-axis, all blue points lie above the red points meaning that the researchers from the "high" co-authorship group were systematically ranked higher when using TC or h-index instead of TCA. In the third and forth panel in Fig. 1 red and blue points overlap in the direction of the y-axis, hence there was no systematic difference in the ranking lists obtained when using TCA and h_l-index or TCA and h_m-index (Fig. 1). However, in the third and fourth panel we can see a clear separation of blue and red points when looking from the direction of the x-axis which suggests that the researchers publishing with a lot of co-authors are ranked worse by TCA and h_l-index as well as TCA and h_m-index than researchers publishing papers with less co-authors, which was expected in this setting.

Distribution of citations across author's publications

The effect of the distribution of citations across author's publications on the agreement is reported in Table 2. Only the results for 5 % $(1 - \pi = 0.05)$ of papers in the "highly" cited

Table 2 Mean (SD) Spearman's rank correlation coefficients for different expected number of citations in the "highly" cited group of papers; 5 % of papers in the "highly" cited group $(1 - \pi = 0.05)$

	Expected number of citations in the "highly" cited group				
	2	5	20	50	100
TC and TCA	0.9646	0.9622	0.9544	0.9566	0.9619
	(0.01)	(0.0108)	(0.0134)	(0.0127)	(0.0111)
TCA and h-index	0.7535	0.7669	0.7625	0.7044	0.6476
	(0.0478)	(0.0476)	(0.0497)	(0.0594)	(0.0665)
TC and h-index	0.7733	0.7898	0.7853	0.7194	0.6576
	(0.0438)	(0.043)	(0.0457)	(0.0576)	(0.0662)
TC and h_l-index	0.4714	0.4922	0.5263	0.4894	0.4526
	(0.0844)	(0.0819)	(0.0799)	(0.0853)	(0.0873)
TC and h_m-index	0.6989	0.7223	0.6915	0.6228	0.5624
	(0.0612)	(0.0587)	(0.0636)	(0.073)	(0.0805)
TCA and h_l-index	0.5706	0.5956	0.6395	0.596	0.5506
	(0.074)	(0.0705)	(0.0668)	(0.0733)	(0.0765)
TCA and h_m-index	0.7364	0.7633	0.7412	0.6715	0.6086
	(0.0556)	(0.0516)	(0.0558)	(0.0654)	(0.0733)
h-Index and h_m-index	0.7095	0.7303	0.7429	0.7438	0.7442
	(0.0587)	(0.0553)	(0.0542)	(0.0531)	(0.0533)
h-Index and h_l-index	0.6011	0.6131	0.6573	0.659	0.6591
	(0.0728)	(0.0697)	(0.066)	(0.0662)	(0.0659)
h_m-Index and h_l-index	0.6871	0.6982	0.7324	0.7335	0.7337
	(0.0588)	(0.0565)	(0.053)	(0.0533)	(0.0527)

group of papers are reported, results for the other values of $1 - \pi$ are reported as additional information, see Online Resource 1.

The results in the first column show the situation where there was no heterogeneity between the groups, hence the same results were obtained for all values of $1 - \pi$ (Online Resource 1). As in the previous simulation setup the best agreement was observed between TCA and TC and the worst between TC and h_I-index. The agreement between TCA and TC was not affected by increasing the proportion of highly cited papers. This was expected since the number of authors per paper was in this setting the same for both groups.

On the other hand the agreement between h-index or h_m-index and TCA or TC was worse when the highly cited papers received more citations; this was less apparent when the proportion of highly cited papers increased (Online Resource 1). This can be explained as in the setting where the proportion of "highly" cited papers was small h-index for the authors in the second group was small, while their total citation count was high, hence the ranking obtained by the TC or TCA and h-index or h_m-index was different. When the proportion of highly cited papers increased, the value of h-index and h_m-index was no
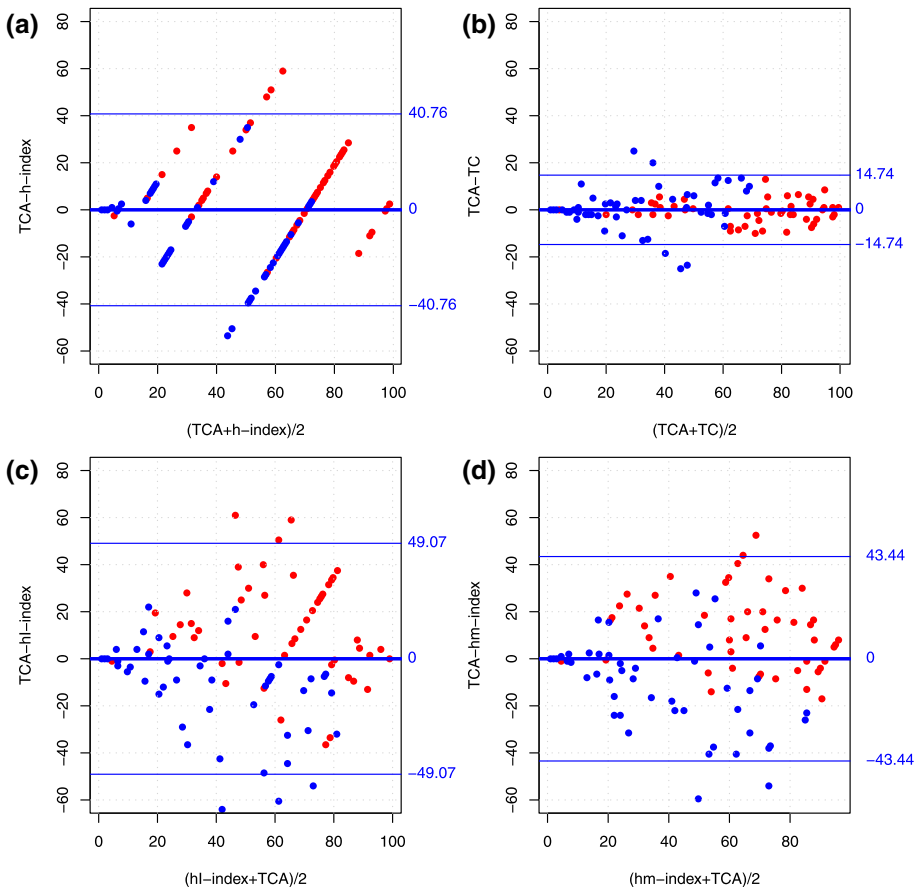

Fig. 2 Bland–Altman plots for a typical simulated data set. *Blue points* represent the "high" citation group (expected value is 50 citations for 5 % of highly cited papers and two citations for the rest of the papers) and *red points* represent the "low" citation group (expected value is two citations for each paper). (Color figure online)

longer limited by the small number of citations in the group of papers with low citation count and the agreement was better. Note also that the agreement between TC or TCA and h_I-index was not affected by the number of citations in the highly cited group of papers. Probable reason for this was poor agreement between these indices even in the setting with no heterogeneity.

Bland–Altman plot for a typical simulated dataset for the situation with 50 citations in the high citation group and 5 % of highly cited papers, revealed that the researchers from the high-citation group were systematically ranked better when using TCA instead of h-index, h_I-index or h_m-index (Fig. 2). There was no apparent systematic difference between the ranking obtained with TCA and TC, however, the researchers from the high-citation group were systematically ranked better than the researchers from the low citation group by the two indices (Fig. 2). While the plot for the comparison of different variants of the h-index also indicated poor agreement, there was no apparent systematic difference in the ranking obtained by different variants of the h-index, data not shown.

Proportion of self-citations

The effect of the proportion of self-citations on the agreement between NCA and other indices is summarized in Table 3; the agreement between the other pairs of indices is reported as additional information, see Online Resource 2.

Increasing the proportion of self-citations did not affect the agreement between all pairs of indices that consider all citations, i.e. TC, TCA, h-index, h_I-index and h_m-index (Online Resource 2). This was expected as these indices do not distinguish between net and self-citations. The agreement between NCA and these indices was however strongly affected by the proportion of net citations (Table 3). When one group of authors published papers where the proportion of net citations was smaller, the agreement was worse. As expected,

Table 3 Mean (SD) Spearman's rank correlations for different expected proportions of net citations

	Expected proportion of net citations					
	0.05	0.1	0.2	0.5	0.75	0.9
NCA and NC	0.9948	0.9947	0.994	0.9892	0.9849	0.9838
	(0.0016)	(0.0016)	(0.0018)	(0.0034)	(0.005)	(0.0053)
NCA and TC	0.4845	0.5358	0.6301	0.8599	0.9607	0.9816
	(0.0817)	(0.0769)	(0.0656)	(0.0296)	(0.0104)	(0.0058)
NCA and TCA	0.4909	0.5428	0.6376	0.8708	0.9744	0.9968
	(0.0814)	(0.0763)	(0.0648)	(0.0269)	(0.0061)	(0.001)
NCA and h-index	0.4446	0.4934	0.5826	0.79	0.872	0.888
	(0.0846)	(0.0806)	(0.0711)	(0.0448)	(0.0309)	(0.028)
NCA and h_I-index	0.3542	0.3952	0.466	0.622	0.6824	0.6946
	(0.091)	(0.0884)	(0.084)	(0.0705)	(0.0629)	(0.0617)
NCA and h_m-index	0.4639	0.5135	0.6051	0.8229	0.9163	0.9353
	(0.0833)	(0.0785)	(0.0684)	(0.0386)	(0.0226)	(0.0191)
NCA and h_n-index	0.9478	0.9381	0.9196	0.8818	0.882	0.8869
	(0.0117)	(0.0144)	(0.0198)	(0.0277)	(0.0287)	(0.0284)

the agreement between NCA and NC was not affected by the expected proportion of net citations (Table 3), while the agreement between NCA and h_n-index increased when the proportion of net citations was smaller.

The Bland–Altman plot for a typical dataset for the situation with 50 % of self-citation citations in the high self-citation group clearly showed that the researchers from the high self-citation group were systematically ranked higher when using h-index, TC or TCA than when using NCA (Fig. 3).

Real data

In this section we report the results based on the real data set. We describe the characteristics of the studied population, compare top lists of researchers obtained when considering different indices and report the level of agreement based on Spearman's rank correlation and Bland–Altman plots.

Characteristics of the studied population

The number of publications increased rapidly in the observed period, from on average 0.61 publications per researcher in 1986 (95 % CI: 0.44–0.77) to 3.94 publications per researcher in 2007 (95 % CI: 3.43–4.45). Similarly, the number of citations and net citations also increased from on average 0.03 in 1986 (95 % CI: 0.00–0.07) to 8.83 in 2007 (95 % CI: 6.65–11.01) and from 0.02 (95 % CI: 0–0.05) to 7.15 (95 % CI: 5.32–8.98), respectively. The number of co-authors also increased in the study period and hence the increase of the mean number of citations per author and net citation per author was much less
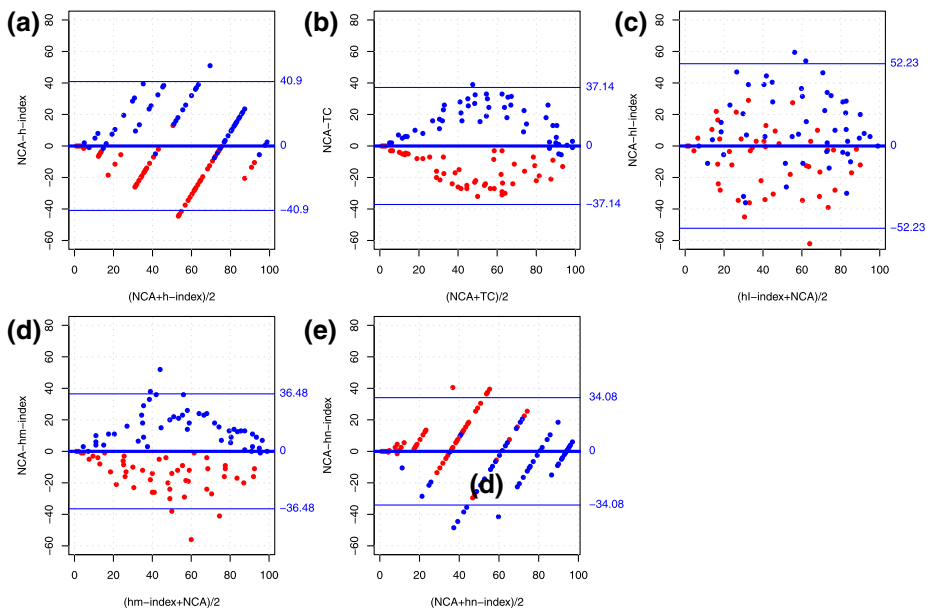
Fig. 3 Bland–Altman plots for a typical simulated data set. *Blue points* represent the "high" self-citation group (50 % of self-citations) and *red points* represent the "low" self-citation group (5 % of self-citations). (Color figure online)

pronounced. For example, the mean number of net citations per author increased from 0.0 in 1986 (95 % CI: 0–0.01) to 1.44 in 2007 (95 % CI: 1.11–1.76).

h-type indices also increased in the study period, the mean h-index was 0.01 in 1986 (95 % CI: 0.00–0.03) and 2.51 in 2007 (95 % CI: 2.18–2.84), but this behavior was expected since h-indices are cumulative.

Some descriptive statistics for the indices used to rank the researchers are reported in Table 4 (see the "Methods" section for more details).

The distribution of all indices showed strong positive skew suggesting that most researchers had very small values of the indices while some (few) researchers could have also very large values. The median total number of publications per researcher was 28.5 (interquartile range (IQR): 8–71) and median h-index and median NCA were both 1 (IQR: 0–7.91 and 0–3 for the number of net citations per author and h-index, respectively). Note that this means that the majority of researchers have very small, in the case of h-index even identical values, which has to be considered when interpreting the results for the agreement between the indices, especially when analyzing the entire set of researchers.

Top lists obtained with different indices

Researchers were ranked according to each index and we calculated the proportion of researchers who were simultaneously ranked in top 10, top 100 or top 500 by the number of net citations per author (NCA) and other indices. The results are shown in Table 5.

When considering the top 10 list only 1 researcher (10 %) who was ranked in the top 10 by the number of publications was ranked in the top 10 also by NCA. The percentage increased to 42 and 66 % in the top 100 and top 500 list, respectively. A similar result was observed also for the number of co-authors, suggesting that the Slovenian researchers with the most publications or the researchers who publish papers with a large number of co-authors are to a large extent not the researchers that have large number of net citations per author. This result can be explained to some extent by noting that the correlation between the number of publications and the number of co-authors when considering the entire sample is very large (Spearman's rank correlation coefficient: 0.96, p value <0.0001).

Very similar top lists were produced by NCA and TCA (80, 94 and 94.2 % for the top 10, top 100 and top 500, respectively), while the lists obtained with h-index were slightly

Table 4 Descriptive statistics for the maximum index achieved in the entire observation period

	Min	Max	Median	Q1	Q3
# of Publications	1.00	483.00	28.50	8.00	71.00
TC	0.00	3001.00	5.00	0.00	44.75
NC	0.00	2412.00	4.00	0.00	34.00
# of Authors	1.00	8277.00	103.00	28.00	282.00
TCA	0.00	506.28	1.38	0.00	10.37
NCA	0.00	352.19	1.00	0.00	7.91
h-Index	0.00	31.00	1.00	0.00	3.00
h_n-Index	0.00	25.00	1.00	0.00	3.00
h_{na}-Index	0.00	10.00	0.00	0.00	1.00
h_f-Index	0.00	15.00	0.00	0.00	2.00
h_{fn}-Index	0.00	15.00	0.00	0.00	1.00
h_{fna}-Index	0.00	10.00	0.00	0.00	1.00

Sum over the observed period was considered for the indices other than h-type indices (see the "Methods" section for more details)

Table 5 Percentage (95 % confidence intervals) of the researchers who were simultaneously ranked in top 10, top 100 or top 500 by the number of net citations per author (NCA) and other indices

	Top 10	Top 100	Top 500
# of Publications	10.00 (0.25–44.50)	42.00 (32.20–52.29)	66.00 (61.66–70.15)
TC	60.00 (26.24–87.84)	77.00 (67.51–84.83)	88.60 (85.48–91.25)
NC	40.00 (12.16–73.76)	75.00 (65.34–83.12)	90.60 (87.70–93.01)
# of Co-authors	10.00 (0.25–44.50)	49.00 (38.86–59.20)	68.00 (63.71–72.07)
TCA	80.00 (44.39–97.48)	94.00 (87.40–97.77)	94.20 (91.78–96.08)
h-Index	60.00 (26.24–87.84)	72.00 (62.13–80.52)	81.60 (77.92–84.90)
h_n-Index	70.00 (34.75–93.33)	82.00 (73.05–88.97)	92.20 (89.49–94.39)
h_{na}-Index	70.00 (34.75–93.33)	78.00 (68.61–85.67)	84.20 (80.70–87.29)
h_f-Index	50.00 (18.71–81.29)	56.00 (45.72–65.92)	74.20 (70.13–77.98)
h_{fn}-Index	50.00 (18.71–81.29)	44.00 (34.08–54.28)	70.20 (65.98–74.18)
h_{fna}-Index	60.00 (26.24–87.84)	53.00 (42.76–63.06)	82.00 (78.35–85.27)

TC total citation count, *TCA* total citation count per author

more different with the percentages in the top 10, top 100 and top 500 lists ranging from 60 72 and 81.6 %, respectively. The lists produced by h_f-index were the most different, with the respective percentages being 50, 56 and 74.2 % in the top 10, top 100 and top 500 lists.

Agreement between the indices

Spearman's rank correlation coefficients calculated for different subsets of the data are shown in Table 6.

The results for all researchers showed strong agreement between NCA and other indices; the exceptions were h_f-type indices where the level of agreement was slightly worse (Table 6). This was confirmed also by the Bland–Altman plots (Fig. 4), where we observed strong agreement between NCA and TCA and slightly weaker agreement

Table 6 Spearman's rank correlation coefficients (95 % CI) for the agreement between the net citations per author and other indices

	All data	Top 100—# of publications	Top 100—# of co-authors
TC	0.9758 (0.9735–0.9779)	0.9426 (0.9155–0.9611)	0.7228 (0.6135–0.805)
NC	0.9861 (0.9848–0.9873)	0.9314 (0.8993–0.9535)	0.6879 (0.5682–0.779)
TCA	0.9884 (0.9874–0.9894)	0.9899 (0.9849–0.9932)	0.9883 (0.9826–0.9921)
h-Index	0.9594 (0.9557–0.9629)	0.947 (0.9219–0.9641)	0.824 (0.7488–0.8783)
h_n-Index	0.9738 (0.9714–0.9761)	0.9397 (0.9113–0.9592)	0.8008 (0.7172–0.8617)
h_{na}-Index	0.9308 (0.9245–0.9366)	0.9338 (0.9028–0.9551)	0.9407 (0.913–0.9598)
h_f-Index	0.8086 (0.7924–0.8237)	0.7001 (0.5833–0.7885)	0.7284 (0.6208–0.8091)
h_{fn}-Index	0.8049 (0.7884–0.8203)	0.7265 (0.6176–0.808)	0.7369 (0.6319–0.8153)
h_{fna}-Index	0.7509 (0.7305–0.77)	0.6278 (0.4913–0.7342)	0.68 (0.5581–0.7731)

between NCA and h-index where 5 % of the researchers differed in relative ranking by more than 0.16. Similar level of agreement was observed also between NCA and h_n-index, while the rankings based on h_f-type indices and NCA were more different; in this case 5 % of the researchers differed in relative ranking by more than 0.34. Also noteworthy is that the researchers who were on average ranked lower by the two indices were systematically ranked worse when using NCA instead of h_f-type indices.

However, strong agreement could be misleading as most researchers in our database had small and identical values of the indices and were therefore ranked equally.

When only 100 researchers with the most publications were considered in the analysis the agreement was slightly weaker, however it remained strong with only few exceptions (see also Fig. 5). The largest difference when compared with the analysis with all researchers was observed when comparing the agreement between NCA and h_n-index (Fig. 5). In this case 5 % of the researchers differed in relative ranking by more than 0.13 (analysis with all researchers) and 0.20 (analysis considering only 100 researchers with the most publications).
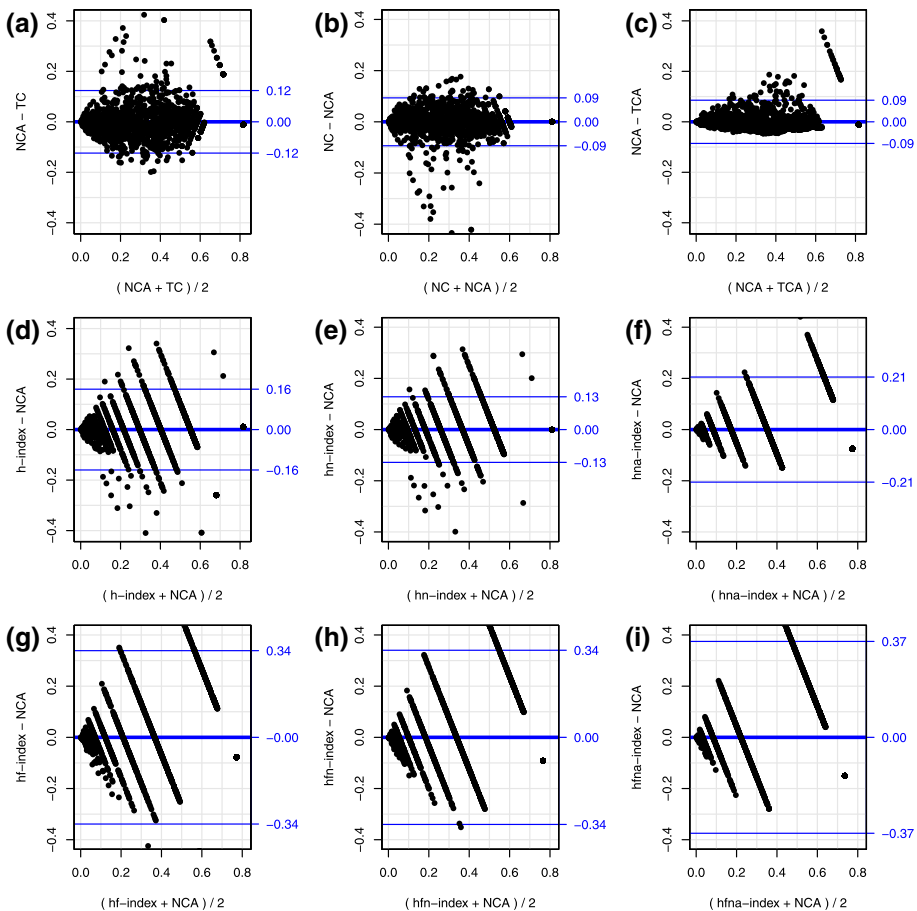

Fig. 4 Bland–Altman plots for all researchers

The reason for strong agreement between the TC and NCA in this analysis is that we observed a strong positive correlation between the number of co-authors and the total citation count (Spearman's rank correlation: 0.8132, $p < 0.0001$) as well as a strong positive correlation between the number of co-authors and the number of net citations per author (Spearman's rank correlation: 0.80135, $p < 0.0001$). Therefore, the ranking lists of indices using either of these variables are expected to be similar for this dataset. Much weaker agreement was obtained when analyzing the data for a subset of 100 researchers that published papers with the most co-authors (Table 6), where the correlation between the number of co-authors and the number of net citations per author was very weak (Spearman's rank correlation: 0.0806, $p = 0.4251$).

Weaker agreement for this subset of the data is evident also from Bland–Altman plots (Fig. 6), with 95 % of the differences being in the range of ± 0.4 for most indices which can be compared with the range ± 0.1 for the entire set of researchers. The exception was the agreement between NCA and TCA as well as NCA and h_{na}-index, where the agreement was similar as in the analysis with all researchers. These indices all use the number of co-
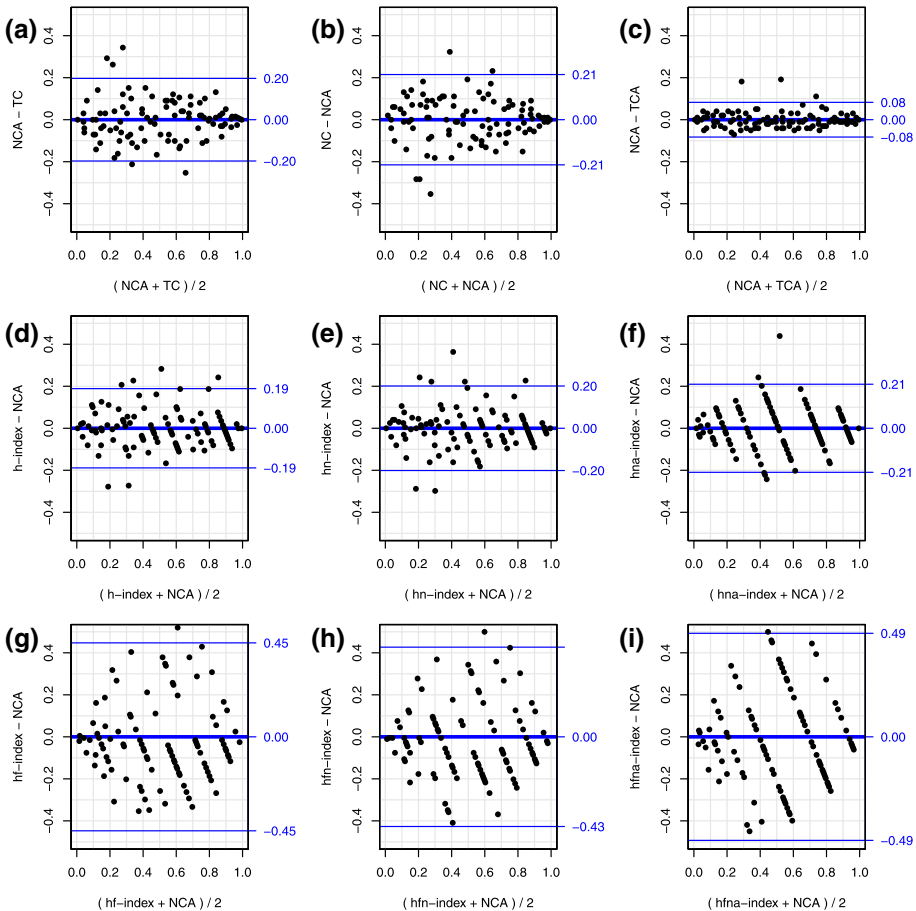

Fig. 5 Bland–Altman plots for 100 researchers with the most publications

authors in their calculation and hence changing the distribution of this variable in the data set does not have an affect on the level of agreement.

Discussion and conclusions

Our simulation study shows that researchers publishing papers with a large number of co-authors are systematically ranked higher when using h-index or TC instead of TCA, but there was no systematic difference in the ranking lists obtained when using TCA and h_l-index or TCA and h_m-index. As expected, different agreement between indices which calculate citations per author and those which don't is observed when varying the number of co-authors. This is a simple consequence of the fact that the information on the number of co-authors in not included in the calculation of TC and h-index.

The simulation study also shows that the researchers who publish a small proportion of papers which receive many citations while the rest of their papers receive only few
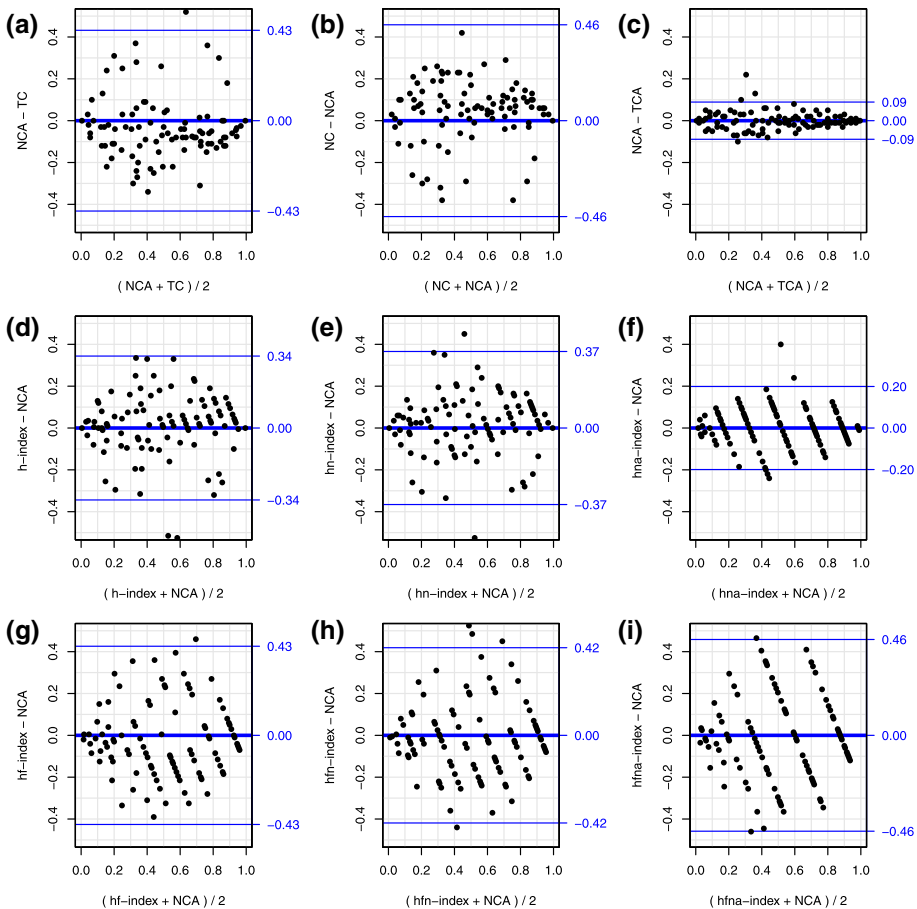
Fig. 6 Bland–Altman plots for 100 researchers publishing papers with the most co-authors

citations are systematically ranked higher when using TCA or TC instead of h-index or h_m-index, while there is no apparent systematic difference between the ranking obtained with TCA and TC.

Similarly, authors who have lower proportion of self-citations are ranked higher when considering indices that include the number of net citations in comparison with indices considering only the total citation count. The ranking of the researchers obtained by TC, TCA, h-index, h_I-index and h_m-index does not depend on the proportion of self-citations, hence the agreement between these indices does not depend on this factor.

The empirical analysis for Slovenian medical Researchers shows good agreement between the indices. One reason is that most researchers have very small and also identical values of different indices and are therefore ranked equally by all indices. For example, the first quartile for all indices is equal to zero, meaning that at least 25 % of the researchers will be ranked completely identically when using different indices. We analyzed also a subset of 100 researchers who publish the most papers. The agreement is slightly worse for this subset of the data, however it remains large with only few exceptions. This result might seem somewhat unexpected but it can be explained for our data set. For this subset of the data we observed a strong positive correlation between the number of co-authors and the total citation count as well as a strong positive correlation between the number of co-authors and the number of net citations per author. Therefore, indices using either of these variables will agree well. However, when there is no correlation between the variables used in the calculation of the indices then the ranking lists can also be considerably different. This is confirmed by analyzing a subset of 100 researchers that publish papers with the most co-authors, where the correlation between the number of co-authors and the number of net citations per author is very weak and consequently the agreement between NCA and other indices is much weaker.

For our empirical dataset it does not make much difference if the researchers are ranked by the total number of citations or the number of citations per author. One probable reason for this is that the Slovenian medical researchers are all very similar in terms of the number of co-authors, median number of co-authors per publication in 2007 for example was 3.7 with interquartile range from 2.6 to 5.2, and hence it made little difference to differentiate between the total citation count and the number of citations per author. This would be very different if we were to rank researchers from two (or more) different science disciplines where publication practices are different. Publication practices could include the common number of co-authors per paper, the proportion of self-citation and similar. For example, if we ranked Slovenian mathematicians, who mostly publish papers alone or with at most 1 co-author, and medical researchers then the mathematicians would generally be ranked relatively worse than medical researchers when using the total citation count instead of the number of citations per author, even when the impact of their publications would have been similar. We argued in the introduction that an author, who published a paper on his/her own and received n citations, should not be equaled to another, who coauthored a paper with the same number of citations, but with more co-authors on the paper. Therefore, care is needed when ranking researchers from different science disciplines, especially when the publication practices vary across disciplines. In this case the use of an inappropriate index can systematically disregard scientists from one discipline.

That h-index cannot be used off-hand to compare researchers of different areas has been pointed out by Hirsch himself (2005), and this was the motivation behind the h_I-index, which should enable the comparison between different research areas (Batista et al. 2006). However, this index only adjusts for one difference between the research areas, namely the difference in the number of co-authors. Our simulation study shows that this index cannot

adjust for the differences between research areas in terms of self-citations or to some extent also the distribution of citations across author's publications. For example, the h_I-index systematically disregards the scientists with a lower proportion of self-citations when compared with NCA. Iglesias and Pecharroman (2006) also addressed the issue of comparing scientists from different fields and concluded that comparison of the h-index is in this case meaningless unless the indices are properly corrected for the fact that different science fields have different average values of citations per paper.

Our results suggest that while the agreement between the h-index after correcting for the factors (such an example is h_{na}-index) instead of using a modified h-type index (such an example is h_m-index) is not perfect, there is no systematic difference between the ranking list obtained by using a modified h-index or h-index corrected for the factors. Hence we prefer to use the later, as these indices are easier to obtain and furthermore, they can be easily adapted to account for more possible factors. As an example, the h_n-index adjusts for possible differences in the proportion of self-citations, and if we want to further adjust the index for possible differences in the number of co-authors, the net citations are divided by the number of co-authors to obtain the h_{na}-index. Note that this modification, while possible, would be much more difficult for the h_m-index. Our study also confirms that any modification of the h-index will systematically disregard scientists with a small number of highly cited papers when compared with TCA or NCA, hence we would prefer to use NCA instead of h_{na}-index in the ranking process anyway, however this view is to some extent subjective.

All variables in our simulation study were simulated independently. The analysis with correlated effects would be straightforward to perform, but it would then be much harder to understand the effect of each factor, similarly as is the case for our empirical dataset. We considered also the exponential distribution to simulate the number of publications and citations, as well as discrete uniform distribution to simulate the number of publications and the number of co-authors per paper and we observed that this did not change our conclusions (data not shown).

In conclusion, if indices use the same amount of information, or speaking in statistical terms, if the same variables are used to derive the indices, the agreement between these indices will be good, and when different variables are used to derive the indices, then the ranking obtained by these indices can disagree if there is some variability in the studied sample in terms of this variable. Obviously, the agreement will be worse when the variability is larger. Also importantly, even when indices use different variables in their derivation, the ranking lists obtained when using the indices can be very similar when there is a strong positive correlation between the variables used to derive the indices as was observed for our empirical dataset.

References

Bartneck, C., & Kokkelmans, S. (2011). Detecting h-index manipulation through self-citation analysis. *Scientometrics, 87*(1), 85–98.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics, 68*(1), 179–189.

Bland, J. M., & Altman, D. G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies, 47*(8), 931–936.

Bornmann, L., & Daniel, H. D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology*, *58*(9), 1381–1385. doi:10.1002/asi.20609.

Bornmann, L., Mutz, R., & Daniel, H. D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, *59*(5), 830–837.

Bornmann, L., Mutz, R., Daniel, H. D., Wallon, G., & Ledin, A. (2009). Are there really two types of h index variants? A validation study by using molecular life sciences data. *Research Evaluation*, *18*(3), 185–190.

Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H. D. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics*, *5*(3), 346–359.

Bras-Amorós, M., Domingo-Ferrer, J., & Torra, V. (2011). A bibliometric index based on the collaboration distance between cited and citing authors. *Journal of Informetrics*, *5*(2), 248–264.

Cole, J. R., & Cole, S. (1973). *Social stratification in science*. Chicago: The University of Chicago Press.

Egghe, L. (1987). An exact calculation of price's law for the law of lotka. *Scientometrics*, *11*(1–2), 81–97.

Egghe, L. (1991). The exact place of zipf's and pareto's law amongst the classical informetric laws. *Scientometrics*, *20*(1), 93–106.

Egghe, L. (1998). Mathematical theories of citation. *Scientometrics*, *43*(1), 57–62.

Egghe, L. (2005). A characterization of the law of lotka in terms of sampling. *Scientometrics*, *62*(3), 321–328.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479.

Glanzel, W. (2006). On the h-index—A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, *67*(2), 315–321.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16,569–16,572.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19,193–19,198. doi:10.1073/pnas.0707962104.

Hristovski, D., Rožić, A., & Adamič, Š. (1996) A decision support system for biomedical research evaluation. In *Medical informatics Europe '96: Human facets in information technologies*, pp 609–613

Iglesias, J. E., & Pecharroman, C. (2006). *Scaling the h-index for different scientific isi fields*. arXiv:physics/0607224.

Imperial, J., & Rodrguez-Navarro, A. (2007). Usefulness of hirsch's h-index to evaluate scientific research in spain. *Scientometrics*, *71*(2), 271–282. doi:10.1007/s11192-007-1665-4.

Nerur, S., Sikora, R., Mangalaraj, G., & Balijepally, V. (2005). Assessing the relative influence of journals in a citation network. *Commun ACM*, *48*(11), 71–74.

Panaretos, J., & Malesios, C. (2009). Assessing scientific research performance and impact with single indices. *Scientometrics*, *81*(3), 635–670.

Pareto, V. (1897). *Course d'économie politique*. Lausanne: F. Rouge.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roediger, H. L. (2006). The h index in science: A new measure of scholarly contribution. *APS Observer 19*.

Saam, N., & Reiter, L. (1999). Lotka's law reconsidered: The evolution of publication and citation distributions in scientific fields. *Scientometrics*, *44*(2), 135–155.

Schreiber, M. (2007). Self-citation corrections for the hirsch index. *EPL (Europhysics Letters)*, *78*(3), 30,002.

Schreiber, M. (2008a). A modification of the h-index: The hm-index accounts for multi-authored manuscripts. *Journal of Informetrics*, *2*(3), 211–216.

Schreiber, M. (2008b). To share the fame in a fair way, h_m modifies h for multi-authored manuscripts. *New Journal of Physics*, *10*(4), 040,201.

Silagadze, Z. (2010). Citation entropy and research impact estimation. *Acta Physica Polonica B*, *41*(11), 2325–2333.

Tol, R. (2011). Credit where credits due: Accounting for co-authorship in citation counts. *Scientometrics*, *89*(1), 291–299.

Wan, J. K., Hua, Ph, & Rousseau, R. (2007). The pure h-index: Calculating an authors h-index by taking co-authors into account. *Collnet Journal of Scientometrics and Information Management*, *1*(2), 1–5.