Scientometrics (2023) 128:6773–6780
https://doi.org/10.1007/s11192-023-04852-w

Some thoughts on transparency of the data and analysis behind the Highly Cited Researchers list

Alexandra-Maria Klein[1] · **Nina Kranke**[1]

Received: 12 June 2023 / Accepted: 2 October 2023 / Published online: 28 October 2023

Abstract
Clarivate's Highly Cited Researchers (HCR) list is one of the most important quantitative performance indicators in science and influences individual careers and also the reputation of research institutions. However, many researchers and representatives of institutions and funding agencies are not familiar with the method that is applied to generate the HCR lists. We therefore provide a detailed description of Clarivate's method and give an example to illustrate how HCR are identified. While Clarivate provides the complete HCR lists and a detailed description of the method used for identifying HCR, the detailed analysis with actual numbers is not published. It is therefore not entirely transparent how exactly the HCR were identified, and which authors were excluded from the initial list, e.g. due to scientific misconduct. It is also impossible to check the analysis for errors. Given the prestige and weight that is attributed to Clarivate's HCR list, we call for transparency of the data and analysis behind the HCR list.

Keywords Highly Cited Researchers · Clarivate · Science performance indicators · Reproducibility

Introduction

Metrics such as the *h*-index, number of publications, and number of citations play an important role in assessments of individual researchers, research groups, and institutions and are viewed as indicators for influence and success (Langfeldt et al., 2021). One of the most prestigious quantitative performance indicators is Clarivate's annually updated Highly Cited Researchers (HCR) list. Researchers awarded as HCR are seen as the most influential researchers globally. Furthermore, research institutions all over the world showcase their HCR to present themselves as institutions that facilitate excellent research and attract other excellent researchers and students. In fact, the number of HCR is important for how universities are perceived and ranked. In the Academic Ranking of World Universities (Shanghai Ranking) for example, the number of HCR constitutes 20% of an institution's score (Shanghai Ranking, 2023). Being honoured as HCR can also boost the careers of

✉ Nina Kranke
nina.kranke@nature.uni-freiburg.de

[1] Chair of Nature Conservation and Landscape Ecology, University of Freiburg, Freiburg, Germany

individual researchers as this award can be advantageous in application procedures for jobs and research grants. While the HCR list is widely used as a performance indicator, many institutions and individuals are unaware of the exact method that Clarivate applies to identify HCR, highlighting the need for increased transparency and accountability in the publication of science performance indicators. We therefore summarise Clarivate's method in the following sections. Finally, we argue that Clarivate should reveal the data and analysis behind the HCR list because the publication of science performance indicators should be subject to the same standards and values as scientific publications in terms of transparency and reproducibility.

Clarivate's method

Within the last few years the methodology for identifying HCR has been revised several times. Here we describe the approach that was used to identify HCR for the 2022 list. Clarivate Analytics collects the data from the Essential Science Indicators (ESI) database, which is updated every other month (Clarivate, 2023a). ESI is a compilation of science performance statistics derived from the Web of Science Core Collection data that includes peer-reviewed original research articles and review articles from indexed journals (Clarivate, 2023b). The database also includes highly cited papers (Clarivate, 2023c). Within this database, papers are assigned to 22 broad categories of research fields based on the journals they appear in. Twenty-one of these categories represent specific research fields (e.g. Chemistry, Engineering, Environment/Ecology) and the twenty-second category is labelled 'multidisciplinary'. The HCR fields correspond to the 21 specific research fields in the ESC database (Clarivate, 2023a).

Each journal and the papers published therein are assigned to only one field except for multidisciplinary journals such as *Science*, *Nature*, *American Scientist* and *Bioengineering & Translational Medicine*. All papers that appear in journals that represent one of the specific research fields are assigned to the journals' respective categories. For example, all papers that appeared in *BMC Immunology* are assigned to the category *Immunology* without reviewing the papers' content or references. Papers that appeared in multidisciplinary journals, however, are individually reviewed and assigned to one of the 21 specific research fields based on the most frequent category of the article's cited references (Clarivate, 2023a). A list of journals that includes the respective categories can be downloaded from http://esi.help.clarivate.com/Content/scope-coverage.htm. The data retrieved from ESI for the 2022 HCR list included approximately 179.000 highly cited papers, i.e., papers that rank in the top 1% by total citations for their ESI field and year (Clarivate, 2023a). Highly cited papers that have been retracted and highly cited papers with more than 30 authors or explicit group authorship are excluded from the analysis (Clarivate, 2023a).

The HCR lists are based on highly cited papers that have been published in an 11-year period (for the 2022 list from 2011 to 2021; Clarivate, 2023d). The approximate number of researchers who are selected as HCR in each of the specific research fields is determined by the square root of the total population of authors of highly cited papers in the respective field. This step in the analysis was introduced to account for the different sizes of the 21 research fields. Authors with highly cited papers in different fields are Highly Cited Researcher candidates in the cross-field category (see below). The threshold number of highly cited papers that a researcher needs to be considered for the list, is the number of papers at the rank of the square root. Not only do highly cited papers need to rank in the

top 1% by total citations for their ESI field and year, but researchers must also have enough total citations of their highly cited papers in the respective field to be considered for the HCR list. All researchers with highly cited papers at the threshold for inclusion and above whose total citations rank among the top 1% in their field, are included in the initial list. Researchers with one fewer highly cited paper than the threshold number are also included in the list, if they have enough citations of their highly cited papers to rank them in the top 50% of researchers at the threshold level and higher (Clarivate, 2023a). The number of researchers who are included in the list may therefore exceed the number of the square root calculation of all authors within the field.

Example

Consider the fictional example for the identification of HCR candidates in research field 1 (Table 1). If the total number of authors in field 1 was 1.800, the square root of 1.800 is the number of researchers included as HCR for field 1. In this case, the number of HCR candidates in field 1 is 42. The threshold of highly cited papers that a researcher needs to be considered for the HCR list, is 7 because the researcher at rank 42 has 7 highly cited papers. This means that all researchers who have 7 or more highly cited papers who also rank among the 1% of total citations are admitted to the initial list. If the threshold for the top 1% of total citations of highly cited papers in field 1 was 650, researchers on ranks 1 to 43 would be included in the list. Although Researcher I has 7 highly cited papers, they would not be admitted to the list because their total number of citations is under the threshold level. Say researchers with 6 highly cited papers in field 1 need more than 989 total

Table 1 List of candidates for HCR in field 1 (fictional example)

Rank	HCR candidates in field 1	Number of highly cited papers in field 1	Total number of citations of highly cited papers in field 1
1	Researcher A	27	3.546
2	Researcher B	27	3.423
3	Researcher C	26	2.899
…	…	…	…
39	Researcher D	8	975
40	Researcher E	8	842
41	Researcher F	7	1.240
42	Researcher G	7	937
43	Researcher H	7	765
44	Researcher I	7	638
45	Researcher J	6	1048
46	Researcher K	6	1.002
47	Researcher L	6	804
48	Researcher M	6	773
49	Researcher N	6	547
50	Researcher O	5	449

Researchers are ranked by number of highly cited papers in field 1 and total number of citations of highly cited papers in field 1

citations of their highly cited papers to rank in the top 50% of researchers at the threshold level or higher and get included in the list. This applies to researchers J and K but not to researchers L, M, and N. The preliminary list of HCR in field 1 therefore includes 45 individuals (researchers on ranks 1–43 plus researchers J and K). In this example the number of researchers included in the list (45) exceeds the number of the square root of the total number of authors in the field (42).

Cross-field category

To honour influential researchers who have published highly cited papers in different fields but not enough papers in any one field to appear in the HCR ranking for the specific research fields, Clarivate introduced the cross-field category in 2018. This measure has led to an increase in HCR from about 3.500 (in 2017) to about 6.000 (in 2018) (Clarivate, 2023e). For the ranking in the cross-field category, the highly cited paper and citation counts are fractionated according to the threshold numbers of the respective field. Researcher O has 5 highly cited papers in field 1. Their paper score for field 1 is thus 5/7ths or 0.714. The citation counts are fractionated in a similar manner. Researcher O's citation score for field 1 is 449/650ths or 0.69. If the sums of all field paper scores and all field citation scores are 1 or higher, the researcher is selected as HCR in the cross-field category (see fictional example on Clarivate's (2023a) website).

In 2022 approximately 45% of the HCR are listed in the cross-field category followed by Clinical Medicine (6.45%) and Biology and Biochemistry (4.19%) (Table 2, Clarivate, 2023f). On the one hand, it could be argued that this distribution represents the growing number of multidisciplinary and interdisciplinary research projects as well as multidisciplinary competences of individual researchers as the cross-field category constitutes almost half of all HCR. On the other hand, this category lumps together researchers from many different research areas. Given the large number of researchers who rank in the cross-field category, it might be useful to further subdivide this category into different cross-fields (Chen, 2022).

Exclusions

The preliminary lists of HCR in the specific research fields and the cross-field category are further scrutinized to find and exclude researchers who have committed scientific misconduct and are involved in gaming strategies to increase the number of publications and citations. Clarivate not only excludes retracted highly cited papers but also uses the publicly available Retraction Watch data base (https://retractionwatch.com/) to identify putative HCR whose publications, that are not highly cited, have been retracted for reasons of scientific misconduct such as plagiarism, image manipulation, and fake peer review (Clarivate, 2023a). Researchers who were involved in scientific misconduct in formal proceedings are also excluded from the HCR list.

Since 2019 Clarivate also excludes researchers with unusually high levels of self-citation or collaborative group citations. To identify authors with high levels of self-citation, the method described by Szomszor et al. (2020) is used. Clarivate analysts also scrutinize authors with outsized output and exclude these researchers from the list, if more than half of their citations derive from coauthors. Additional filters are used to identify and exclude

Table 2 Numbers of HCR in the cross-field category and the specific research fields (source: https://clarivate.com/highly-cited-researchers/analysis/, 17.03.2023)

HCR category	Number of HRC	%
Cross-field	3244	44.90
Clinical Medicine	466	6.45
Biology and Biochemistry	303	4.19
Chemistry	270	3.74
Social Sciences	270	3.74
Neuroscience and Behaviour	225	3.11
Materials Science	222	3.07
Immunology	214	2.96
Molecular Biology and Genetics	206	2.85
Environment and Ecology	202	2.80
Psychiatry and Psychology	191	2.64
Plant and Animal Science	185	2.56
Physics	176	2.44
Engineering	153	2.12
Pharmacology and Toxicology	153	2.12
Geosciences	148	2.05
Microbiology	129	1.79
Agricultural Sciences	116	1.61
Computer Science	115	1.59
Space Science	93	1.29
Economics and Business	92	1.27
Mathematics	52	0.72
Total	7225	100.00

authors with "suspect citation activity", but these are not revealed "in the interest of staying ahead of those attempting to game [the] identification of Highly Cited Researchers" (Clarivate, 2023a). According to Clarivate (2023a), about 300 authors were excluded from the 2021 list. The number of excluded researchers in 2022 increased to about 550 as a result of the implementation of additional filters. Unfortunately, Clarivate does not publish lists with excluded papers and authors, so that it remains non-transparent who has been excluded for which reason.

Transparency and reproducibility

We explained how Clarivate proceeds to identify HCR. While the current and past HCR lists can be downloaded from Clarivate's website and a detailed description of the method that was used to generate the current HCR list is available, the data and analysis with actual numbers, is not published. According to Clarivate, this is "to prevent gaming or manipulation of the system" (personal communication via email on 23.12.2022). It is thus not transparent how exactly the HCR were identified and which authors were excluded from the initial list. It is also impossible to check the analysis for errors. We know from our experiences and discussions with colleagues that Clarivate analysts make mistakes. We do

not criticize this, as mistakes do happen but it would be helpful, if everyone could review the analysis.

The ESI thresholds (e.g. the number of citations received by the top 1% of authors) and the total number of authors in each field are available and can be downloaded from the ESI website (Clarivate, 2023f, 2023g). The highly cited papers of HCR candidates and the categories that they are assigned to can also be retrieved from the ESI database (Clarivate, 2023h). People who have access to the ESI database (e.g. via their institutions) could thus reconstruct Clarivate's analysis to a certain extent. It would be a great deal of effort to collect these data and process them based on Clarivate's description of their method, but it would theoretically be possible to reconstruct their initial list.

With these numbers one could calculate field paper thresholds and identify the researchers that were selected for the initial lists. Some of the exclusions could also be reconstructed with the Retraction Watch data base and the method to identify unusually high levels of self-citation described by Szomszor et al. (2020). Since Clarivate does not reveal their additional filters to exclude authors with suspect citation activity, it is impossible to completely reproduce the analysis. Given the prestige and weight that is attributed to Clarivate's HCR list, we call for transparency of the data and analysis. This way, the list could be checked for errors more easily and it would also be easier for authors to understand why they were not included in the list and compare themselves to other researchers. It would furthermore help to understand why female researchers are underrepresented in HCR lists and advise Clarivate how to change their methods to mitigate gender bias (Shamsi et al., 2022; Langfeldt et al., 2021; see also Bradshaw et al., 2021). It would also be possible to see how many papers in what fields were published by authors who were ranked in the cross-field category which could make a further subdivision of the cross-field category unnecessary.

We recognize the importance of addressing gaming and misconduct and it is understandable that Clarivate does not reveal the additional filters used to exclude authors from the initial list to stay ahead of gaming strategies. We nevertheless argue that at least the actual numbers and the detailed analysis that produce the initial list should be made public. Given that Clarivate already provides a rather detailed general description of their method, they might as well publish the actual numbers to facilitate a review of their results. We argue that not publishing actual numbers will not prevent people from trying to manipulate the system, as the increased number of exclusions shows. If someone is willing to manipulate their citations or game the system to become HCR, they would do this whether or not Clarivate publishes the data and analysis.

Transparency and reproducibility are considered standards for good scientific practice and journals as well as funding agencies increasingly require or at least encourage public data (raw and processed), method, algorithm, software, and code sharing (e.g. Nature, 2014; Nature Geoscience, 2014; Wiley, 2023). Ideally, the same standards should be applied to science performance indicators while taking care of "explicit biases" in science performances. Gaming and misconduct are encouraged by research institutions and funding agencies that attach great importance to metrics (Biagioli, 2016) and will not be curbed by keeping data and analysis of quantitative performance indicators hidden. Since the problem is systemic, its solution needs to be systemic as well. If researchers were not only rewarded for publication output and citations, but also for behaviour that strengthens research integrity (see Moher et al., 2020), gaming and cheating activities would likely decrease.

Preventing gaming might be one of the reasons why the actual numbers and detailed analysis are not made public. However, being a publicly traded company, Clarivate is driven by commercial motives to safeguard their reputation and promote their products.

Modifications in the algorithm, coupled with the shifts resulting from the 11-year cut off, introduce a certain amount of unpredictability, thereby preserving the element of novelty associated with the annual release of the HCR list. From an economic standpoint, Clarivate undeniably has an incentive to maintain the irreproducibility of the analysis or at least make it difficult to reproduce and review the analysis. This unconscious or conscious incentive ensures that users of the list continue to rely on Clarivate as the authority on HCR status. It's worth noting that Clarivate faces competition from various free or low-cost alternatives to their products. Hence, even if the issue of gaming were to be resolved, it remains unclear if Clarivate would choose to release the actual data and complete analysis, considering that they valuably already disclose substantial detailed information on their methodology.

HCR status relates only to a small elite, but the use of metrics to evaluate researchers is a widespread practice. A more general aim of this letter is to promote a critical examination of metrics that are used to evaluate and compare researchers. As each metric has its strengths and biases (Bornmann et al., 2008; Bradshaw et al., 2021), it is important to understand what exactly is measured and how. We have done this exercise using the example of Clarivate's HCR list, because the analysis is rather complex compared to other metrics. When using of referring to the HCR list, one should keep in mind that it likely serves commercial interests more than other scientometric indicators. However, in general, it is advisable not to overestimate the importance and relevance of quantitative performance indicators but instead use them carefully, e.g. as starting points for a more thorough evaluation. It should also be noted that most metrics do not correct for gender bias and other biases that are still prevalent in academia (see Bradshaw et al., 2021) and therefore overestimate the performance of privileged individuals and groups.

Acknowledgements We thank the two anonymous reviewers for valuable comments and feedback.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded by the German Research Foundation (DFG) project number 452861007/FOR 5281.

Declarations

Conflict of interest The authors have no relevant non-financial interests to disclose.

References

Biagioli, M. (2016). Watch out for cheats in citation game. *Nature, 535*(7611), 201–201. https://doi.org/10.1038/535201a

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. https://doi.org/10.1002/asi.20806

Bradshaw, C. A. J., Chalker, J. M., Crabtree, S. A., Eijkelkamp, B. A., Long, J. A., Smith, J. R., Trinajstic, K., & Weisbecker, V. (2021). A fairer way to compare researchers at any career stage and in any discipline using open access citation data. *PLoS ONE, 16*(9), e0257141. https://doi.org/10.1371/journal.pone.0257141

Chen, X. (2022). Does cross-field influence regional and field-specific distributions of highly cited researchers? *Scientometrics, 128*, 825–840. https://doi.org/10.1007/s11192-022-04584-3

Clarivate. (2023a). *Highly Cited Researchers—methodology*. Retrieved March 23, 2023, from https://clarivate.com/highly-cited-researchers/methodology/

Clarivate. (2023b). *Essential Science Indicators Help: Scope and coverage*. Retrieved April 12, 2023, from https://esi.help.clarivate.com/Content/scope-coverage.htm

Clarivate. (2023c). *LibGuides. Essential Science Indicators: Learn the basics*. Retrieved March 23, 2023, from https://clarivate.libguides.com/c.php?g=593878&p=4107958

Clarivate. (2023d). *Essential Science Indicators Help: Highly cited thresholds*. Retrieved March 23, 2023, from http://esi.help.clarivate.com/Content/highly-cited-thresholds.htm. Accessed 23 March 2023

Clarivate. (2023e). *Highly Cited Researchers—past lists*. Retrieved March 23, 2023, from https://clarivate.com/highly-cited-researchers/past-lists/

Clarivate. (2023f). *Highly Cited Researchers—analysis*. Retrieved March 23, 2023, from https://clarivate.com/highly-cited-researchers/analysis/

Clarivate. (2023g). *InCites Essential Science Indicators: Citiation thresholds*. Retrieved March 23, 2023, from https://esi.clarivate.com/ThresholdsAction.action

Clarivate. (2023h). *InCites Essential Science Indicators: Top papers by research fields*. Retrieved March 23, 2023, from https://esi.clarivate.com/IndicatorsAction.action

Langfeldt, L., Reymert, I., & Aksnes, D. W. (2021). The role of metrics in peer assessments. *Research Evaluation, 30*(1), 112–126. https://doi.org/10.1093/reseval/rvaa032

Meho, L. I. (2022). Gender gap among highly cited researchers, 2014–2021. *Quantitative Science Studies, 3*(4), 1003–1023. https://doi.org/10.1162/qss_a_00218

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., Coriat, A.-M., Foeger, N., & Dirnagl, U. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLOS Biology, 18*(7), e3000737. https://doi.org/10.1371/journal.pbio.3000737

Nature. (2014). Code share. *Nature, 514*(7524), 536–536. https://doi.org/10.1038/514536a

Nature Geoscience. (2014). Towards transparency. *Nature Geoscience, 7*(11), 777–777. https://doi.org/10.1038/ngeo2294

Shanghai Ranking. (2023). *Shanghai Ranking's Academic ranking of world universities methodology 2022*. Retrieved March 23, 2023, from https://www.shanghairanking.com/methodology/arwu/2022

Shamsi, A., Lund, B., & Mansourzadeh, M. J. (2022). Gender disparities among highly cited researchers in biomedicine, 2014–2020. *JAMA Network Open, 5*(1), e2142513. https://doi.org/10.1001/jamanetworkopen.2021.42513

Szomszor, M., Pendlebury, D. A., & Adams, J. (2020). How much is too much? The difference between research influence and self-citation excess. *Scientometrics, 123*(2), 1119–1147. https://doi.org/10.1007/s11192-020-03417-5

Wiley. (2023). *Wiley's data sharing policy*. Retrieved March 27, 2023, from https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html