Scientometrics
https://doi.org/10.1007/s11192-024-05158-1

The definition of highly cited researchers: the effect of different approaches on the empirical outcome

Rainer Frietsch[1] · Sonia Gruber[1] · Lutz Bornmann[2]

Received: 2 May 2024 / Accepted: 12 September 2024

Abstract

In 2001 onetime and since 2014 annually, Clarivate (and the former Thomson Reuters) has used publication and citation data to identify exceptional researchers—highly cited researchers (HCRs)—in nearly all disciplines. The approach used by Clarivate has not been without criticism. HCRs can be defined differently; the approach of Clarivate is one possibility among several others. HCRs can be identified by considering field-normalized citation rates or absolute numbers of citations; inclusion or exclusion of self-citations; full counting or fractional counting of publications; all authors, only corresponding authors or only first authors; short, long or varying citation windows; and short or long publication periods. In this study, we are interested in the effect different approaches have on the empirical outcomes. One may expect HCRs lists with large overlaps of authors, since all approaches are based on the same (bibliometric) data. As we demonstrated with five different variants of defining HCRs, the selection among these options has a significant influence on the sample of selected researchers and their characteristics that are thereby defined as highly cited. Some options have a stronger influence on the outcome than other options such as the length of the citation window or the focus on all authors versus only the corresponding author. Based on the empirical results of this study, we recommend that the user of HCR lists should always be aware of the influence these options have on the final lists of researchers.

Keywords Bibliometrics · Highly cited researchers · Clarivate · Scopus

✉ Rainer Frietsch
Rainer.Frietsch@isi.fraunhofer.de

Sonia Gruber
Sonia.Gruber@isi.fraunhofer.de

Lutz Bornmann
bornmann@gv.mpg.de

[1] Competence Center Innovation and Knowledge Economy, Fraunhofer Institute for Systems and Innovation Research, Breslauer Str. 48, 76139 Karlsruhe, Germany

[2] Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

Published online: 21 January 2025

Introduction

Scientometric research found very skewed distributions in output and citations of researchers (Seglen, 1992). Although the number of publishing researchers grew at an exponential rate over the years (Dong et al., 2017), the law of scientific productivity means that "the number of scientists producing N papers is proportional to $1/N^2$" (van Raan, 2019, p. 238). Empirical analyses of 15 million researchers show correspondingly that only 1% was able to publish at least one paper annually. This small core of researchers published yet around 40% of all papers and around 90% of all papers with high citation counts (more than 1000 citations) (Ioannidis et al., 2014). Other empirical research suggests that the small core of very productive researchers is a relatively stable group over years (Abramo et al., 2017). According to Wang and Barabási (2021) "sustained high productivity is rare, but it correlates with scientific impact and eminence. Given this evidence, it may appear that productivity is the key indicator for a meaningful career in science. Yet … among the many metrics used to quantify scientific excellence, productivity is the least predictive. The reason is simple: While great scientists tend to be very productive, not all scientists who are productive make long-lasting contributions" (p. 15). To separate outstanding from great researchers, citation data can be applied in addition to output data.

In 2001 onetime and since 2014 annually, Clarivate (and the former Thomson Reuters) has used publication and citation data to identify exceptional researchers—highly cited researchers (HCRs)—in nearly all disciplines. The purpose of Clarivate with the analysis of highly cited papers is "to recognize scientists and social scientists with community-wide influence" (Szomszor et al., 2020, p. 1124). The ranking list of HCRs worldwide is annually published at https://clarivate.com/highly-cited-researchers/. The annual publication of the list is associated with high public attention (which can be observed, e.g., on Twitter, see da Silva & Bernes, 2018), and the institutional number of HCRs is a regularly used indicator in the well-known Academic Ranking of World Universities (ARWU, see https://www.shanghairanking.com).

Clarivate admits that its method is not perfect and that it represents a particular selection of authors that fit their pre-defined criteria. In the description of its method, they stress therefore: "each measure or set of indicators, whether total citations, h-index, relative citation impact, mean percentile score, etc., accentuates different types of performance and achievement. Here we arrive at what many expect from such lists but what is unobtainable: that there is some optimal or ultimate method of measuring performance" (Clarivate, 2021). However, this is a tautological statement in its methodological implications—as right will always differ from wrong (Klein & Kranke, 2023). Since the core question is whether the method keeps what the concept promises, the choice of the method is a matter of reliability and validity in the specific empirical application.

It is therefore not surprising that the approach used by Clarivate has not been without criticism. For example, it has been criticized that Clarivate does not consider authorship positions in its definition: a paper is assigned to a researcher with full counting without considering that the researcher might have contributed only a marginal part (as reflected in their middle position in the list of co-authors) or significant part (as reflected in the first or corresponding authorship position) (see, e.g., Baerlocher et al., 2007; Fox et al., 2018; Tscharntke et al., 2007). Clarivate has adapted its method over the years to tackle (some) of these criticisms. Most recently, for example, it introduced the following—at first sight–minor change: the exclusion of papers with more than 30 authors. Previously, Clarivate excluded only papers with more than 30 distinct affiliations. For some organizations and some countries like Switzerland or Germany, this "minor" change of the method has a great influence on the national numbers of HCRs.[1]

HCRs can be defined differently; the approach of Clarivate is one possibility among several others. In this study, we are interested in the effect different bibliometric approaches have on the empirical outcomes. One may expect HCRs lists with large overlaps of authors, since all approaches are based on the same (bibliometric) data, referring to the same period of time and the same overall population of authors. The above example with the differing results for Germany and Switzerland (depending on the consideration of papers with more than 30 authors or not) demonstrates, however, that this is probably not the case. We employ four alternative, justifiable approaches for identifying HCRs and compare their empirical outcomes with the outcome from the approach used by Clarivate. We do not believe that a "universally true" list of HCRs exists and therefore do not intend to suggest a single approach as "best approach". We rather intend to sensitize users of HCRs lists on the effect the methodology has on the outcome. In order to assess an approach, it is necessary to have a clear, solid, and reasonable definition of what a HCR essentially should represent in a particular analytical situation.

In this study, we are interested in the similarities and differences of the HCRs lists from the various approaches: Do we receive very similar or different lists? What is the effect of the different approaches on aggregated results (i.e., national and institutional numbers of HCRs)? We use four indicators to analyze the characteristics of the HCRs (resulting from the different approaches): the average scientific age of HCRs, their gender and institutional affiliation as well as the average team size they publish with. The final analysis deals with the application of the different approaches in a certain context: How are the different approaches able to identify Nobel laureates? Do the different HCRs lists contain a similar number of laureates, and what are the reasons for possible different results?

Clarivate's highly cited researchers database

Methods used for the identification of highly cited researchers

Docampo and Cram (2019) and Schwartz et al. (2010) provide a detailed overview of the development of the methods used by Clarivate to identify HCRs since 2001. Basically, Clarivate applies an algorithm to cluster highly cited papers by authors followed by a cleaning process based on visual inspection. According to Szomszor et al. (2020) "the publication clusters associated with Highly Cited Researchers have a high level of manual curation" (p. 1125). The non-transparent cleaning process leaves room for speculation on the exclusion of certain researchers from the list of HCRs. While criteria like scientific

[1] Clarivate changed the methodology in 2022, but did not report HCRs lists based on the changed methodology retrospectively. Therefore it is not possible to directly assess the effect of the methodology change on the lists. We have therefore used our own approximation of Clarivate's method based on Scopus data to estimate the effect. All other factors being equal, we have only altered the selection of highly-cited papers, excluding papers with 30 or more authors versus 30 or more affiliations. For the period 2010–2018, the number of HCRs in Germany increases by 2.1% from 1,543 to 1,576 authors. For Switzerland, the number decreases by 8.5% from 800 HCRs based on the exclusion of papers with 30 or more affiliations to 732 HCRs (when papers with 30 or more authors were excluded). Smaller countries show even larger relative effects than larger countries. For the USA, the effect is rather moderate with an increase of 0.6% from 7653 to 7701 HCRs. These results show that small variations in the methodology might have considerable effects on HCRs lists and that these effects are not equally distributed across countries.

fraud or misconduct as well as withdrawn papers might lead to exclusion, it is unclear which other criteria for manual curation have been used over the years. In the list from 2023, three criteria are explicitly mentioned: extreme levels of hyper-authorship, excessive self-citation, and unusual patterns of collaborative group citation activity (i.e. suspicions for citation cartels).

For the identification of HCRs in the current form (since 2014), Clarivate selects papers (articles and reviews) from various disciplines that have appeared in a period of eleven years (Clarivate Analytics, 2021). For the 2021 HCRs list, e.g., the publication years 2010 to 2020 have been considered. Within every discipline and publication year (between 2010 and 2020), Clarivate identified those papers that belong to the 1% most frequently cited. The authors of the selected highly cited papers were sorted within every discipline: the more highly cited papers there were for a researcher, the higher their rank in the discipline. For the finally published list of HCRs, two criteria are relevant: (1) researchers with a rank less than or equal to the square root of the population consisting of all authors in a discipline with at least one highly cited paper; (2) researchers with enough citations to their highly cited papers in a discipline "to rank among all authors in the top 1% by total citations" (Clarivate Analytics, 2021, p. 18). For the identification of the early 2001 HCRs list, the second criterion was the only criterion used (Docampo & Cram, 2019). The 2021 HCRs list includes about 3,800 researchers in 21 single disciplines.

In 2018, Clarivate started to identify researchers not only within one discipline, but also across several disciplines (Clarivate Analytics, 2021). The most important reason for the additional list of HCRs was that the discrimination of researchers publishing in several disciplines should be avoided (that unavoidably results from the focus on only papers in one discipline to build the lists). The method for selecting HCRs across different disciplines is explained by Clarivate Analytics (2021) as follows: "The challenge for us was finding a method that took account of the different threshold number of highly cited papers in each field so that those contributing papers in several fields could be compared in an equal manner with those selected in one or more ESI [Essential Science Indicators] fields. The solution chosen was to fractionally count the credit for each highly cited paper such that a paper in a field with a high threshold number of papers was weighted less than a paper in a field with a lower threshold number of papers" (p. 20). The 2023 HCRs list includes about 3332 HCRs across several disciplines.

Studies based on the highly cited researchers database

Since the introduction of Clarivate's HCRs database, several studies have been published that are based on the database. These studies have investigated the database (focusing on critical points of the approach) and/or have used the database as data source for empirical studies.

One of the earliest studies based on the 2001 HCRs list was published by Bauwens et al. (2011). The authors generated institutional and national aggregates based on HCRs counts. Their results reveal high concentrations of HCRs on only a few institutions and countries worldwide. Bornmann and Bauer (2015b) used the 2014 HCRs list to analyze the worldwide institutional distribution of HCRs. Since the affiliations of the HCRs in the 2014 list were presented with various institutional name variants, Bornmann and Bauer (2015b) disambiguated the data in an expensive process. They identified the University of California with its different locations as the institution with the most HCRs. In a first follow-up study based on the same 2014 HCRs list, Bornmann et al. (2015) analyzed the distribution of men and women among HCRs. They found that "there are significantly more men (87%) than women (13%) among the scientists" (p. 2715). Gender distributions among HCRs have also been investigated by Shamsi (2022) based on the 2018 HCRs list. The author found with about 11% a similar share of women as the study by Bornmann et al. (2015). In a second follow-up study of Bornmann and Bauer (2015b), Bornmann and Bauer (2015a) focused on the German scientific landscape and analyzed the distributions of HCRs by federal state and city. The tables in Bornmann and Bauer (2015a) reveal that most HCRs from Germany are located in Bavaria (federal state) and Munich (city). The results also show that HCRs are regionally concentrated in Germany.

Some other studies have used the HCRs database—similar to Bornmann and Bauer (2015a)—to undertake national empirical analyses focusing on Brazil, China, and South Africa:

Martinez and Sá (2020) took a closer look at HCRs from Brazil. The study is based on nine HCRs from the 2018 HCRs list. For the HCRs, Martinez and Sá (2020) did not only search for all papers published by the HCRs, but also academic CVs using a database provided by the Brazilian National Council for Scientific and Technological Development (Brasília). The empirical results of the study have been summarized by the authors as follows: "connections with core countries in the global North are critical for most of Brazil's highly cited researchers. Most of these scientists work at prestigious public universities and are inserted in international research networks. While all but one of them obtained their PhDs from Brazilian universities, they have been internationally mobile from the early stages of their careers" (p. 51).

For investigating Chinese research success, Li (2018) analyzed HCRs lists from 2001, 2014, 2015, and 2016 and additionally government funding data from the Chinese National Science Foundation (Peking). Similar to the enormous Chinese growth rates in numbers of papers and researchers, the number of HCRs has also increased: from seven in 2001 to 160 in 2014 (with a primary affiliation located in mainland China). On the institutional level, Li (2018) not only counted the institutional number of HCRs in China, but also the number of HCRs per 1000 academic faculty staff. This ratio reveals, e.g., a relative high value for the Tsinghua University in 2016, although the Chinese Academy of Science employs significantly more HCRs.

Diko (2015) analyzed the 2014 HCRs list with a focus on South Africa. On this list, South Africa was the only country from the African continent hosting HCRs. The results of the study emphasize the problem of dealing with more than one affiliation for a single HCR in the institutional or national aggregation of HCRs data. Among ten researchers associated with an institution in South Africa, four researchers indicate the South African affiliation as second affiliation with the first affiliation in the USA, Norway, Belgium, or Switzerland. A focus on only the first affiliation in aggregating HCRs data would underestimate South Africa's achievements; the additional consideration of further affiliations of HCRs would overestimate the national achievement.

Two other papers focused on HCRs from Iran (Kamali et al., 2022) and Japan (Nagane et al., 2018). Kamali et al. (2022) investigated the single papers published by the Iranian HCRs and found many retracted papers in the database: about one-tenth of the HCRs had retracted papers. Based on their results, Kamali et al. (2022) demand from Clarivate that "unethical conduct by any researcher must be considered before nominating them as an HCR. Undertaking fake peer-review, manipulating any peer-review, or duplications are sufficient reasons to rescind a researcher's HCR status". Nagane et al. (2018) used HCRs lists published in 2014, 2015, and 2016, and included 119 HCRs from Japan in their study. The two institutions in Japan with the most HCRs are the Institute of Physical and Chemical

Research (RIKEN) and the Osaka University. The concentration of HCRs on only a few institutions in Japan is explained by Nagane et al. (2018) as follows: “first, researchers in Japan tend to perform collaborative research within universities or institutes. Second, labor mobility is low within Japanese universities. Third, apprenticeship relations in Japan often last for a lifetime”.

Butler et al. (2018) did not focus on HCRs from a specific country, but from a specific discipline. The results are, however, also relevant for interpreting the results for other disciplines. The authors investigated a sample of 29 HCRs from the 2017 HCRs list in “Social Sciences, general”. They found that many medicine-based researchers are among the HCRs. The results point thus to the problem of identifying HCRs in rather broad disciplines: Researchers working in certain social sciences fields with basically higher citation rates than other fields have an advantage to appear on the list. In the case of the selected HCRs by Butler et al. (2018), “medical related research gains a higher rate of citation compared to fields within social science, [it] is … any wonder than that the Social Science HCR list is dominated by medical researchers.” This problem that has been identified for the social sciences can be generalized to other disciplines: researchers publishing in certain fields in broad HCRs disciplines will have an advantage of entering the HCR status than in other fields.

Must (2020) matched data from the 2014 HCRs list with ResearcherID data to build a sample of n = 329 researchers. The results show that “HCRs have greater mobility, in the majority of cases they belong to the group who has been in the field for 11 to 15 years, they are highly productive, publishing on average over 100 articles each and interdisciplinary research plays a definite role in their careers” (Must, 2020, p. 198). Sinay et al. (2020) also analyzed a sample (random sample) of HCRs (n = 198); this time from the 2018 HCRs list yet. They found that “the profile of the highly cited scholars, as established by Clarivate Analytics, is so narrow that it may compromise the validity of scientific knowledge, because it is biased towards the perception and interests of male scholars affiliated with very-highly developed countries where English is commonly spoken and of their sponsors. This highly cited scholars accounted for 76% of the random sample analyzed, absent were women from Latin-America, Africa, Asia, and Oceania, and scholars affiliated with institutions in low-human-developed countries. Also, 98% of the published research came from institutions located in very-highly developed countries”.

Li (2016) extended the analyses by Bornmann and Bauer (2015b). Li (2016) not only revealed institutional distributions of HCRs counts, but also of HCRs efficiency whereby efficiency is measured by linking institutional number of researchers and number of HCRs. The same has been done by Li (2016) on the national level by linking Gross Domestic Product (GDP) and population with the number of HCRs. Li (2016) points out in this context that “the more meaningful correlation is between average GDP in 100 billion, and number of highly-cited researchers, with a correlation coefficient of approximately 0.93. This strong correlation signifies that countries with more funds tend to do higher quality and quantity of research, therefore resulting in more highly-cited researchers” (pp. 70/77).

In one of the most recent studies, Aksnes and Aagaard (2021) investigated a sample of HCRs from the 2018 HCRs list (n = 150 of about 6000 HCRs). The authors were interested in the question whether HCRs are really—as Clarivate claimed—a “small fraction of the researcher population that contributes disproportionately to extending the frontier and gaining for society knowledge and innovations that make the world healthier, richer, sustainable, and more secure” (Aksnes & Aagaard, 2021, p. 43). The authors found that HCRs are very productive in terms of paper output compared to “median” researchers, and reviews (published by more than one author) were disproportionally frequent among their highly cited papers. Thus, it seemed that not articles with primary research results were crucial for being on the HCRs list, but frequently cited summaries of the literature. On average, highly cited papers published by the HCRs in the sample of Aksnes and Aagaard (2021) have 59 co-authors. Their results also show that "15% of the HCRs published highly cited papers with an average of over 100 authors, and 7% with 50–100 authors. Thus, these scientists tend to be members of large research consortia and their presence on the [2018 HCRs] list can be attributed to such memberships" (p. 54). These results indicate that the full counting approach used by Clarivate for identifying HCRs may be questioned. Only those papers should be counted for being awarded as an exceptional researcher with a significant contribution of the researcher themselves.

Studies providing other methods to identify HCRs

Li et al. (2019) considered the top 5% of authors in a certain discipline and period– in terms of citation counts—to identify HCRs, which—once on the list—stay there for the rest of their academic life. They add further criteria for the selection of top scientists, namely career start between 1980 and 1998 and at least 20 years of activity. Their research interest was to examine if the collaboration with HCRs in early phases of a career is a good predictor of future academic success of young scholars, based on their academic impact. They find empirical evidence for their hypotheses.

Ioannidis et al. (2019) criticized Clarivate's method for identifying HCRs as follows: (1) Clarivate uses a coarse classification of science in only 21 disciplines. (2) Clarivate selects "only about 6000 scientists (https://hcr.clarivate.com/worlds-influential-scientific-minds), i.e., less than 0.1% of the total number of people coauthoring scholarly papers". (3) Clarivate does not exclude self-citations for measuring citation impact. Based on the three critical points, Ioannidis et al. (2019) produced a freely available dataset including about 100,000 HCRs based on the Scopus database. The HCRs were selected based on "their ranking of a composite indicator that considers six citation metrics (total citations; Hirsch h-index; coauthorship-adjusted Schreiber hm-index; number of citations to papers as single author; number of citations to papers as single or first author; and number of citations to papers as single, first, or last author)". The composite indicator is explained in more detail in Ioannidis et al. (2016). An updated version of the HCR dataset has been presented by Ioannidis et al. (2020).

Rodríguez-Navarroa and Brito (2021) compared HCRs data on the national level based on the definition by Clarivate and Ioannidis and co-authors (Ioannidis et al., 2019, 2020). Although the authors found a high correlation between the number of HCRs based on both definitions, they also found notable differences for countries with only a few numbers of HCRs. Rodríguez-Navarroa and Brito (2021) concluded that their empirical results validated the definitions of HCRs by Ioannidis and co-authors, and that "the number of WoS [Web of Science]-HCR [by Clarivate] fails to assess correctly the success of research in some specific countries".

Methods

Given the criticism of Clarivate's HCRs approach, the aim of the following analyses is not to suggest an optimized approach. The aim rather is to compare different approaches of identifying HCRs, thereby assessing the effects of different implementations on the empirical outcomes. Next to an approximation of the approach by Clarivate (first variant in the following), we employed these additional approaches: focusing on the top-1% of authors (second variant), applying a fractional accounting of citations (instead of a full counting) (third variant), considering only correspondence authors (instead of all authors of a paper) (fourth variant). The fifth variant—the approach by Ioannidis et al. (2020)—has not been implemented by ourselves using Scopus data. We used the dataset provided by the authors themselves.

Dataset used

As a data source for our bibliometric analyses we used Scopus raw data in the form of a snapshot from the end of April 2021, covering all publications up to the cohort of the publication year 2020. The raw data are implemented as an Oracle-SQL database with different add-ons (e.g., field classifications and gender information).[2] We are aware of the coverage differences between WoS and Scopus in terms of journals, disciplines, or author countries (see, e.g., Schmoch et al., 2011; Stahlschmidt et al., 2019). This is the reason why we approximated Clarivate's HCR approach with our own data instead of employing existing HCRs lists as they are annually published by Clarivate.

The use of Scopus is beneficial for the purpose of identifying HCRs, as an author disambiguation (AU-ID) as well as an institutional disambiguation (AFF-ID) is provided by Scopus at a global scale. A reliable disambiguation, especially on the author level, is mandatory for identifying HCRs in the database, since one can expect that all publications of an author are traced under a single ID. Research about the Scopus AU-ID has shown that the reliability of the paper assignments to authors is given (Aman, 2018; Moed et al., 2013). The fact that the AU-ID exists for all authors offers the possibility to focus not only on a selected set of papers or authors, but to take all publications of all authors into account for identifying HCRs. For generating the list of HCRs, Clarivate focuses on a selected set of papers, i.e., only on papers (and their authors) that belong to the top-1% papers in the WoS database.

Generating different highly cited researchers lists to be compared

In this study, we used five different approaches to identify HCRs in the Scopus database—the approximation of the Clarivate method and four alternative implementations. For each of these five, the Scopus AU-ID has been used as an identifier of individual authors. For the four approaches implemented ourselves, we excluded self-citations throughout the analyses and used a 3-year citation window. We identified HCRs on an annual basis for the period 2010 until 2018. We excluded more recent years to ensure the 3-year citation window for every paper; our citation data covers publications up to 2018. We restricted any selection and analysis to substantial items: article, letter, note, and review.[3] As field-classification

[2] The database is maintained by the Fraunhofer Institute for Systems and Innovation Research (ISI) and supported via the German Competence Network for Bibliometrics. The network is funded by the Federal Ministry of Education and Research (Grant: 16WIK2101A).

[3] We decided to use the four document types as they are substantial contributions by authors. We could have restricted our dataset on articles and reviews, but letters and notes are sometimes substantial papers and the number of letters and notes in the Scopus database is rather small. We did not include books or book chapters as these document types are biased towards certain countries and fields. In addition, we did not take conference proceedings into account – in the awareness of introducing a potential field bias with a discrimination against computer science and electrical engineering (Michels & Fu, 2013). Even after field-

schema, we used the Fields of Science and Technology (FOS)[4] suggested by the Organisation for Economic Co-operation and Development (OECD). Of the 42 fields listed in the revised version of the FOS, we were able to assign 35 to distinct fields in the Scopus classification system. The missing seven fields are miscellaneous fields in the main fields (natural sciences, engineering and technology, medical and health sciences, agricultural sciences, and social sciences). We assigned all journals in Scopus based on their All Science Journal Classification Codes (ASJC)[5] to the 35 fields.

We did not introduce any further restrictions to the population of authors and therefore kept authors from any country and any institutional type in the dataset. We calculated all indicators for all countries and all institutions. When we compared institutional effects on the outcomes (HCRs lists), we restricted the analyses to the German research landscape. We have a consolidated affiliation disambiguation at the German organizational level in our in-house database. Since we have considerable background knowledge of the institutional landscape in Germany, we were in the best position to interpret the findings.

There is a methodological difference between our and Clarivate's field definition approach. Clarivate uses its Essential Science Indicators (ESI) classification of 22 fields.[6] ESI assigns each journal to exactly one field. We could not find any description of the assignment procedure by Clarivate.[7] Instead of forcing each journal—and all papers in that particular journal—into exactly one field, we fractionally assigned each journal (which belongs to more than one field) to the corresponding fields.

Five approaches for identifying highly cited researchers

In this study, we compared five approaches (variants) for identifying HCRs:

Variant-0 represents the approximation of Clarivate's approach. This means, we restricted the identification of HCRs to the authors who have published the top-1% of highly cited papers per field. However, we decided to adapt the implementation of the methods employed by Clarivate in some parts, mainly to make them comparable to the implementations of the other approaches. Different to Clarivate, (1) we employed a 3-year citation window to give older and younger publications the same chance of making it into the list, (2) we excluded self-citations, (3) we used the FoS classification, and (4) we allowed for multiple fields per publication instead of forcing it into only one category (see the previous section). This procedure resulted in a list of slightly more than 1% of papers in Scopus, due to the different sizes of the individual fields. Similar to Clarivate, we excluded

Footnote 3 (Continued)

and document type-specific normalization some differences between the citation impact of conference proceedings and of substantial contributions might remain. The difference possibly result – among other factors – from a time difference between the conference event and the publication of the proceedings. As our paper intends to show structural effects of different definitions of HCRs, the results of our analyses should not be affected by the inclusion or exclusion of certain document types. All lists of HCRs would differ in a similar way depending on publication sets including or excluding conference proceedings.

4 see https://www.oecd.org/science/inno/38235147.pdf. The assignment of the classes to the FOS can be requested from the corresponding author.

5 see https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/

6 see http://esi.help.clarivate.com/Content/scope-notes.htm

7 see http://esi.help.clarivate.com/Content/scope-coverage.htm

papers with more than 30 affiliations,[8] which reduced the set of papers for the period 2010 to 2018 by 1455 (= 0.7%) to 218,638.

For each of the 35 fields, we calculated the square root of the number of authors and took the number of publications of that particular author in the sorted (descending) list as the threshold per field.[9] Cross-field HCRs were taken into account by summing the fractions of the number of papers divided by the field-specific threshold across all fields an author is active in. If the sum was larger than (or equal to) one, an author entered the HCRs list.

Variant-0 led to a total HCRs list of 14,349 authors for the period 2010 to 2018. This number equals a share of 0.07% of all distinct authors in the period. 5,342 authors entered the list via the field-specific calculations and 9,007 via the cross-field calculations.[10]

The second approach (Variant-1) that we prepared for our comparison follows a slightly different procedure than the approach by Clarivate: instead of searching for HCRs only in the set of the 1% highly cited papers, we calculated the absolute citation numbers of each author for each publication year and field, based on all publications in the database. This approach only focuses on the number of citations, independent of the number of published papers. For example, an author with 10 papers, each cited 10 times, got the same weight as an author with only one paper that was cited 100 times. While we started with about 1 million authors in Variant-0 of which we selected the HCRs, Variant-1 started from all about 20 million authors in the period 2010 to 2018. The reasoning behind the use of this approach for identifying HCRs is the emphasis of high individual visibility: Variant-1 interprets the term "highly cited" as most often cited in absolute and not in relative terms (i.e., per paper published as in Variant-0).

In each year of our observation period, all authors in a field with at least the same number of citations like the one at the 1%-margin joined the HCRs list. This procedure led to slightly higher shares of HCRs than 1%, namely (1) between 1.5% and 1.6% of the author population in each individual year, and (2) 2.1% of all authors in the observation period from 2010 to 2018. This "full" list of HCRs included 318,913 authors. This is a substantially higher number of HCRs than we received with Variant-0.

For the comparison of Variant-1 with the other approaches in this study, we took the full list. Further additional criteria could be used in future studies to possibly "optimize" the HCR selection. For example, "constant producers" could be identified by taking only authors into account, who are present, e.g., in two consecutive years or in two out of five years in a particular period. The advantage of Variant-1 as used in this study is, however, that the total impact of an author is seen as relevant for being a HCR (without considering how this impact has been arisen). An important disadvantage of Variant-1 may be that highly productive authors with papers that have gathered considerable, but not necessarily very high citation numbers have good chances to enter the HCRs list: the many citations are the result of high productivity. Another disadvantage of Variant-1 may be that each

[8] Our analyses build on the methods published in 2021.

[9] Clarivate includes all authors with one publication less than the threshold number.

[10] Clarivate identified 6,400 HCRs for the 2020 list in the WoS: 3,900 HCRs entered the list via the field-specific calculations, and 2,500 via the cross-field calculations. The reason for the higher author numbers identified in this study (compared to Clarivate) may be the larger number of publications in Scopus than in the WoS. The reason for the higher share of cross-field entries probably lies in the fact that Clarivate assigns each journal exactly to one field. We employed fractional assignments to each of the 35 FOS (see footnote 4).

author in large author groups accounts for the full set of citations. In consequence, citations are counted multiple times (among co-authors). This disadvantage led us to Variant-2.

The third approach—Variant-2—is identical to Variant-1. The only difference is that we applied fractional instead of full counting of citations according to the number of authors. In other words, all authors of a paper share the number of received citations. The top-1% of authors per field entered the list: any author with the same number of fractional citations was considered above the 1%-margin (including the authors at the 1%-margin). Variant-2 led to a slightly lower HCRs number than Variant-1: 313,475 authors represent 2% of all authors in the period from 2010 to 2018. On a year-by-year basis, we found an almost constant rate of 1.5% among the authors, increasing from about 46,000 to 77,000 between 2010 and 2018. The advantage of Variant-2 (compared to Variant-1) is that each citation is counted only once. Its disadvantages may be that (1) all authors are treated equally in any group of authors, and (2) highly cited publications with large numbers of authors counted with low citation impact. Especially the first disadvantage led to our third variation—Variant-3.

Variant-3—the fourth approach—could be called "the corresponding author takes it all" variant. The variant is based on the fact that the correspondence author is the main contributor to a paper (and most of the co-authors have only limited contributions). To identify HCRs, we took the top-1% of corresponding authors in terms of total citation counts. Total citation counts is (again) defined as the number of citations that all papers of a particular author receive within a 3-year citation window. The advantages of the focus on the corresponding author are obvious: all citations and all publications are counted only for one (the main) author. The assumption that the corresponding author makes a considerable contribution to a publication is evident (de Moya-Anegón et al., 2013). The most important disadvantage is the ignorance of any other authors than the corresponding author and the differences in disciplinary or institutional habits of corresponding authorships. In some disciplines and institutional settings, the role of the corresponding author is taken over by junior researchers, who have—as a matter of fact—less publications than senior researchers and therefore have lower probabilities to enter the HCRs lists in some of the variants, especially Variant-0. In addition, in some disciplines and institutional settings it is not the main researcher or contributor who acts as the corresponding author, but the most senior researcher. In some countries, e.g., China, first and/or corresponding authors are particularly remunerated, so that they have a material interest to take over this role in international collaborations.

In total, the Variant-3 approach for identifying HCR led to 94,415 (corresponding) authors in the period from 2010 to 2018. This number represents 2% of all corresponding authors in the considered period and 0.6% of all authors.

The fifth approach (Variant-4) is based on the list of authors from the standardized citation indicators database provided by Ioannidis and co-authors (Ioannidis et al., 2016, 2019, 2020). The first version of the list was published in 2016. Updates have been published in 2019 and in 2020 including the top 100,000[11] authors. Ioannidis and co-authors (Ioannidis et al., 2016, 2019, 2020) used the following indicators: 1) citation score (field-normalized), 2) h-index, 3) adjusted h-index (Schreiber's h-index), 4) citations to single-authored papers, 5) single- or first-authored papers, 6) citations to single-, first- or last-authored papers.

[11] In the 2016 version, the authors provided a list of top 30,000 authors. Since 2019, the list has been extended to the top 100,000 authors. We are very grateful to Jeroen Baas, who provided us with a file of authors generated in August 2021 that also included Scopus author IDs.

These six indicators were used for the selection of top authors. In addition, they provided an equally weighted composite index of normalized (0, 1; based on the maximum value of each indicator) scores after log-transformation. In their 2020 paper, the authors offer rankings of top 100,000 authors (considering citation impact with and without self-citations) as well as career-long citations versus single-year-citations (citations in Scopus received in 2019) in 22 fields and 176 sub-fields (of the Scopus ASJC). Of the different lists from Ioannidis et al. (2020), we decided to employ the career-long list without self-citations (since it corresponds best to the lists based on the other variants for identifying HCRs). We selected the top 100,000 authors of this list.[12] However, to make it comparable to the other variants applied in this study, we focused on authors for whom the last year of publication was 2010 or later. This left us with a list of 94,364 authors, who account for about 0.6%[13] of all (worldwide) authors publishing in the period from 2010 to 2018.

Variables used

The first part in the following results section focuses on the overlaps in authors between the variants: do the variants identify similar or different authors as HCR?

The second part deals with the characteristics of the HCRs resulting from the different approaches: (1) average scientific age of HCRs, (2) their gender, (3) institutional affiliation, and (4) the average team size. We are interested in structural similarities and differences resulting from the different variants:

(1) Scientific age: As information on the real age of authors is not available in the data, we resort to the difference between the first appearance of an author in the database and 2022. We call this difference the “scientific age” of authors.

(2) Gender: We assigned the binary gender (male, female) to the HCRs based on the first names with the help of country-specific names lists (see Michael, 2007). However, in these lists some names are not covered (missings) or are ambiguous with respect to the gender (e.g., Alex, Kim or Sasha) and are therefore coded as “ambiguous”. In the gender analyses, we only refer to those persons with names that are unambiguous and can be assigned to one of the sexes. While the overall population of authors that entered the calculation of the five HCRs variants was not restricted to certain countries (we did not use a subset), we used a subset for the gender analyses of HCRs. We took only papers from countries that have a male/female-coverage of at least 90% of the papers in the period from 2010 to 2020. These countries have a high share of identified names and a low share of ambiguous gender information: AT, BE, BG, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IL, IS, IT, JP, LT, LU, LV, MT, MX, NL, NO, NZ, PL, PT, RO, SE, SI, SK, TR, US. We assume that missing and ambiguous names (from other countries) follow the same gender distribution as the one we are able to assign.

[12] Ioannidis and co-authors (Ioannidis et al., 2016, 2019, 2020) did not restrict their analyses to certain document types; they included – in contrast to our study – books, book chapters, and conference proceedings. However, we restricted our excerpt from their list of HCRs to authors who had published articles, letters, notes, or reviews in our observation period.

[13] The denominator is the number of distinct author IDs in Scopus in our observation period. Ioannidis and co-authors (Ioannidis et al., 2016, 2019, 2020) restricted their analyses to author profiles with at least 5 articles, reviews or conference proceedings to avoid arbitrary and false author profiles. Our calculations might therefore represent lower bounds of author shares than Ioannidis and co-authors (Ioannidis et al., 2016, 2019, 2020).

(3) Institutional affiliation: Since we have disambiguated[14] data on the institutional level in Germany, we additionally took a look on the institutional distribution of HCRs across the four large German non-university research organizations—Helmholtz Association, Fraunhofer Society, Leibniz Association, and Max Planck Society—as well as German universities as two groups: universities of applied sciences (univ of appld. sc.) and universities. The consideration of the German universities and organizations are particularly interesting, since they are differently aligned. Whereas the Fraunhofer Society is an applied research organization prioritizing relevant technologies and transferring its findings, the institutions in the Leibniz Association address problems that have societal and international relevance. The Max Planck Society is most active in basic research, and the Helmholtz Association develops and operates complex research infrastructures.

(4) Team size: We calculated the team size per author as the average of the number of co-authors per paper within the observation period. For the comparison of the five HCR variants, we took the average of this number per variant and publication year.

The third part in the following results section deals with the application of the different HCR variants in a certain context. We are interested in the question of how the different approaches are able to identify Nobel laureates. Do the different HCRs lists contain a similar number of laureates, and what are the reasons for possible different results? For the analyses, we identified Nobel laureates from the areas of physics, chemistry, medicine, and economics in the various HCRs lists. For this purpose, we identified the author profiles of all laureates since the year 1985.[15] We expected that the HCRs lists contain many Nobel laureates, since both HCRs and laureates can be denoted as exceptional scientists.

Results

Overlaps between the variants

We started our analyses by calculating the absolute numbers of HCRs and their shares among all authors. We found that the absolute numbers and shares significantly differ between the five variants. Variant-0 (Clarivate) led to the smallest set of HCRs with n = 14,349 and a share of 0.09% of all authors in the period from 2010 to 2018. Variant-1 resulted in the largest list of HCRs in absolute and relative terms (318,913 and 2%). This approach takes the total number of citations gathered by the papers of the top 1% authors per field and year into account. Variant-2 (fractional counting of citations) led to a slightly smaller list than Variant-1 with a number of 313,475 authors (2% of all authors in the period). The list of HCRs based on the corresponding authors approach (Variant-3)

[14] The disambiguation process is a semi-automatic and manual curation of German affiliations provided by the University of Bielefeld in the context of the German Competence Network for Bibliometrics funded by the Federal Ministry of Education and Research (Grant: 16WIK2101A) (Rimmert et al., 2017).

[15] We had to make a decision to reduce the efforts necessary for the manual identification of the Nobel laureates. We made a practical decision and selected the year 1985 as the earliest year of considered Nobel laureates (that we took into account). As our analyses are based on publications from the years 2010 to 2018, we set the threshold to 1985 based on the assumption that laureates before this year were hardly scientifically active during our observation period.

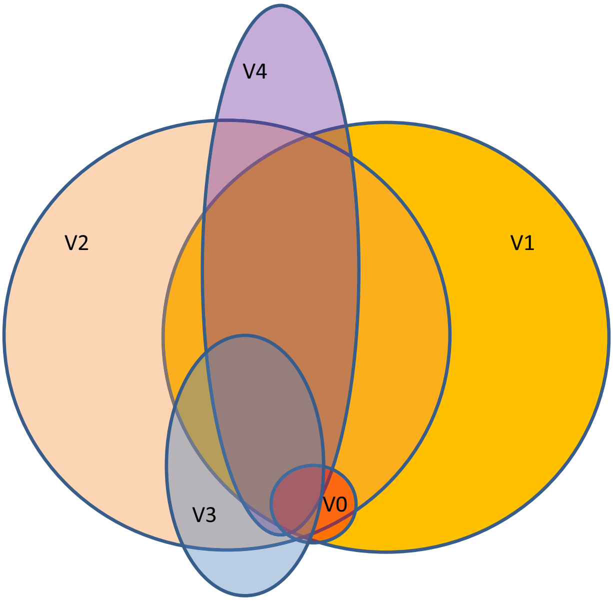

Fig. 1 Schematic representation of overlaps between HCRs lists from five approaches – Variant-0 (V0) to Variant-4 (V4) – to identify HCRs. *Source* Elsevier—Scopus; Fraunhofer ISI calculations and representation

contains more than 94,415 authors who represent 0.6% of all authors and 2% of all corresponding authors[16] in the observation period.

Figure 1 shows the overlaps of the five variants in terms of HCRs coverage. The bilateral overlap and bilateral exclusivity are depicted. It becomes evident that almost all authors in the list of Variant-0 are also part of the lists from Variant-1 (98%) and Variant-2 (92.2%). The list from Variant-3, however, only partly overlaps with Variant-0 having 60.4% of Variant-0's authors in common or 9.2% of Variant-3's authors. In consequence, if we accept that corresponding authors are the researchers who contributed most or at least considerably to the papers, we have to determine that a large number and a relevant share of them is not represented by Variant-0. It seems that many authors profit from papers although they were not the main actors in the research process.

The overlap between the lists of Variant-1 and Variant-2 is only slightly above 50%. This means that full counting versus fractional counting of citations has significant influence on the composition of the HCRs list. Variant-3 (corresponding authors) is able to capture only about two-thirds of the HCRs identified by Variant-0 and Variant-1, but almost three-quarters of HCRs identified by Variant-2. It seems that fractionally counting of citations is able to identify corresponding authors. The HCRs list from Variant-4 (the list provided by Ioannidis and co-authors) seems to be the list with the greatest difference compared to the other four lists. There is only a small coverage of 43% with respect to Variant-0 authors, while—the other way around—the list resulting from Variant-0 only contains 6.6% of the list from Variant-4. 11% of Variant-1 authors are covered by Variant-4, and vice versa 48% of the authors in Variant-4 can also be found in the list resulting from Variant-1. Accordingly, the

[16] In the period from 2010 to 2018, we identified about 16 million distinct authors and 4.8 million distinct corresponding authors in total.

overlaps between Variant-2 and Variant-4 are 14% and 61%, respectively. Variant-3 and Variant-4 overlap in 23% and 29% of the cases, respectively.

Characteristics of highly cited researchers

Institutional affiliation

What already becomes obvious at this stage of our comparison is that Variant-0 is highly selective. The other insight is that the list from Variant-4 is clearly distinct from the lists of the other variants. We expect that these similarities and differences also have an influence on the characteristics of the HCRs in the different lists. Table 1 shows the shares of the HCRs by German research organizations and universities. The shares are calculated as the number of HCRs over the total number of authors of the particular organization in a particular year. As can be seen in Table 1 for the year 2018, the Max Planck Society ranks first in all variants' lists hosting relatively the most HCRs in Germany. This result is expectable, since the organization can be denoted as successful in basic research (e.g., the organization has received many Nobel prizes in the last years). However, the Leibniz Association and Helmholtz Association compete for the second rank, depending on the HCRs list we focus on. Using Variant-0 and Variant-3, the Leibniz Association takes the second rank, while using the other variants, the Helmholtz Association takes the second rank.

In the case of full counting of citations as adopted in Variant-1, the Helmholtz Association performs better than the other organizations. This seems to be due to the fact that the Helmholtz Association has established a rather intensive collaboration network, especially with universities (Frietsch et al., 2022; Frietsch et al., 2016). Many of the publications by the Helmholtz Association are in collaboration with authors from other organizations so that the full counting approach is favoring the Helmholtz Association. Variant-4 also puts the Helmholtz Association in the second rank and the Leibniz Association in the third rank. Different to the approach based on full counting of citations (Variant-1), however, it does not exaggerate the Helmholtz Association that much and keeps the relation to the Max Planck Society at a similar level like the other variants.

Although the ranks of the other organizations (except for the Helmholtz Association and Leibniz Association) are relatively consistent over the five variants (as displayed in Table 1), the relative distances between the organizations differs significantly: the index in Fig. 2 shows the relative positions of the six German organizations or groups of organizations (in case of universities and univ of appld. sc.) to the Max Planck Society for the

Table 1 Shares (in %) of HCRs by German research organizations and universities in 2018

	Variant-0	Variant-1	Variant-2	Variant-3	Variant-4
Max Planck Society	1.2	5.1	4.8	3.7	5.3
Leibniz Association	0.8	2.8	2.6	2.0	3.1
Helmholtz Association	0.6	4.0	2.9	2.4	3.5
Universities	0.4	2.4	2.2	1.6	2.9
Fraunhofer Society	0.2	1.0	1.3	1.1	1.4
Univ of appld. sc	0.2	0.5	0.8	0.9	0.5

Source Elsevier—Scopus; Fraunhofer ISI calculations

HCRs in 2018. The index has been calculated as the individual value per organization (or group of organization) divided by the value of the Max Planck Society, which acts as the benchmark in this exercise. We calculated the index to better demonstrate the similarities of and differences between the organizations and universities.

Variant-1 and Variant 2 relatively disfavor the Leibniz Association, whereas Variant-0 disfavors the Helmholtz Association, when comparing all five variants. The same holds for the universities, which reach lower shares of HCRs at Variant-0. Next to Variant-0, Variant-1 is most unfavorable for the Fraunhofer Society. For the universities of applied sciences, Variant-1 and Variant-4 are unfavorable.

The various analyses of the HCRs on the organizational level show that the institutional characteristics of the HCRs lists are different – if we take a closer look at German institutions. We find rather huge differences and impacts of the method on the outcome.

Scientific age

The shares of four different scientific age groups for all authors in the Scopus database (between the years 2010 and 2018) as well as for the five variants is depicted in Table 2. It is evident that authors in the younger age groups are strongly underrepresented in all variants of HCRs under examination here. In Variant-4, almost all authors (99%) belong to the oldest age group of more than 16 years in the science system. Clarivate's Variant-0 as well as Variant-3 (corresponding authors) reach levels of 70% of the oldest age group, while Variant-1 (full counting) and Variant-2 (fractional counting) contain larger shares of authors who belong to younger age groups.

Table 3 shows the average scientific age of authors in the period 2010–2018 broken down by scientific field. The first column contains the shares of the age groups for all authors in Scopus. All five variants reach values well above the average age of all authors

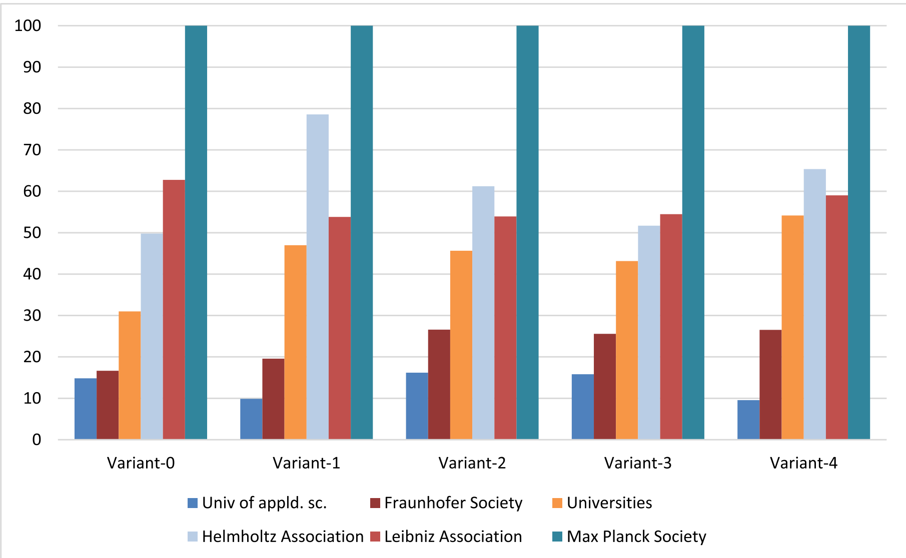

Fig. 2 Index (Max Planck Society = 100) with shares of HCRs by German organizations and universities, 2018. *Source* Elsevier—Scopus; Fraunhofer ISI calculations

Table 2 Shares (in %) of HCRs by scientific age groups, 2010–2018

Scientific age	Total	Variant-0	Variant-1	Variant-2	Variant-3	Variant-4
< = 5 years	18	1	4	3	1	0
6 to 10 years	41	11	17	16	9	0
11 to 15 years	22	17	21	20	19	1
> = 16 years	19	70	58	61	70	99
Total	100	100	100	100	100	100

Source Elsevier—Scopus; Fraunhofer ISI calculations

in the database in this period. This means that HCRs are—on average—more experienced scientists than the average. This is an obvious result, since usually (substantially) more than one paper is needed to become a HCR. What can also be seen is that the average age of the HCRs from Variant-4 is highest among all variants, followed by Variant-0. It seems that Clarivate fails to meet their aim: "recognizing early and mid-career as well as senior researchers is one of our goals in generating Highly Cited Researchers lists" (Clarivate Analytics, 2021, p. 18). The average of 19.6 years is rather high and only HCRs from Variant-3 (corresponding authors) come close to a similar average age. Variant-1 and Variant-2 identify authors that are on average 2 to 3 years younger than the HCRs from Variant-0. This holds not only true for all authors, but also for the individual fields of science.

In three fields (biological sciences, chemical engineering, and material engineering) the average age is considerable lower (more than 1 year lower) for HCRs from Variant-0 than for HCRs from Variant-3. HCRs from Variant-4, however, are well beyond even this level and reach a total average of 25.4 years. The explanation for this may be the selection of career-based lists of Ioannidis et al. (2020), who used total citation counts without considering a citation window. As a matter of fact, this favors highly productive and especially longer active authors. The longer time back an author started their career, the more papers they will have produced that will have had more time to collect citations. The other variants that we employed are based on a 3-year citation window. The window compensates for the long-term effects of careers and gives younger authors equal chances as older ones.

We also took a look at the aggregated level of some countries; we wanted to know whether there are country-specific variations in the age distribution. When looking at four countries (see Table 4), we find the same pattern for Germany, the USA, China, and the UK. Although China's authors are on average much younger than the authors from the other countries, Variant-4 identifies the oldest HCRs among the Chinese authors compared to the other variants. The USA has—on average—slightly younger HCRs than Germany in the case of Variant-0, while their HCRs are about one year older than the German authors in Variant-3. In effect, when using a highly selective list of HCRs in the 1% most highly cited papers (the Clarivate approach) the method selects much younger researchers in case of the USA, while there are older researchers when the corresponding authors are selected. The reason might be: on average, the USA has a much higher average citation rate than Germany. It seems that in the very selective approach of Clarivate this has a structural influence on the age: they receive relatively early in their career a level of citations that puts them at the top notch. The level of citations is more balanced when a broader set is taken into account. Based on Variant-1, Variant-2, and Variant-4, the HCRs from the USA and Germany are rather similar.

Table 3 Average scientific age of the HCRs resulting from the five variants as well as of all authors (total) in Scopus broken down by field, 2010–2018

	Field	Total	V0	V1	V2	V3	V4
	Total	10.8	19.6	17.4	18.0	19.4	25.4
Natural sciences	Mathematics	13.7	20.6	17.6	17.8	18.0	25.1
	Computer and information sciences	12.9	20.5	17.4	17.4	17.3	25.1
	Physical Sciences	12.9	19.5	17.5	18.1	19.2	25.3
	Chemical Sciences	12.3	19.5	16.6	17.2	20.4	25.4
	Earth and related environmental sciences	13.0	21.0	18.8	19.8	19.1	25.2
	Biological sciences	12.5	20.2	18.2	18.8	21.5	25.4
Engineering and technology	Civil engineering	12.3	19.6	16.9	16.9	17.4	24.7
	Electrical engineering., information engineering	12.5	18.5	17.2	17.5	17.9	24.9
	Mechanical engineering	12.2	18.4	17.0	16.8	18.9	25.1
	Chemical engineering	12.4	18.4	16.7	17.1	19.6	25.1
	Materials engineering	12.1	18.4	16.3	17.0	19.8	25.2
	Environmental engineering	12.8	19.7	17.9	17.6	18.9	25.2
	Environmental biotechnology	13.5	18.8	16.9	17.1	19.8	25.0
	Other engineering and technologies	12.6	18.9	16.4	16.7	18.9	25.1
Health	Basic medicine	12.4	21.3	17.5	19.9	21.9	25.5
	Clinical medicine	11.9	21.9	20.4	20.3	21.8	25.5
	Health sciences	12.9	22.1	19.1	18.9	19.5	25.4
Agric. sciences	Agriculture, forestry, and fisheries	13.4	21.9	18.8	19.0	19.9	25.4
	Animal and dairy science	13.5	22.1	19.5	19.6	19.9	25.5
	Veterinary science	13.2	22.2	18.9	19.5	20.5	25.5
Social sciences	Psychology	13.3	21.2	18.9	18.8	18.6	25.2
	Economics and business	12.3	20.2	17.4	17.3	17.6	24.8
	Educational sciences	12.7	21.2	17.4	17.2	17.2	25.1
	Sociology	13.0	20.7	18.6	17.9	18.2	25.0
	Law	12.7	20.4	17.4	17.2	17.5	24.9
	Political science	13.2	20.6	17.7	17.7	18.2	24.9
	Social and economic geography	13.1	20.6	17.8	17.3	17.6	24.9
	Media and communications	13.0	19.6	16.2	16.2	16.2	24.7
Humanities	History and archaeology	13.2	22.2	19.0	18.2	18.9	25.2
	Language and literature	12.2	20.8	18.0	17.7	17.7	25.0
	Philosophy, ethics and religion	12.5	21.0	16.9	17.1	17.3	25.0
	Arts	13.4	21.5	17.6	17.4	17.9	25.1
	Other humanities	13.5	20.3	17.4	17.3	17.5	24.8
	Multidisciplinary	14.3	20.8	17.3	17.5	21.5	25.4

Source Elsevier—Scopus; Fraunhofer ISI calculations

Table 4 Average scientific age for HCRs from the five variants for all authors and broken down by four frequently* publishing countries, 2010–2018

	Total	Variant-0	Variant-1	Variant-2	Variant-3	Variant-4
Germany	12.2	21.1	18.6	18.9	19.6	25.3
USA	12.4	20.6	18.4	18.7	20.4	25.4
China	8.8	17.3	14.6	15.1	17.6	24.7
UK	12.5	21.4	19.0	19.1	19.8	25.3

*defined as the largest publishing countries in absolute terms

Source Elsevier—Scopus; Fraunhofer ISI calculations

Gender

Figure 3 shows the share[17] of females among the HCRs from the five variants as well as the (Scopus) worldwide total. All variants reach shares of female authors well below the worldwide shares of females. The results point out that the HCRs lists are male dominated. However, there are large differences in the representation of female authors between the five variants under consideration. The lowest share of females is covered by Variant-4—the dataset from Ioannidis et al. (2020). It reaches an almost constant share of about 15%. Variant-0—the approximation of the Clarivate approach—reaches the second lowest shares at a rate of about 20%.

The other three variants perform better in representing female authors and exhibit increasing trends over time, which also reflect the overall trend of female authorship in journal publications. Variant 1 reaches the highest shares among the variants; the variant that uses full counting. The comparison of the HCR variants shows that women are less often corresponding author than their male collaborators.

The five variants also perform very differently with respect to the female representation in different fields (the results are not shown here). Female authors are relatively better represented in engineering fields and in social sciences. In most natural sciences, the female representation in HCRs lists is much lower, compared to the share of females among all authors in the corresponding field.

Team size

Co-publications (international) receive (on average) higher citation rates than publications without collaborations (Michels & Schmoch, 2014; Puuska et al., 2014; Schmoch & Schubert, 2008; Thelwall & Maflahi, 2020). Since for most HCR variants which we compare in this paper the contribution of each author is either fully or fractionally taken into account, the team size—defined as the average number of authors across all papers of an author – may be connected to the selection of HCRs.

[17] The data underlying the gender analyses is restricted to the following countries where a male/female-coverage of research papers exceeds 90% in the period from 2010 to 2020: AT, BE, BG, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IL, IS, IT, JP, LT, LU, LV, MT, MX, NL, NO, NZ, PL, PT, RO, SE, SI, SK, TR, US.

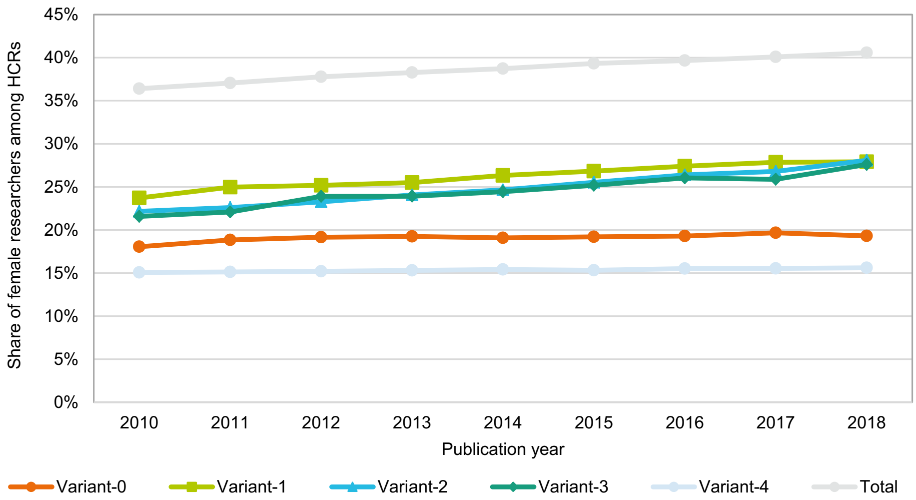

Fig. 3 Share of females among the HCRs from the five variants and all authors, 2010–2018. *Source* Elsevier—Scopus; Fraunhofer ISI calculations

Figure 4 shows the average[18] number of co-authors of the HCRs in each publication year for each HCRs variant. The team sizes increase over time for the HCRs identified by the five variants. Fractional counting (Variant-2) of citations per author leads to the lowest average number of co-authors among the HCRs. In other words, the inclusion in this HCRs list is least dependent on the team size compared to the other variants. Variant-3 as well as Variant-4 is based on similar team size levels, while Variant-0 and Variant-1 exhibit the highest average number of co-authors of the HCRs. The growth is steepest for Variant-0, which almost catches up with Variant-1 in the most recent year of our observation period. However, Variant-1 stays ahead of all variants, indicating that in case of full counting the impact of co-authorship plays an important role for the selection of authors for the HCR list.

The ability of different variants to identify Nobel laureates among the highly cited researchers

This section deals with the application of the different approaches to identify HCRs in a certain context: We are interested how the different approaches are able to identify Nobel laureates. Do the different HCRs lists contain a similar number of laureates, and what are the reasons for possibly different results? Fig. 5 displays the shares of identified Nobel laureates in the HCRs lists who have actively published in a particular year. It is evident that Variant-0 is able to identify only about 10% to 13% in the period from 2010 to 2018, while the four other variants are able to cover between 61% and 92% of all publishing Nobel laureates. Variant-4 (the HCR list from Ioannidis and co-authors) is performing best in terms

[18] As the distribution of the number of authors might be skewed, we also calculated the median number of authors per HCR variant. The results and conclusions are very similar using both methods. One reason for the similarity may be that we excluded papers with more than 30 authors from the HCR identification in general.

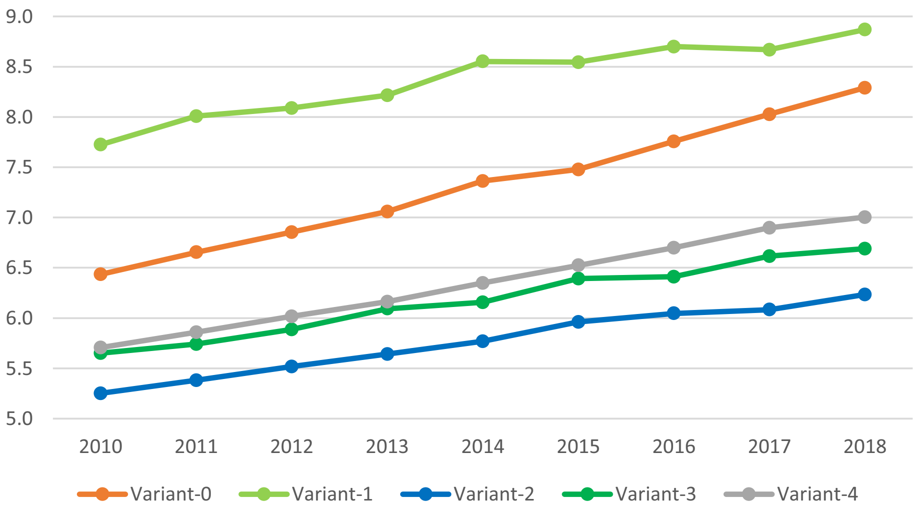

Fig. 4 Average number of co-authors of the HCRs. *Source* Elsevier—Scopus; Fraunhofer ISI calculations

of Nobel laureates' coverage. It reaches a constant share of all publishing Nobel laureates per year of about 90%. Variant-2 (fractional counting of total citations) is performing second best in this respect (76% to 87%).

The results in Fig. 6 show that the total coverage of HCRs from the five variants is considerably different, with Variant-0 being most selective. This explains the low absolute numbers and therefore low shares of laureates in the HCR lists as such. In another analysis, therefore, we took the absolute numbers into account and calculated the share of Nobel laureates among all selected authors (by a variant) in relation to the share of Nobel laureates in all authors in the database. As the results in Fig. 6 indicate, the representation of laureates is poorest with respect to Variant-0. The probability of a Nobel laureate to be in the list of HCRs is only between 32 and 51 times higher than for any other author to be selected. In case of Variant-3, which performs best in this respect, the same calculations

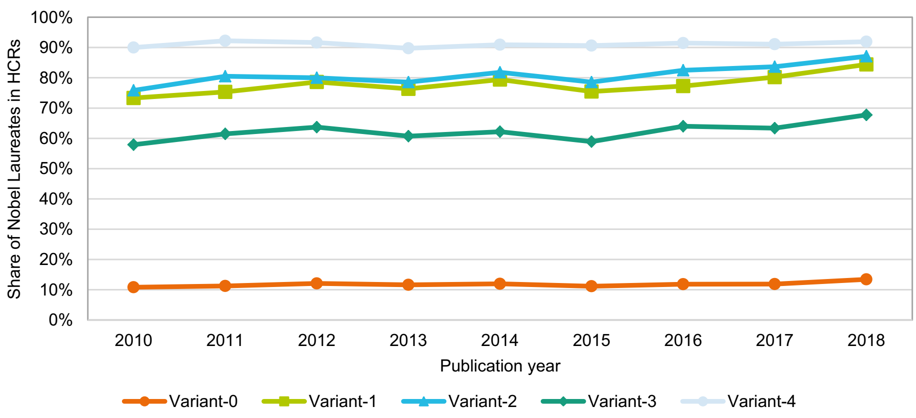

Fig. 5 Share of publishing Nobel laureates identified by each variant. *Source* Elsevier—Scopus; Fraunhofer ISI calculations

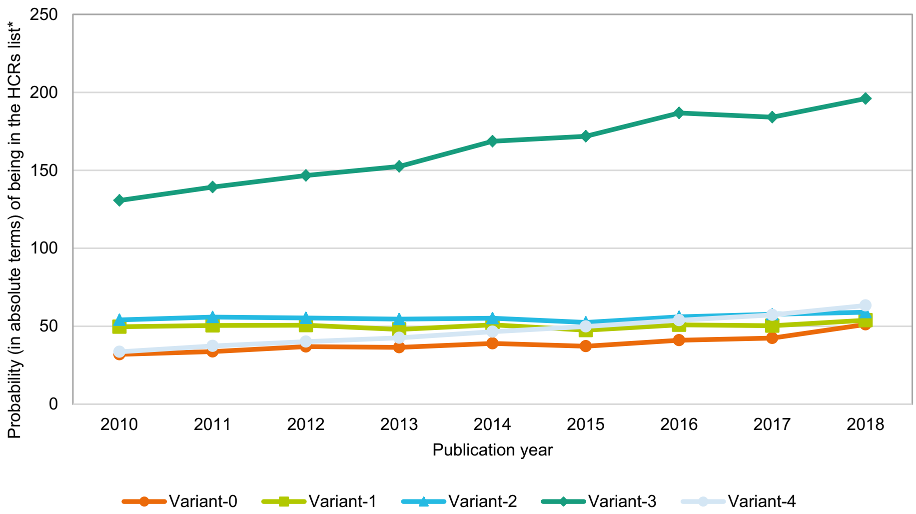

Fig. 6 Probability of covering Nobel laureates in the HCRs lists produced by each variant**.** *The probability is calculated as the share of Nobel laureates in the particular variant over the share of Nobel laureates in the total database in each particular year. *Source* Elsevier—Scopus; Fraunhofer ISI calculations

reveal a probability that is between 131 and 196 times higher than for any author to be selected by random sampling. The other three variants are only slightly above Variant-0 and reach probabilities that vary between 34 and 63 times higher than the average over the period from 2010 to 2018.

The empirical results show that the five variants are differently able to identify Nobel laureates. Since the empirical analyses in this section were intended to investigate the effect of using different variants in a certain evaluation context on the results, the results on Nobel laureates show that the variant does indeed matter. The results are not independent from the methods used. If one is interested, e.g., to identify laureates with bibliometric indicators, another variant than Variant-0 should be used.

Discussion

Although a great number of researchers is involved in research activities, the (great) contributions result from only a small part of their shoulders. Price's law (Price, 1963) implies that only about 25% of researchers are responsible for about 75% of publications (van Raan, 2019). One important reason for the unequal distribution among the shoulders might be the high number of temporary researchers in the system. Only a small part of researchers spend their full career in science (Milojevic et al., 2018). Amara et al. (2015) did not only differentiate researchers with respect to their publication output performance, but also with respect to their citation impact performance. The authors found unequal performance distributions by differentiating between (i) non-publishing researchers, (ii) low performing researchers, (iii) frequently publishing researchers, (iv) frequently cited researchers, and (v) researchers with many highly cited papers. The study found that low and high performers (in terms of publication and citation counts) differ as follows: high performers are "full professors, they dedicate more time to their research activities, they receive all their research funding from research councils, and, finally, they are located in top tier universities" (Amara et al., 2015, p. 489).

Although much of the publication output and citation impact in science depends on the few high performers and they are in the focus of interest within science and beyond (e.g., for reviewing of large grant applications or expert advices in politics), it is frequently not clear how they can be identified. For example, Li et al. (2019) defined a "top scientist" as someone who "belongs to the top 5% of cited authors in her discipline for that same year". One of the most prominent definitions of high performing researchers has been introduced by Clarivate: it defines HCRs as those who have published the most papers that belong to the 1% most frequently cited papers in a field. Although a standard method does not exist in bibliometrics to identify HCRs, the results by Docampo and Cram (2019) show that "the use of HCR lists [published by Clarivate] as a basis for recruitment can be a legitimate element of sound academic policy. HCRs are by and large talented and there are many reasons for a university to attract them beyond the pursuit of higher ranking" (p. 1022). The use of HCRs lists in recruitments can be questioned, if different plausible methods exist, and the methods come to different results.

We have demonstrated in this study that HCRs can be differently identified: there is no standard definition of HCRs. Depending on the methods used and the perspectives taken, very different HCRs lists can be expected—as the empirical results of our study demonstrate. HCRs can be identified by considering field-normalized citation rates or absolute numbers of citations; inclusion or exclusion of self-citations; full counting or fractional counting of publications; all authors, only corresponding authors or only first authors; short, long or varying citation windows; and short or long publication periods. As we demonstrated with five different variants of defining HCRs, the selection among these options has an influence on the sample of selected researchers and their characteristics that are thereby defined as highly cited. Some options have a stronger influence on the outcome than other options such as the length of the citation window or the focus on all authors versus only the corresponding author. The user of HCRs lists should always be aware of the influence these options have on the final lists of researchers. For example, if the lists are used for recruitments, one should be aware of the dependence of the HCRs list from the method for identifying the HCRs. For recruitments, especially the HCRs list that is only based on corresponding authors may be relevant, since the list focuses on the main actors in a particular research project. This recommendation is supported by the influence of the number of co-authors on the inclusion in HCRs lists.

The coverage of the variants significantly differs. While the approach suggested by Clarivate identifies only about 0.07% of all authors as highly cited researchers, the variants that refer to the absolute number of citations a researcher received are able to mark about 2% of all authors within a 10 year period as being highly cited. The other variants examined here—focusing on corresponding authors and the suggestion by Ioannidis and co-authors—reached levels in between. In this study, we not only proposed several variants for identifying HCRs, we also compared the outcomes of these variants. Structural analyses on the representation of German organizations and universities, (scientific) age-groups, female researchers, and team size revealed strong differences between the variants examined here.

The ranking and the relation between German institutions is altered, depending on the definition of HCRs. The analysis of the scientific age and the distribution of scientific age groups revealed that all five variants under examination in this study select—on average—older researchers. The variant provided by Ioannidis and co-authors almost only selected researchers that have been in the system for at least 15 years and therefore belong to the oldest age group. Variant-0 also resulted in a HCRs list that contains a large share of oldest (more than 15 years) researchers.

In terms of female representation, all variants performed below the worldwide share of female authors in the database. The HCR lists are male dominated. One reason for this result could be the age structure, as the HCR lists are biased towards older authors. Hardly any country reached parity between men and women in terms of scientific output, but in many countries the participation rate of women and also their contribution to publications have considerably increased in recent years. This structural change cannot be captured by approaches that use open citation windows, resulting in a selection—on average—of older authors. Citation windows—especially if they are rather short—are better able to capture the short-term or recent changes in the science system. If the focus in the definition of HCRs is on long-term citation counts, open citation windows are more appropriate. However, the users of the lists need to be aware that short-term changes play a minor (statistical) role. Their interpretation should be rather long-term oriented and not on recent policy implications or structural trends.

The average number of co-authors varies significantly between the different HCR variants indicating that the selection of authors is dependent on the team size. One explanation may be that papers by larger author teams and especially internationally co-authored papers tend to receive higher citation counts. This effect may increase the inclusion probability of authors publishing in larger teams in HCR lists. Since the overlap between the HCR variants is rather small and the structural differences between the variants are large, our empirical findings particularly challenge the use of full counting methods in the selection of HCRs. These findings are in line with results by Aksnes and Aagaard (2021).

We applied the HCR variants in a certain context and investigated whether the various lists of HCRs contain Nobel laureates. The results by Li et al. (2020) show that "Nobel laureates were energetic producers from the outset, publishing almost twice as many papers as scientists in our comparison group … Yet, compared with this productivity difference, more impressive is the gap in impact. Indeed, the future laureates had a more than six fold increase over the comparison group in terms of the rate of publishing hit papers, defined as papers in the top 1%". Thus, one can expect that many Nobel laureates are among the identified HCRs. Our findings show that the different implementations of HCRs lead to very different representation of Nobel laureates in the identified sample. It is evident, however, that the larger the number of HCRs identified by a variant is, the larger is the number of Nobel laureates covered by the list. The extremes in this respect were the variants suggested by Clarivate (Variant-0), which only covered about 10% of the publishing laureates, in contrast to the approach suggested by Ioannidis and co-authors. This approach covered about 90% of the publishing laureates. Beyond the mere coverage, we were also able to show that the probability of a Nobel laureate to be represented in the list of HCRs is much higher in Variant-3, which focuses on corresponding authors only. In essence, a focus on corresponding authors is able to offer a selective set of authors, but at the same time probably capture highly reputable (and active) researchers.

The possibility of different HCRs definition variants and the analyses of their outcome characteristics clearly suggest two main insights (from our point of view). First, one needs to be very cautious with the use of HCRs lists because the definition for identifying the HCRs has a strong influence on the outcomes. The second point follows from the first: policy recommendations and evaluations of organizations or individuals based on HCRs lists, or even the use of HCRs lists as a decision-making tool in recruitment processes should be used with care and should be put into perspective. We demonstrated this by analyzing the number of Nobel laureates among the HCRs. To make reasonable and reliable contributions to any of these processes based on HCR lists, the variant should be selected target-oriented according to the evaluation criteria. For identifying future Nobel laureates, the HCRs definition should be used that has proven most reliable in identifying laureates in the past. For selecting active researchers in recruitment processes, the HCR approach based on corresponding authors seems to be most appropriate.

Acknowledgements The bibliometric data used in this paper are from an in-house database of Elsevier's Scopus maintained by the Fraunhofer Institute for Systems and Innovation Research ISI and supported via the German Competence Network for Bibliometrics funded by the Federal Ministry of Education and Research (Grant: 16WIK2101A).

Author contributions Conceptualization: LB, RF; Data curation: SG; Formal Analysis: SG; Investigation: RF, SG; Methodology: RF, SG; Project administration: RF; Visualization: RF, SG; Writing—original draft: LB, RF, SG; Writing—review & editing: LB, RF, SG.

Funding We have not received external funding.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

Abramo, G., D'Angelo, C. A., & Soldatenkova, A. (2017). How long do top scientists maintain their stardom? An analysis by region, gender and discipline: Evidence from Italy. *Scientometrics, 110*(2), 867–877. https://doi.org/10.1007/s11192-016-2193-x

Aksnes, D. W., & Aagaard, K. (2021). Lone geniuses or one among many? An explorative study of contemporary highly cited researchers. *Journal of Data and Information Science, 6*(2), 41–66. https://doi.org/10.2478/jdis-2021-0019

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics, 117*(2), 705–720. https://doi.org/10.1007/s11192-018-2895-3

Amara, N., Landry, R., & Halilem, N. (2015). What can university administrators do to increase the publication and citation scores of their faculty members? *Scientometrics, 103*(2), 489–530. https://doi.org/10.1007/s11192-015-1537-2

Baerlocher, M. O., Newton, M., Gautam, T., Tomlinson, G., & Detsky, A. S. (2007). The meaning of author order in medical research. *Journal of Investigative Medicine, 55*(4), 174–180. https://doi.org/10.2310/6650.2007.06044

Bauwens, L., Mion, G., & Thisse, J.-F. (2011). The resistible decline of European science. *Louvain Economic Review, 77*(4), 5–31.

Bornmann, L., & Bauer, J. (2015). Evaluation of the highly-cited researchers' database for a country: Proposals for meaningful analyses on the example of Germany. *Scientometrics, 105*(3), 1997–2003. https://doi.org/10.1007/s11192-015-1619-1

Bornmann, L., & Bauer, J. (2015). Which of the world's institutions employ the most highly cited researchers? An analysis of the data from highlycited.com. *Journal of the Association for Information Science and Technology, 66*(10), 2146–2148.

Bornmann, L., Bauer, J., & Haunschild, R. (2015). Distribution of women and men among highly cited scientists. *Journal of the Association for Information Science and Technology, 66*(12), 2715–2716. https://doi.org/10.1002/asi.23583

Butler, A., Xu, G., Musial, K. (2018). *Research performance reporting is fallacious.* Paper presented at the 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), Kaohsiung, Taiwan.

Clarivate. (2021). Highly cited researchers. Methodology. Overview. Retrieved February, 24, 2023, from https://recognition.webofscience.com/wos-op/awards/highly-cited/2021/methodology

Clarivate Analytics. (2021). *Highly cited researchers 2021*. Clarivate Analytics.

da Silva, J. A. T., & Bernes, S. (2018). Clarivate analytics: Continued omnia vanitas impact factor culture. *Science and Engineering Ethics, 24*(1), 291–297. https://doi.org/10.1007/s11948-017-9873-7

de Moya-Anegón, F., Guerrero-Bote, V., Bornmann, L., & Moed, H. (2013). The research guarantors of scientific papers and the output counting: A promising new approach. *Scientometrics, 97*(2), 421–434. https://doi.org/10.1007/s11192-013-1046-0

de Price, D. J. S. (1963). *Little science, big science*. Columbia University Press.

Diko, M. L. (2015). South African scholars make Thomson Reuters 'Highly cited researchers 2014.' *South African Journal of Science, 111* (9–10). https://doi.org/10.17159/sajs.2015/a0121C

Docampo, D., & Cram, L. (2019). Highly cited researchers: A moving target. *Scientometrics, 118*(3), 1011–1025. https://doi.org/10.1007/s11192-018-2993-2

Dong, Y., Ma, H., Shen, Z., Wang, K. (2017). *A century of science: Globalization of scientific collaborations, citations, and innovations.* Paper presented at the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, Association for Computing Machinery

Fox, C. W., Ritchey, J. P., & Paine, C. E. T. (2018). Patterns of authorship in ecology and evolution: First, last, and corresponding authorship vary with gender and geography. *Ecology and Evolution, 8*(23), 11492–11507. https://doi.org/10.1002/ece3.4584

Frietsch, R., Gruber, S., Blind, K., Neuhäusler, P. (2022). *Erfassung und Analyse bibliometrischer Indikatoren im Rahmen des Pakt-Monitorings zum Pakt für Forschung und Innovation IV, Studie beauftragt vom Bundesministerium für Bildung und Forschung (BMBF)*. Fraunhofer ISI

Frietsch, R., Gruber, S., Helmich, P., & Neuhäusler, P. (2016). *Analysis of bibliometric indicators as part of the Pact for Research and Innovation, phases I and II. Background report for the Federal Ministry for Education and Research (BMBF).* Institut für Forschungsinformation und Qualitätssicherung iFQ, Fraunhofer Institute for Systems and Innovation Research ISI, University Bielefeld, Institute for Interdisciplinary Studies of Science I^{2}SoS

Ioannidis, J. P. A., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLOS Biology*. https://doi.org/10.1371/journal.pbio.3000384

Ioannidis, J. P. A., Boyack, K. W., & Baas, J. (2020). Updated science-wide author databases of standardized citation indicators. *PLOS Biology*. https://doi.org/10.1371/journal.pbio.3000918

Ioannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PLOS One, 9*(7), e101698. https://doi.org/10.1371/journal.pone.0101698

Ioannidis, J. P. A., Klavans, R., & Boyack, K. W. (2016). Multiple citation indicators and their composite across scientific disciplines. *PLOS Biology, 14*(7), e1002501. https://doi.org/10.1371/journal.pbio.1002501

Kamala, N., Rahimi, F., & Talebi Bezim Abadi, A. (2022). Learning from retracted papers authored by the highly cited Iran-affiliated researchers: Revisiting research policies and a key message to clarivate analytics. *Science and Engineering Ethics, 28*, 18. https://doi.org/10.1007/s11948-022-00368-3

Klein, A.-M., & Kranke, N. (2023). Some thoughts on transparency of the data and analysis behind the highly cited researchers list. *Scientometrics, 128*, 6773–6780. https://doi.org/10.1007/s11192-023-04852-w

Li, J. T. (2016). What we learn from the shifts in highly cited data from 2001 to 2014? *Scientometrics, 108*(1), 57–82. https://doi.org/10.1007/s11192-016-1958-6

Li, J. T. (2018). On the advancement of highly cited research in China: An analysis of the highly cited database. *Plos One*. https://doi.org/10.1371/journal.pone.0196341

Li, J., Yin, Y., Fortunato, S., & Wang, D. (2020). Scientific elite revisited: Patterns of productivity, collaboration, authorship and impact. *Journal of the Royal Society Interface, 17*(165), 20200135. https://doi.org/10.1098/rsif.2020.0135

Li, W., Aste, T., Cacciolo, F., et al. (2019). Early coauthorship with top scientists predicts success in academic careers. *Nature Communications, 10*, 5170. https://doi.org/10.1038/s41467-019-13130-4

Martinez, M., & Sá, C. (2020). Highly cited in the south: International collaboration and research recognition among Brazil's highly cited researchers. *Journal of Studies in International Education, 24*(1), 39–58. https://doi.org/10.1177/1028315319888890

Michael, J. (2007). 40000 Namen. *c't*(7), 182.

Michels, C., Fu, J. (2013). *Systematic analysis of coverage and usage of conference proceedings in Web of Science*. Fraunhofer ISI.

Michels, C., & Schmoch, U. (2014). Impact of bibliometric studies on the publication behaviour of authors. *Scientometrics, 98*(1), 369–385. https://doi.org/10.1007/s11192-013-1015-7

Milojevic, S., Radicchi, F., & Walsh, J. P. (2018). Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences of the United States of America, 115*(50), 12616–12623. https://doi.org/10.1073/pnas.1800478115

Moed, H. F., Aisati, M., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics, 94*(3), 929–942. https://doi.org/10.1007/s11192-012-0783-9

Must, U. (2020). The highly cited researchers with researcher ID: Patterns of behavior through time. *Journal of Scientometric Research, 9*(2), 195–199. https://doi.org/10.5530/jscires.9.2.23

Nagane, H. S., Fukudome, Y., Maki, K. (2018, 17–20 June 2018). *An analysis of star scientists in Japan.* Paper presented at the 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), Stuttgart, Germany.

Puuska, H. M., Muhonen, R., & Leino, Y. (2014). International and domestic co-publishing and their citation impact in different disciplines. *Scientometrics, 98*(2), 823–839. https://doi.org/10.1007/s11192-013-1181-7

Rimmert, C., Schwechheimer, H., Winterhager, M. (2017). *Disambiguation of author addresses in bibliometric databases – technical report*. Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I^2SoS)

Rodríguez-Navarroa, A., Brito, R. (2021). Validated research assessment based on highly cited researchers. Retrieved September 13, 2021, from https://arxiv.org/abs/2109.01366

Schmoch, U., Michels, C., Neuhäusler, P., Schulze, N. (2011). *Performance and structures of the German science system. Studien zum deutschen Innovationssystem No. 9–2012*. Expertenkommission Forschung und Innovation (EFI).

Schmoch, U., & Schubert, T. (2008). Are international co-publications an indicator for quality of scientific research? *Scientometrics, 74*(3), 361–377.

Schwartz, C., Hernon, P., Bornmann, L., Daniel, H.-D., Hall, H., Pomerantz, J., & Thelwall, M. (2010). Some thoughts on peer review. *Library & Information Science Research, 32*(1), 13–15. https://doi.org/10.1016/j.lisr.2009.12.003

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628–638.

Shamsi, A. (2022). Gender of highly cited researchers focused on the cross-filed category. *Gaceta Sanitaria, 35*(5), 506–507. https://doi.org/10.1016/j.gaceta.2020.03.007

Sinay, L., Carter, R. W., & de Sinay, M. C. F. (2020). In the race for knowledge, is human capital the most essential element? *Humanities and Social Sciences Communications, 7*(1), 20. https://doi.org/10.1057/s41599-020-0521-5

Stahlschmidt, S., Stephen, D., Hinze, S. (2019). *Performance and structures of the German science system. Studien zum deutschen Innovationssystem No. 5–2019.* Expertenkommission Forschung und Innovation (EFI).

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

Thelwall, M., & Maflahi, N. (2020). Academic collaboration rates and citation associations vary substantially between countries and fields. *Journal of the Association for Information Science and Technology, 71*(8), 968–978. https://doi.org/10.1002/asi.24315

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author sequence and credit for contributions in multiauthored publications. *PLOS Biology, 5*(1), e18. https://doi.org/10.1371/journal.pbio.0050018

van Raan, A. F. J. (2019). Measuring science: Basic principles and application of advanced bibliometrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 237–280). Springer International Publishing.

Wang, D., & Barabási, A.-L. (2021). *The science of science*. Cambridge University Press.