European Journal of Clinical Investigation

ESCI

Volume 43, Issue 12
December 2013
Pages 1339-1365

Perspective | Free Access

A list of highly influential biomedical researchers, 1996–2011

Kevin W. Boyack, Richard Klavans, Aaron A. Sorensen, John P.A. Ioannidis

First published: 03 September 2013 | https://doi.org/10.1111/eci.12171 | Citations: 27

Metrics

Citations: 27

124

Details

Keywords

Bibliometrics biomedical researchers citation counts h-index scientific impact scientometrics

Publication History

Issue Online:
19 November 2013

Version of Record online:
21 October 2013

Accepted manuscript online:
03 September 2013

Manuscript accepted:
29 August 2013

Manuscript received:
29 August 2013

Abstract

We have generated a list of highly influential biomedical researchers based on Scopus citation data from the period 1996-2011. Of the 15,153,100 author identifiers in Scopus, approximately 1% (n=149,655) have an h-index >=20. Of those, we selected 532 authors who belonged to the 400 with highest total citation count (>=25,142 citations) and/or the 400 with highest h-index (>=76). Of those, we selected the top-400 living core biomedical researchers based on a normalized score combining total citations and h-index. Another 62 authors whose focus is outside biomedicine had a normalized score that was at least as high as the score of the 400th core biomedical researcher. We provide information on the profile of these most influential authors, including the most common Medical Subject Heading terms in their articles that are also specific to their work, most common journals where they publish, number of papers with over 100 citations that they have published as first/single, last, or middle authors, and impact score adjusted for authorship positions, given that crude citation indices and authorship positions are almost totally orthogonal. We also show for each researcher the distribution of their papers across 4 main levels (basic-to-applied) of research. We discuss technical issues, limitations and caveats, comparisons against other lists of highly-cited researchers, and potential uses of this resource.

Introduction

The world of modern science has become increasingly competitive in recent years. Literature-based metrics are playing a greater role in decision-making than in the past [1]. Many researchers are highly aware of their so-called status, and check diverse metrics related to the impact of their work on a regular basis. The differing dimensions of impact and reasons for citing are receiving renewed discussion and analysis [2-6]. Although citation counts and related metrics (e.g. journal impact factor, h-index for individual researchers) are typically considered as proxy for impact [7], the nature of that impact is rarely, if ever, specified.

It is within this context of creating a better understanding of impact(s) that we have created a list of highly influential biomedical researchers. Ranking of scientists is explicitly not the main purpose of this list. Rather, we wanted to identify a pool of researchers who have had sustained success in highly influential work and who would thus presumably have substantial insight into differing features that could be associated with high impact. This list is being used in an ongoing survey where highly cited researchers are asked about the features of their most-cited articles. However, the list may be of use for many other purposes, as we discuss below.

Method and construction of database

We created a list of 400 highly influential biomedical scientists using an XML copy of the entire Scopus database obtained from Elsevier in June 2012. Scopus data contain author identifiers for each individual researcher (http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authoridentifier). We used these Scopus author identifiers rather than attempting to solve the author identity or disambiguation problem [8] independently. Scopus author identifiers do suffer to some extent from the two main problems associated with the author identity problem – polysemy (multiple authors merged in a single identifier) and synonymy (multiple identifiers for a single author). Based on our experience, as many as 10% of prolific authors have more than one Scopus author identifier (unpublished observation). However, in the majority of these cases, the papers are split between one very large profile that is weighted towards older publications and one that is much smaller containing a few newer publications. Thus, few cases of synonymy have a large or deleterious effect on metrics. Polysemy occurs far less often, but is much more problematic from a metrics point of view because the works of multiple authors are counted together. Polysemy is most often associated with common names. Many of the cases are easy to identify due to an unnaturally large number of papers associated with an author identifier.

The method we used to identify highly influential researchers assumes that the Scopus author identifiers have been correctly assigned to individual papers, and that each Scopus author profile contains only papers authored by that researcher. The method used was as follows. For each of the 15 153 100 Scopus author identifiers:

1. The number of articles published between 1996–2011, along with all citations to those articles as of the end of 2011, was counted.
2. These articles and citations to them were used to calculate an h-index [9] as of the end of 2011.
3. Using a local copy of the PubMed database covering the same time period, and for which we had previously linked Scopus records with PubMed records [10], we determined the fraction of publications from (1) that also appear in PubMed.

Of the over 15 million author identifiers, 149 655 (corresponding to 1% of the total) had an h-index of at least 20. These authors were ranked by total citation counts and by h-index. The number of authors with an h-index of at least 30, 40, 50, 60, 70 and 80, was 45 752; 15 385; 5185; 1773; 717, and 281, respectively. We focused on those authors who were ranked in the top 400 by either h-index or by total citation count, that is, those that had either h-index of at least 76 or total citation count of at least 25 142. This resulted in a list of 532 authors. 268 authors were ranked in the top 400 using both metrics, while 264 were ranked in the top 400 using one metric or the other.

Sorting of the 532 researchers was accomplished by first normalizing their total citation counts and h-index values to the top such values (100 939 citations for Shizuo Akira; h-index of 156 for Walter Willett), and then averaging these two normalized values. There are many ways of measuring impact. Rather than choosing a single value upon which to sort, we chose to include both total citation counts and h-index in our process.

There are obviously highly influential researchers across all scientific fields. We chose to generate a focused list of 400 researchers who publish primarily in biomedicine. Upon inspection, we found that all highly cited researchers with at least 80% of their publications linked to PubMed could be clearly classified as core biomedical researchers. In addition, for those with PubMed linked fractions between 60–80%, roughly half could be considered as core biomedical researchers while the rest have a focus outside biomedicine. The proportion of PubMed linkage of the articles of these authors was apparently lower than the true values, because of potential inaccuracies in the linkage process or uneven coverage in one source or the other. For these cases, biomedical researchers were definitively separated from nonbiomedical researchers by inspection of Medical Subject Heading (MeSH) terms associated with their work and the journals in which they mainly publish. Those with PubMed fractions below 60% were clearly nonbiomedical researchers.

The disposition of the 532 authors was as follows. The top 407 core biomedical researchers (seven deceased, 400 living) form our list of highly influential biomedical researchers. These are listed in Table 1 along with numbers of articles, total citation counts, h-index and the normalized score. Institutions in Table 1 are best estimates of the authors' primary affiliations, but may not be completely accurate due to mobility or joint appointments. Table 2 lists the 62 nonbiomedical researchers whose normalized score is higher than the lowest normalized score of the researchers listed in Table 1. These authors would have appeared on the list if we were making no distinction between biomedical and nonbiomedical subject areas.

Table 1. Highly influential biomedical researchers (Scopus 1996–2011)

Researcher	Institution	Main Journal	#papers	#cites	h	score	#FS100
Cohen, Philip[a]	University of Dundee	Biochemical Journal	231	29 040	83	0·410	12
Colditz, Graham A.	Washington University St. Louis	Cancer Epidemiology Biomarkers and Prevention	725	60 962	132	0·725	8
Collen, Désiré	Katholieke Universiteit Leuven	Thrombosis and Haemostasis	375	25 706	80	0·384	1
Collins, Francis S.	US NIH	Nature Genetics	504	63 972	107	0·660	12
Collins, Rory	University of Oxford	Lancet	201	41 715	63	0·409	6
Colombo, Antonio	Universita Vita-Salute San Raffaele	Catheterization and Cardiovascular Interventions	632	29 351	84	0·415	5
Cook, Deborah J.	McMaster University	Critical Care Medicine	481	29 612	86	0·422	9
Cooper, Cyrus	University of Southampton	Osteoporosis International	630	24 663	82	0·385	7

#papers: number of papers in 1996–2011; #cites: number of total citations of papers published in 1996–2011 as of end-2011; H: Hirsch h-index; Score: normalized score; #FS100: number of papers with at least 100 citations authored as single or first author; #L100: number of papers with at least 100 citations authored as last author; #M100: number of papers with at least 100 citations authored as middle author; AAS: authorship-adjusted score.

[a] Some errors noted for the respective Scopus author identifier, based on communication from the author.

Table 2. Highly influential researchers with a focus outside biomedicine (Scopus 1996-2011)

Researcher	Institution	Journal	#papers	#cites	h	score	#FS100	#L100
Alivisatos, A. Paul	University of California at Berkeley	Nano Letters	225	45 610	87	0·505	6	57
Antonietti, Markus	Max Planck Institute of Colloids and Interfaces	Chemistry of Materials	453	22 468	83	0·377	8	38
Barabasi, Albert-László	Northeastern University	Physical Review Letters	173	37 064	67	0·398	10	31
Bawendi, Moungi G.	Massachusetts Institute of Technology	Physical Review B	223	25 008	78	0·374	0	40
Brinkmann, Jonathan V.	Apache Point Observatory	Astronomical Journal	319	37 540	104	0·519	0	14
Buchwald, Stephen L.	Massachusetts Institute of Technology	Journal of the American Chemical Society	280	27 886	96	0·446	2	91
Caruso, Frank	University of	Langmuir	337	22 775	78	0·363	22	34

Abbreviations as per Table 1.

The remaining 63 authors are not included in either list for the following reasons: 12 were removed from the list because the numbers of articles associated with those author names, combined with them having common names, made us suspicious that these were cases of polysemy; seven had fewer than three papers cited at least 100 times for which they were first, single or last authors, likely indicating that others were principal investigators on the vast majority of the work associated with their publications; and 44 had normalized scores less than those of the biomedical authors in Table 1.

Tables 1 and 2 also list the dominant specialty journal (excluding the multidisciplinary journals *Science*, *Nature*, *PNAS* and *PLoS One*) for the articles of each highly cited scientist. This information may offer insight for the main field(s) where each scientist is working. To offer some further information, we have also selected differentiating MeSH terms from their publications. To do this, we first remove a list of the most common terms (e.g. human, male, female, etc.), check terms and geographical terms; after this step, the top remaining terms for each scientists may still be too generic (a common issue with any thesaurus); therefore, we use the following formula to rank MeSH terms: $\log(1 + n/\text{nptot})*n/\text{nkwd}$ where n = #times the MeSH term occurs (for an author), nptot=number of papers by the author, nkwd = #times the MeSH term occurs across all of PubMed. The formula thus rewards a MeSH term that is dominant for the author, but penalizes that term if it is common. To avoid selecting MeSH terms that are highly specific to the author but represent only a tiny proportion of his/her work, we limit this exercise to MeSH terms that appear in at least 10% of the author's articles. MeSH terms for the researchers of Table 1 appear in the longer online version of Table 1 (http://www.mapofscience.com/?page_id=761) for the whole period 1996–2011 and also limited to 2005–2011. MeSH terms are not provided for the researchers in Table 2, because MeSH coverage of nonbiomedical sciences is relatively sparse and most researchers would not be adequately represented by the minority of their papers indexed also in PubMed.

Limitations and caveats

Due to the above-mentioned issues associated with the author identity problem, we are well aware that not all of the author records for those listed in Tables 1 and 2 are completely accurate. Scopus author profiles are also subject to change over time as Elsevier improves assignment algorithms and responds to author requests for changes. Despite the lack of total accuracy, the methodology used is sufficient for our purposes of identifying a set of highly influential biomedical authors. By choosing researchers with the highest total citation counts and/or h-indexes, even if some of those researcher profiles contain a few papers from other researchers with the same name, this does not diminish the fact that these are highly influential researchers.

We are also well aware that many researchers have published highly significant works prior to 1996. This is an inherent limitation to Scopus data. However, using this common time window of 1996–2011 reduces the effect of age and offers a more level ground for comparing productivity and impact over a similar period of time. Scientists whose key work was published exclusively or predominantly before 1996 are not captured in our list. Relatively, young scientists may also be at a disadvantage because they may not have reached full productivity by 1996 and very young scientists who started publishing late in the 1996–2011 window are at a major disadvantage. Nevertheless, the selected window captures highly influential scientists with a wide range of ages, most of whom are still highly active and relevant for the current evolution of science. The total citation counts and h-indexes listed in Tables 1 and 2 are of course lower than those that can be calculated from sources that include materials from before 1996, such as Google Scholar or the Web of Science. Table 3 shows comparatively data on the h-index of a sample of highly cited authors from Scopus including papers published before 1996 and Google Scholar – the h-indices have been updated to August 2013. As shown, when all years are considered Scopus h-indices increase modestly, while Google Scholar h-indices can be substantially higher. However, the relative ranking of scientists is not affected substantially (rank correlation coefficients 0·89 for Scopus 1996–2011 vs. Scopus all years, 0·80 for Scopus 1996–2011 vs. Google Scholar, and 0·96 for Scopus all years vs. Google Scholar, $P < 0{\cdot}001$ for all).

Table 3. Comparative data on h-index for a sample of highly influential researchers calculated with different databases

Name	h-index per Table 1	h-index per Scopus (all years)	h-index per Google Scholar
Braunwald, Eugene	109	142	214
Croce, Carlo M.	104	130	168
Grundy, Scott M.	93	123	154
Ferrara, Napoleone	88	120	139
Mantovani, Alberto	91	119	144
Miller, Webb	58	67	81
Schwartz, Michael W.	88[a]	84[a]	95
Bouter, Lex M.	88	104	130
Holgate, Stephen T.	84	107	127
Kadowaki, Takashi	80	96	120
Gibbs, Richard A.	61	78	86
Buchler, Markus W.	76	83	102
Schomig, Albert	79	89	105
Hood, Leroy	71	96	146

[a] Some problems noted with attribution of papers in the Scopus author identifier (communication with author).

We also note that while both total citation counts and h-index were used in the normalized score, total citation counts are more highly skewed than h-index values. Thus, the h-index typically accounted for greater than 50% of the normalized score for a researcher.

Finally, the normalized scores that we used did not account for multiple authorship of papers and for author position and relative contributions in each paper. The final columns in Tables 1 and 2 provide some additional insight into authorship aspects. Three columns show the numbers of papers with at least 100 citations split by author position. First and single authored papers are counted together (#FS100), while last and middle authored papers are considered separately (#L100 and #M100, respectively). Exact contributions are not listed in many papers and are difficult to quantify with a single simple metric. To attempt to account for authorship patterns and to show how different metrics can change relative rankings, the final column in Tables 1 and 2 shows an authorship-adjusted score (AAS) calculated as the normalized score times the proportion of papers cited 100 times or more that are first, single or last authored papers (pFSL). The rank correlation coefficient is 0·106 ($P = 0{\cdot}033$) between total citations and pFSL, 0·012 ($P = 0{\cdot}81$) between h-index and pFLS, suggesting that citation indices and authorship positions are almost totally orthogonal. Eventually, the rank correlation coefficient is 0·42 ($P < 0{\cdot}001$) between the normalized score and AAS.

The sortable online version of Table 1 (http://www.mapofscience.com/?page_id=761) also contains information on the proportion of papers of each author that is categorized in each of 4 levels of research (applied to basic), as well as the dominant type for each scientist. Research levels were first used in 1976 by Narin *et al*. 11 when they classified biomedical journals into four types: (i) clinical observation, (ii) clinical mix, (iii) clinical observation and (iv) basic research. Machine learning was recently used to train a classifier on title and abstract words using journal research level data. This classifier was then used to assign research levels to individual articles 12. It is those research level assignments that are used to populate Table 1. All metrics and descriptors in Tables 1 and 2 are based on the time period 1996–2011.

Comparison with other lists of highly cited scientists

Thomson Reuters generates lists of the top 1% highly cited scientists for 21 different fields based on total citation counts for articles belonging to each field (in the module Essential Science Indicators) and it also used to generate lists of the 300 most-cited scientists in each of these fields, but the latter option is no longer updated. Other focused lists have been published from time-to-time in various fields 13, 14. While division per field causes more granularity and adjusts for potential differences in citation density per field, scientists who work in several of these more granular fields are at a major disadvantage vs. scientists concentrated in a more narrow focus. Moreover, the Thomson Reuters classification uses only total citations, while our score incorporates also the h-index that may be more appropriate.

Microsoft Academic Search also allows generating of lists of scientists in distinct fields and also in even more narrow subfields based on total citations or h-index, but not on a combined score. Moreover, the inclusion of papers and the respective citation data for biomedical sciences are still more limited in Microsoft Academic Search than in Scopus.

None of the previously developed lists has accounted for multiple authorship and author contributions. Moreover, they have not provided information on dominant/specific MeSH terms that may provide more accurate information on the focus of each scientist's work.

Potential uses

The list of highly influential researchers was developed so as to identify a reproducible set of researchers with high citation metrics. We are currently using this list in an ongoing survey of these scientists regarding what features of scientific papers define major impact. The list may be used in additional surveys pertaining to issues where these scientists are likely to offer a highly knowledgeable viewpoint, for example issues of funding, conduct and reporting of scientific research. Lists of influential researchers may also be a resource for identifying suitable scientists for leadership, advisory and reviewer positions. Finally, impact metrics may be used for funding decisions based on appraisal of investigator excellence instead of or in addition to specific grant proposals 15. We see the generated list as a resource that can be used creatively by other interested investigators for very diverse purposes and we welcome suggestions for improvements.

Address

SciTech Strategies, Inc., Albuquerque, NM 87122, USA (K. W. Boyack); SciTech Strategies, Inc., Berwyn, PA 19312, USA (R. Klavans); Temple University School of Medicine, Philadelphia, PA 19140, USA (A. A. Sorensen); Stanford University School of Medicine, Stanford, CA 94305, USA (J. P. A. Ioannidis).

References

Citing Literature