SCIENCE & TECHNOLOGY POLICY INSTITUTE

Identifying Outstanding Scientists Using Bibliometric Indicators

Xueying Han
Sally S. Tinkle
Pavel Panko
Nathan N. L. Dinh
Gabriella G. Hazan
Gifford J. Wong
Abby R. Goldman
Luba Katz

March 2022

IDA Document D-33048

Log: H 22-000142

IDA SCIENCE & TECHNOLOGY
POLICY INSTITUTE
1701 Pennsylvania Ave., NW, Suite 500
Washington, DC 20006-5805

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the IDA Science and Technology Policy Institute under contract NSFOIA-0408601, project TP-20-1005.DD, "S&T Bibliometric Review," for the Office of Science and Technology Policy. The views, opinions, and findings should not be construed as representing the official positions of the National Science Foundation or the sponsoring agency.

For More Information

Xueying Han, Project Leader
xhan@ida.org, 202-419-5498

Kristen M. Kulinowski, Director, Science and Technology Policy Institute
kkulinow@ida.org, 202-419-5491

Copyright Notice

SCIENCE & TECHNOLOGY POLICY INSTITUTE

IDA Document D-33048

Identifying Outstanding Scientists Using Bibliometric Indicators

Xueying Han
Sally S. Tinkle
Pavel Panko
Nathan N. L. Dinh
Gabriella G. Hazan
Gifford J. Wong
Abby R. Goldman
Luba Katz

Executive Summary

In September 2021, the White House Office of Science and Technology Policy (OSTP) asked the IDA Science and Technology Policy Institute (STPI) to evaluate whether outstanding scientists can be identified using existing bibliometric indicators and to assess the limitations of using bibliometric indicators for this purpose. STPI was also asked to develop a new analytical approach that takes advantage of bibliometric indicators to determine which individuals are outstanding scientists. Specifically, the two research questions addressed in this study are:

1. *Can bibliometric indicators be used to accurately identify outstanding scientists within a scientific discipline through time, and what are the limitations of their use?*
2. *Can new analytical approaches be developed to identify outstanding scientists within a scientific discipline through time?*

To address the first research question, STPI reviewed select publication- and citation-based indicators in addition to other types of bibliometric indicators such as the h-index and alternative metrics (i.e., altmetrics). Conventional bibliometric indicators such as total count of publications—or some fraction thereof—provide a quick way of assessing the productivity of a researcher or research group, with the assumption that more publications correlate with more research impact. While arguably the simplest bibliometric indicator to calculate, publication-based indicators can potentially overlook or inadvertently highlight a number of factors that obfuscate the quality of individual or research group contribution. Citation-based indicators have been used to estimate research impact and performance since the introduction of the Science Citation Index. These types of analysis typically involve counting the number of citations—or some fraction thereof—for a particular journal article for a period of years after its publication. Limitations to citation-based indicators include bias of self-citations, non-traditional career paths for which citation in scientific journals is not likely, and time lags between a publication and its citations.

Following our assessment, STPI found that there is not an agreed upon set of bibliometric indicators that can or should be used to identify outstanding scientists. The simplicity and ease-of-use that define bibliometric indicators precludes the inclusion of many of the factors such as career stage or having non-traditional academic career paths that determine research impact based on publications alone.

To address the second research question, STPI conducted a pilot study to determine whether publication and citation trajectories could be used to identify patterns of research impact and distinguish high impact researchers from those with low and mid-level impact for two different scientific disciplines—genetics and artificial intelligence. We found that it is possible to cluster researchers into low, medium, or high research impact groups based on individuals' publication and citation trajectories and that this clustering approach provides an alternative method to identify outstanding scientists. STPI also calculated into which impact group individuals would be clustered for select bibliometric indicators using a ranking analysis and found that there was a high degree of discordance in the results between the cluster and ranking analysis, and across bibliometric indicators within the ranking analysis. Individuals identified as outstanding, *high impact*, researchers in the cluster analysis were not strictly confirmed by the ranking analysis. This finding confirmed the challenge of using bibliometric indicators to identify outstanding scientists reported by other studies.

Based on the review of bibliometric indicators and findings from the pilot study, STPI identified several limitations as well as data challenges that would hinder any in-depth analyses looking at research impact. Specifically, additional data such as demographics, patents, funding, awards, and recognition are needed to provide additional information and context when considering who is an outstanding scientist. An overview of additional data sources that could be combined with bibliometric data is provided.

STPI identified the inability to verify outstanding scientists as a general challenge because there is no consensus on a set of criteria that characterizes an outstanding scientist. One method valuing a certain set of factors in its selection criteria will generate a different list of outstanding scientists than another method that has prioritized a different set of criteria, although some criteria will likely overlap.

STPI identifies three potential next steps for OSTP to consider that could refine a process to identify outstanding scientists. First, additional bibliometric analyses could be conducted to better identify undercited scientists who have outsized research impact. In addition, STPI could explore options to verify a group of outstanding scientists such as using expert opinion or various reputational indicators. The final follow-up direction is to examine the possibility of identifying *teams of scientists* who produce exceptional work rather than individual scientists. This strategy could be both more robust and is consistent with a generally accepted notion that many advances in science involve contributions from teams of scientists.

Contents

1. Introduction

A. Background

Classic economic theory states that a country's competitive advantage is born out of its natural endowments like its land, location, resources, labor pool, interest rates, or currency value (Krist 2013). However, in 1990, economist and business strategist Michael Porter posited that a country can actually create and sustain its own competitive advantage through four attributes (Porter 1990):

1. *Factor conditions*: factors of production such as skilled labor and infrastructure that can and should be created by a country instead of inherited such as natural resources. In particular, Porter argues that countries succeed in industries where they excel at factor creation.
2. *Demand conditions*: meeting the home demands of a given industry so that companies can better perceive, interpret, and respond to domestic buyer needs. This, in turn, will help companies innovate faster and achieve competitive advantages over their foreign counterparts.
3. *Related and supporting industries*: the presence of related and supporting industries in the nation that are also internationally competitive. Having internationally competitive suppliers and end-users in the same nation can increase the pace of innovation by having shorter communication times and providing faster feedback.
4. *Firm strategy, structure, and rivalry*: the organization and management style and practices that are favored in a country combined with the sources of competitive advantage in the industry. There is not one universally appropriate organization and management style or practice—what works for one country may not work for another country. Local rivals within a country also act to stimulate innovation and competitiveness, thereby enabling a country to have competitive advantage within an industry.

The U.S. Federal Government, along with many other countries, has focused on the first attribute of developing and maintaining human capital, particularly in STEM fields (National Science and Technology Council 2018, 2021). However, as Porter noted, having a workforce that is high school or even college educated may not confer competitive advantage to a country. Rather, a factor must be highly specialized to an industry's needs to convey advantage. Such factors are scarce and require long-term investment to sustain.

Scientists who have outsized research impact compared to that of their peers, particularly in STEM fields, arguably are a factor condition that may yield competitive advantage for a country. The ability to identify outstanding scientists within a particular field may be important to designing and implementing policies that attract and retain these highly specialized individuals.

To understand better the relationship between outstanding scientists and competitive national advantage, the White House Office of Science and Technology Policy (OSTP) asked the IDA Science and Technology Policy Institute (STPI) to first assess whether outstanding scientists, within a scientific discipline, can be identified using existing bibliometric indicators. In addition, STPI was asked to develop new indicators or an analytical approach using existing bibliometric indicators to identify which individuals are outstanding scientists. Specifically, the two research questions addressed in this study are:

1. *Can bibliometric indicators be used to accurately identify outstanding scientists within a scientific discipline through time, and what are the limitations of its use?*
2. *Can new bibliometric indicators or analytical approaches be developed to identify outstanding scientists within a scientific discipline through time?*

B. Context for This Study

Following our literature review and internal discussions, STPI developed several definitions for key constructs in the study.

Outstanding Scientists. STPI defined an outstanding scientist as an individual who has significantly higher than average research impact compared to that of their peers. These individuals could be laboratory scientists in industry or academia, scientific entrepreneurs, theoreticians, or outside-of-the-box thinkers.

In this study, STPI selected the academic scientific community as most likely to have publications and citations amenable to bibliometric analysis. The term *peers* refers to fellow researchers in an individual's scientific discipline.

Research impact. An individual's research impact is the demonstrable effect of their scientific contributions that are assessed through advances in a research field or to general scientific knowledge (research impact), and contributions to the general economic and social capital of the nation (economic impact, social impact), all of which have implications for the security of the nation (national security impact; Moravcsik 1977; Martin and Irvine 1983; Penfield et al. 2014; Wilsdon et al. 2015; Abramo et al. 2017; Bu et al. 2021). Research impact is evaluated through traditional research outputs (journal articles, conference publications, book chapters, and datasets) and non-traditional factors

(professional networks and public recognition; Wilsdon et al. 2015; Ravenscroft et al. 2017; Siudem et al. 2020).

STPI focused on research impact and traditional research outputs for this study and set aside other forms of impact (e.g., social, economic, natural security impact). A summary of these other forms of impact can be found in Appendix A.

Bibliometric Indicators. Bibliometric indicators are statistical measures of the quantity and quality of publications and other research outputs. There are two types of bibliometric indicators considered in this study: *quantity indicators*, which measure the productivity of a particular researcher; and *quality indicators*, which measure the quality (or "performance") of a researcher's output (Durieux and Gevenois 2010). Specifically, publications are used as quantity indicators and citations as the quality indicator.

C. Organization of the Report

In addressing the first research question, STPI will show that there is no universal set of bibliometric indicators that can be used to identify outstanding scientists. We also discuss limitations in the use of bibliometric indicators such as field dependence; inability to accurately predict the research impact of early career researchers and non-traditional academicians; and perpetuation of gender biases.

In addressing the second research question, STPI demonstrates a novel cluster analysis approach that uses researchers' publication and citation trajectories as an alternative approach to identifying outstanding scientists. Specifically, the cluster analysis grouped scientists into *high*, *medium*, or *low* research impact groups with those in the *high* impact group being considered outstanding scientists. The two cases investigated—AI and genetics—suggest science discipline-specific patterns for research impact, and further testing of the approach could investigate the challenges associated with data availability, the appropriate timeline for publications and citations, and internal inconsistencies between indicators and indices.

Chapters 2 and 3 address research question one and Chapter 4 addresses research question two. The report is organized as follows:

- Chapter 2 provides a review of select existing publication- and citation-based indicators, as well as other types of bibliometric indicators.
- Chapter 3 presents an overview of select limitations to the use of bibliometric indicators.
- STPI introduces a new method to identify outstanding scientists in Chapter 4 using a clustering approach based on researchers' publication and citation trajectories through time, which is tested on two scientific disciplines, genetics and artificial intelligence (AI).

- Chapter 5 identifies additional data sources that when combined with bibliometric data may mitigate existing data challenges, help identify outstanding scientists, and answer broader questions about U.S. competitiveness.
- Lastly, Chapter 6 proposes possible additional studies that would expand on the work performed thus far.

2. Review of Bibliometric Indicators

In this section, we review select publication- and citation-based indicators along with other types of bibliometric indicators. The following is not a comprehensive review of all bibliometric indicators but rather a selection of those deemed most appropriate to identify outstanding scientists, and that are easy to moderately easy to calculate and interpret. A more comprehensive review of 108 bibliometric indicators can be found in Wildgaard et al. (2014).

A. Publication Indicators

The total number of publications produced by an individual is often seen as an indication of research productivity, with the number of publications directly proportional to the magnitude of productivity (King 1987; Gauffriau et al. 2007). Since the advent of databases on scientific publications in the 1960s, both the contents and range of data covered by databases have increased (e.g., Larsen 2008). A 2014 review identified a number of publication-based methods of assessing researcher impact (

Table 1) including whole counting (i.e., counts of total publications; Wildgaard et al. 2014); first author publications (Cole and Cole 1973); weighted publications; patent applications (Okubu 1997); all public contributions that include tv, radio, and websites (Mostert et al. 2010); and fractional counting of contribution (Price 1976); proportional (Van Hooydonk 1997); geometric (Egghe et al. 2000); and harmonic (Hodge and Greenberg 1981).

Table 1. List of Publication-Based Bibliometric Indicators and How They're Calculated

Publication-Based Indicator	Description
Whole Counting	Each *N* author of a paper receives equal credit
First Author Counting	Only the first listed author receives credit for a publication
Weighted Publication Counting	Applies a weighted score to the type of output
Patent Application Counting	Count of patent applications only
Counting of All Public Contributions, Including TV, Radio, and Websites	Count of all contributions disseminated in the public sphere

Publication-Based Indicator	Description
Fractional Counting	Each of the *N* authors is credited by a value equal to 1/*N*
Proportional Counting	Author with rank R in by-line with *N* co-authors (e.g., R=1, … *N*) receives score *N*+1-R
Geometric Counting	Author with rank R with *N* co-authors receives credit of 2**N*-R
Harmonic Counting	Ratio of credit allotted to *i*th and *j*th author is *j:i* regardless of total number of co-authors

According to the literature, there are three publication-based counting methods: whole counting, fractional counting, and first author counting (Gauffriau et al. 2007; Table 2). For the purposes of this study, we focused on whole counting and fractional counting. In whole counting, all authors contributing to a publication receive one credit regardless of the number of authors. In fractional counting, all authors contributing to a publication share one credit with equal fractions assigned to each listed author. For example, if a publication has six authors, each author receives a publication count of 1/6.

While arguably the simplest bibliometric indicator to calculate, publication-based assessments can potentially overlook or inadvertently highlight a number of factors that obfuscate the quality of the actual contribution, including: the variety of publication practices across fields and between journals (e.g., King 1987), the trend in shorter papers that lead to the "least publishable unit" phenomena (e.g., King 1987; Budd and Stewart 2015), or the difficulty in determining an author's contribution unless a statement describing their level of contribution is included (Bennett and Taylor 2003).

Table 2. Common Publication-Based Bibliometric Indicators and Their Ease of Calculation

Publication-Based Indicator	Description	Easy to Calculate	Data Are Readily Available
Whole Counting	Each *N* author of a paper receives equal credit	✓	✓
Fractional Counting	Each of the *N* authors is credited by a value equal to 1/*N*	✓	~
First Author Counting	Only the first listed author receives credit for a publication	✓	✓

Tildes (~) denote moderate agreement with the category.

B. Citation Indicators

The total number of citations received by a publication over a period of years after its publication is often seen as an indication of research quality, with the number of citations directly proportional to the magnitude of impact (King 1987). Citation analysis began in earnest with the publication of the Science Citation Index (SCI) in 1961 (MacRoberts and MacRoberts 1989) and is widely used to this day.

Waltman (2016) describes five basic citation indicators: total number of citations, average number of citations per publication, number of highly cited publications, proportion of highly cited publications, and the h-index (described subsequently). For the purposes of this study, STPI selected to focus on total number of citations, average number of citations per publication, and fractional counting of citations per publication (Table 3).[1] The average number of citations per publication returns the calculated mean number of citations per publication. The total number of citations considers all citations including self-citations. The fractional counting of citations per paper assigns each author contributing to a publication a share of the citation count with equal fractions of the total citations to each listed author.[2]

While efforts have been made to normalize citation counting to a particular scientific field or set of publications (Wildgaard et al. 2014), limitations to citation-based measures persist and include:

- data completeness and consistency (e.g., not every database provides citation counts, different cited reference counts can occur depending on which source is used);
- qualitative issues regarding how to best account for researcher impact—for example, considering how different citation-based indices, like i10[3] or fractional citation (Egghe 2008) reflect different aspects of a researcher's impact;
- timeliness of a citation is often not accounted for (e.g., MacRoberts and MacRoberts 1989);
- variation of citation rates between document types and research fields (e.g., MacRoberts and MacRoberts 1989);

---

1 This citation impact indicator was chosen by STPI because it mirrors the fractional counting of publications as another way to avoid increasing the total weight of a single paper.

2 In other words, fractional counting of citations per paper considers crediting an author of an *N*-authored paper with *c* citations a value equal to *c/N*.

3 The i10-Index is the number of publications from an author with at least 10 citations. It was created by and used exclusively in Google Scholar. A more detailed description can be found from the University of California San Diego Library at https://ucsd.libguides.com/c.php?g=704382&p=5000890#.

- the most significant paper may not necessarily be the paper with the most citations (e.g., Wildgaard et al. 2014);
- self-citations (e.g., Szomszor et al. 2020); and
- it may be inappropriate to treat citations equally due to the many factors that govern citing, such as an author's intellectual or social motivations, and ultimately promotes articles—incorrect, controversial, or retracted—regardless of its actual relevancy (e.g., see Zhang et al. 2013; Hernández-Alvarez and Gomez 2015; Tahamtan and Bornmann 2019).

Table 3. Common Citation-Based Bibliometric Indicators and Their Ease of Calculation

Citation-Based Indicators	Description	Easy to Calculate	Data are Readily Available
Total Number of Citations	Considers all citations including self-citations	✓	✓
Average Number of Citations per Publication	Considers the mean number of citations per paper	✓	✓
Fractional Counting of Citations per Publication	Considers crediting an author of an *N*-authored paper with *c* citations a value equal to *c*/*N*	✓	~

Tildes (~) denote moderate agreement with the category.

C. H-index and Variations

The h-index is an author-level metric that incorporates both an author's productivity and impact as measured by the citation rate (Hirsch 2005). The h-index is defined as:

"A scientist has index h if h of his or her N_p papers have at least h citations each and the other ($N_p - h$) papers have $\leq h$ citations each."

The set of papers that have $\leq h$ citations are also referred to as an author's h-core (i.e., those publications that determine the h-index). For example, as seen in Table 4, this example author has papers a–g that are ranked by their citation counts. Based on this ranking, this author has an h-index of 5, which is the largest rank magnitude that is less than or equal to the paper's number of citations ($5 < 7$ citations).

Table 4. Example to Calculate the H-Index

Title	Rank	Citation Count
Paper c	1	200
Paper f	2	65

Title	Rank	Citation Count
Paper d	3	24
Paper g	4	12
Paper b	5	7
Paper a	6	3
Paper e	7	2

While the h-index is easily understood and calculated, multiple studies have highlighted its shortcomings (for example, Bihari et al. 2021). These include:

- the lack of adjustment for different citation practices across disciplines
- the inclusion of self-citations in the computation of the h-index
- the inability of the h-index to adjust for decreases in productivity
- the lack of credit for future publications unless they become a part of the h-core
- the lack of extra credit to highly cited articles
- the disregard for all citations in the h-tail (i.e., all citations outside of the h-core)
- the equal weight given to all authors of each publication regardless of actual contribution level

Because of these shortcomings, an entire field of study has been developed that adjusts the h-index or create more tailored indices. These variations on h-index attempt to account for the shortcomings of the h-index by incorporating various mathematical adjustments. For example, a 2006 study attempted to allocate more credit to highly cited articles by defining the g-index as "the highest number, *g*, of papers that together received g^2 or more citations" (Egghe 2006).

By squaring all citation values, the g-scores of highly cited publications will be higher than their h-scores (i.e., their total number of citations) and therefore incorporate more of the highly cited articles. For example, a rank 1 paper with citations = 100 has a $g^2 = 10{,}000$. As this is repeated for all papers, the rank magnitude that determines the *g*-index score will be naturally larger than the h-index score. However, in doing so, the g-index has been noted to provide *too much* credit to highly cited articles, where a single very highly cited article may skew the g-index and dilute all other publications. Accordingly, a 2010 study introduced the hg-index, which attempts to balance the credit given to highly cited articles (or lack thereof) by the h- and g-indices (Alonso et al. 2010). These continual adjustments by various researchers have spawned a plethora of variations to the h-index that all attempt to produce a single output to rank author impact, productivity, and quality of work. Some of these variations include adjusting for the average number of co-authors that an author has (h_i-index; Batista et al. 2006), penalizing authors for self-citations (discounted h-index;

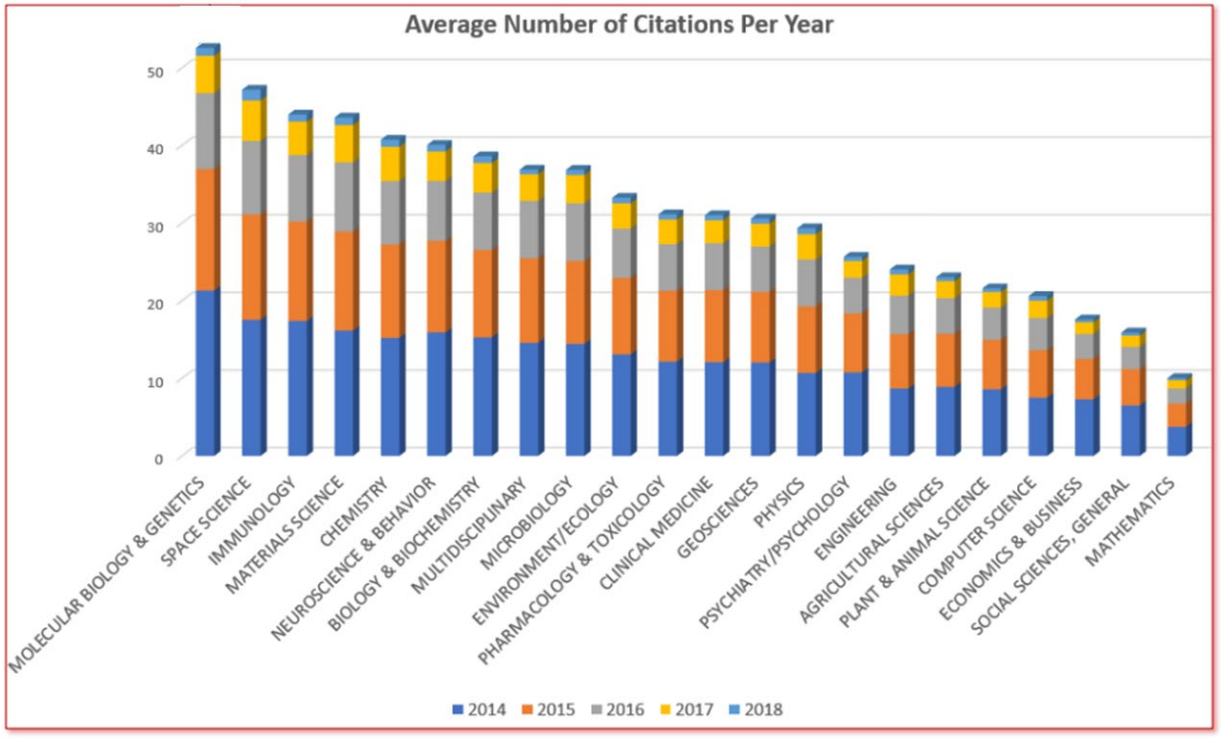
Ferrara and Romero 2013), accounting for research career age (v-index and m-quotient; Vaidya 2005; Burrell 2007), and level of contribution to articles (normalized h_i-, fractional h-, and fractional g-indices; Egghe 2008; Wohlin 2009). For a more in-depth analysis of selected variations, please see Appendix B.

During this h-index review, STPI noted that despite the emergence of many alternatives to the traditional h-index, most of these measures have not been broadly adopted. Citation platforms that incorporate modifications to h-index are, therefore, limited. For example, Publish or Perish (PoP), a citation management program, integrates the h-index, g-index, contemporary h-index, Zhang's e-index, AR-index, and several other publication-level indices into its platform (Harzing 2016). Google Scholar incorporates the h5-index and h5-median, which only consider publications from the last 5 years. Web of Science (WoS), a citation database, only displays an author's h-index in addition to their total number of publications and citations.

There is also limited research into the predictive power of h-index variations as these are generally used to rank researchers based on existing work. The literature that attempts to calculate the predictive power of h-index variations is generally limited to one database (e.g., WoS or Google Scholar), which may vary greatly in quality and depth—or to one science discipline, which may have different publishing and citing patterns from other disciplines. These differences in citation rates across disciplines can be observed in

Reprinted from Green (2019)

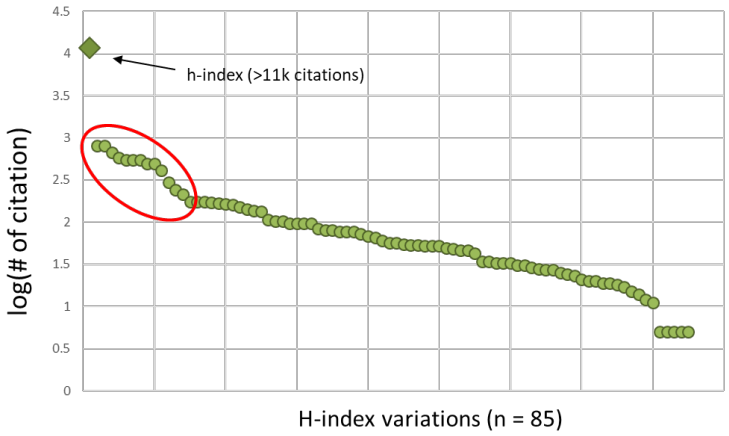
Figure 1.



Reprinted from Green (2019)

Figure 1. Citations Rates across Science Disciplines

To select variations for further analysis from the wide range of options, STPI devised a prioritization strategy to focus on the most widely cited indices for initial analysis. Using the list of h-index variations reviewed in Bihari et al. (2021), the 85 options were ranked by the number of citations that their original publications received according to Google Scholar (Figure 2). The publications of 14 h-index variations with more than 200 citations (red circle in Figure 2) were then reviewed in-depth.



Note: Each datapoint represents an individual h-index variation that was reviewed. Each variation is graphed based on the log number of citations that its original publication garnered. The variations that were analyzed in depth are circled in red.

Figure 2. H-index and Variations by Publications' Citation Count

Following review of publications of the 14 highest cited h-index variations, 5 indices were selected for the following properties: they were either easy to calculate or had already been calculated (e.g., PoP or Google Scholar); demonstrated predictive power; adjusted for career age; or used in prior studies to identify outstanding scientists. The selected indices are the AR-index, m-index, contemporary h-index, hg-index, and A-index (Table 5). A full explanation on how to calculate the h-index and these five variations can be found in **Error! Reference source not found.**.

Table 5. Characteristics of Five Selected Variations

Index	Existing calculator	Easy to calculate	Predictive power demonstrated	Adjusts for productivity over time	Used in the context of identifying outstanding scientists
AR-index	✓	~		✓	
m-index	✓	✓	✓	~	✓
Contemporary h-index	✓	~	✓	✓	✓
hg-index		✓	✓	~	
A-index		✓		~	

Tildes (~) denote moderate agreement with the category.

D. Alternative Metrics

While the h-index and other citation-based indicators are based on published literature, alternative metrics (altmetrics) consider citations and references by non-academic publications such as public policy documents; discussions on research blogs; coverage on mainstream media; and mentions on social media such as Twitter and Facebook (Altmetric 2022). Altmetrics consider how widely disseminated an article is beyond the publishing journal and immediate scientific community and how much attention an article receives from the public sphere (Altmetric 2022). Traditional citation-based indicators are unable to measure the immediate impact of an article because of the lag time between when an article is submitted and accepted to when it receives its first journal citation. Therefore, altmetrics complement the h-index and other traditional measures.

Altmetrics have their own limitations, however, such as the possibility of being gamed through the use of bots or humorous titles; the lack of standardization or quality control in collection methods; and the lack of coverage for a vast majority of publications (Michael Thelwall 2020). Further, altmetrics have not been studied within the context of systematically identifying outstanding scientists; instead, most studies have been conducted on publication-level rather than at the author-level (Akella et al. 2021; Bornmann and Haunschild 2018; Costas, Zahedi, and Wouters 2015; Ringelhan, Wollersheim, and Welpe 2015; Mike Thelwall et al. 2013; Mike Thelwall and Nevill 2018).

Although there is limited research for the use of altmetrics within the context of identifying outstanding scientists, the current body of literature may help to inform potential prediction strategies should a standardized, author-level metric arise. A brief overview of alternative metrics is provided below in Table 6. Altmetrics should not be

viewed as an alternative to traditional, citation-based indicators but rather, as a complementary measure of impact on the general public.

Table 6. Alternative Metrics

Metric	*Pros*	*Cons*
Altmetric compilation sites (Altmetric.com, PlumX, Crossref Event Data, ImpactStory)	• Comprehensive and curated aggregation of public policy documents, mainstream media, blogs, online reference managers, patents, Wikis, Facebook, Twitter, and other sources	• Limited coverage (only 15% of WoD DOIs covered between 2011 and 2014 (Costas et al. 2021)) • Exact calculations may not be made publicly available • Coverage of each source differs by platform • Scores are only provided at the publication-level rather than author-level
Mendeley readers	• Reflects scholarly and partly educational impact • Could be used to measure impact before citations begin via readership • Prior studies by Thelwall (2018), Nuzzolese (2019), and Akella (2021) have found some predictive power for future publication-level citations	• May undercount all readers • Does not necessarily reflect societal impact
Health website citations	• High quality websites provide direct evidence of societal impact	• Low production number per article
Google Books citations	• Supplemental citation count that is missed by academic journal article counts	
Online syllabus mentions	• Direct educational impact measure	• Most syllabi are likely private (although search engines may be able to find a select few)
Wikipedia citations	• Small but significant correlation between Wikipedia and Scopus citation counts • Measure of “information impact”	• Low production number (5% of academic articles)
Blogs	• Weak positive correlation with citation counts	• Low production number (6% of recent articles)
Patents	• Direct commercial impact measure • Easy to collect • Positive correlation with citation counts	• Low production value (<1% of journal articles)
Grey literature citations	• Collection method exists for some government websites	• Collection may be difficult to collect
Tweets	• One of the most common platforms for publication sharing	• Lack of explanation for use • Low positive or negative correlations with citation counts • Impact may not necessarily be tied to academia

Note: Data in these tables are derived from Thelwall (2020) and Ortega (2020)

3. Limitations to the Use of Bibliometric Indicators to Identify Research Impact

STPI's review of bibliometric indicators identified several limitations to their use in assessing research impact, which include differences across disciplines, lack of baseline comparison, bias, and misuse.

A. Field-Dependent Publication and Citation Patterns

Research questions, processes, standards for authorship inclusion, and rate of progress vary by scientific discipline, which influence publication and citation patterns. A 2021 study that analyzed more than 4.1 million documents in Scopus and found substantial differences in the types of publications used across seven disciplines (Mendoza 2021). Specifically, conference papers made up 61% of computer science publications but only 38% in engineering and 19% in physics. The authors also found temporal differences in highly cited publications: the top papers in medicine, physics, biochemistry, and chemistry were published in the 1990s and 2000s whereas the top papers in mathematics and psychology had longer timespans from the 1970s to the 2000s. Similarly, a 2017 review found that the citation time window that could serve as a predictor of future citation rates varied by field: mathematics had the longest citation half-life; citations in biology, biomedical research, chemistry, clinical medicine, and physics peaked quickly after publication; whereas Earth and space science along with engineering followed a more regular and slower-growing trend (Abramo et al. 2017). A 2015 study compared metrics across astronomy, environmental science, public health, and philosophy and found significant differences between the number of publications and citations between disciplines and between publication databases (Wildgaard 2015). Consequently, if impact is being considered for scientists in multiple fields, any potential indicator would have to be adjusted.

A specific example of how various bibliometric indicators are typically used includes the different analyses conducted on published h-index variations. Of the five h-index variations that STPI reviewed (the m-, contemporary h-, hg- A-, and AR-indices), only three variations tested the predictive power of the indices (m-, contemporary h-, and hg-indices). These three publications used highly varied methodologies to test the predictive powers. The m-index was used to test the likelihood of a scientist to be awarded a Boehringer Ingelheim postdoctoral biomedical fellowship (Bornmann et al. 2008); the hc-

index analyzed the ranking of computer scientists (Sidiropoulos et al. 2007); and the hg-index analyzed the ranking of astrophysicists (Alonso et al. 2010).

B. Name Disambiguation

Accurately associating names with publications is a major challenge in bibliometric analysis. For example, authors with the same or similar names often had combined WoS profiles (Milhaljevic et al. 2019). This was particularly evident for researchers of Chinese descent. A census conducted by the Chinese Public Security Bureau found that 21.4% of the Chinese population had the surname Li, Wang, or Zhang, which makes disambiguation very difficult findings (Fish 2013). Current name disambiguation methodologies may also negatively affect those who change their names. STPI is not aware of any datasets that automatically account for name changes without direct involvement by the author themselves. There are efforts currently underway at WoS and with the University of Michigan's UMETRICS initiative to improve name disambiguation, but the level of accuracy across population groups is unknown. Another approach to resolve name disambiguation is through the use of unique, personal identifiers that can be used to link an individual researcher to his/her research outputs, funding, or any other professional distinctions such as those provided through the Open Researcher and Contributor ID (ORCID), which are increasingly in use.

C. Gender Disparities

Gender differences in publication patterns are well documented (Cech and Blair-Loy 2019; Duma 2020; Huang et al. 2020; Viglione 2020) and could influence the results of bibliometric analyses. A study conducted in 2013 that analyzed over 5 million papers with more than 27.3 million authorships found that when a woman is either the sole author, first-author, or last-author, the paper garnered fewer citations than in cases when a man was in one of these positions (Lariviere et al. 2013). Women and men were also more likely to be overrepresented in certain disciplines. Women dominated authorship in nursing; midwifery; speech, language, and hearing; education; social work; and librarianship. Men dominated fields that included military sciences, engineering, robotics, aeronautics and astronautics, high-energy physics, mathematics, computer science, philosophy, and economics. Finally, at least one study documented differences in publication patterns by career stage: female mathematicians were less likely to publish than men at the beginning of their careers and more likely to leave academia (Mihaljević-Brandt et al. 2016). Given these differences, the use of bibliometric indicators to assess research impact—especially for purposes of hiring, funding, or tenure decisions—could provide inaccurate information that perpetuates these gender disparities.

D. Early and Mid-career Scientists

The use of traditional data on early- or mid-career scientists to predict later research impact may discount those with non-traditional career paths.[4] For example, those who enter research careers later in life or take a leave from academia would likely have lower early career publication and citation rates. Accordingly, any predictive indicator that is dependent on early career citation patterns could underestimate any future contributions of these individuals.

E. Matthew Effect

It is generally assumed that academia is a meritocratic system in which everyone is equitably awarded for their contributions. Some studies challenge this view, however. For example, a 2018 study showed that a small percentage of researchers receive the majority of research funding and that researchers who received the funding did not necessarily submit better proposals (Bol et al. 2018). The idea that success begets success in science was first introduced in 1968 and is referred to as the Matthew effect (Merton 1968) and applies to various indicators of recognition, including receiving grants and receiving prestigious academic appointments and awards. The Matthew effect may be exacerbated in fields that have a large number of papers published per year because the most-cited papers receive a disproportionate share of future citations at the expense of new, innovative ideas (Chu and Evans 2021). The Matthew effect is important to consider when trying to identify outstanding scientists as it can disadvantage individuals with lower public profiles.

F. Additional Considerations

Multiple studies identified other biases in evaluating research impact through bibliometric indicators (Belter 2015; Simko 2015; Wang et al. 2017). For example, a 2015 study noted that various strategies can be used to artificially inflate citation counts, such as citing papers that are focused on methodologies, crediting experts, or discussing flawed results (Belter 2015).

---

[4] A *traditional* career path is typically considered to involve acquisition of a terminal degree, placement within a research career, and continued growth of publications and citations.

4. Alternative Approach to Identify Outstanding Scientists: A Pilot Study

A. Introduction

The review of the literature performed for this study has not yielded a reliable existing strategy to identify outstanding scientists. To address OSTP's interest in developing new indicators or analytical tools to achieve this goal, STPI explored a novel bibliometric approach that took advantage of cluster analysis. Cluster analysis is a statistical method that organizes items into groups, or clusters, on the basis of how closely they are associated. Unlike many other statistical methods, it is typically used when there is no assumption made about the likely relationships within the data. Cluster analysis provides information about where associations and patterns in data exist, but not what they are or might mean (Tan et al. 2013; Aggarwal 2018).

To test this approach, STPI developed a pilot study for academic researchers in two disciplines: genetics (biological sciences) and artificial intelligence (physical sciences). STPI developed two problem statements, and used publication and citation data as the basis for clustering. A random selection of researchers within the clusters were then ranked by bibliometric indicators to assess the robustness of the cluster analysis. Additional discussion of the assumptions and the rationale for performing this study can be found in Appendix D.

The hypotheses tested were as follows:

- Publications and citations will provide a recognizable pattern of research impact.
- The trajectory of publications and citations will, over time, distinguish high impact researchers from those with low- and mid-level research impact.

The remainder of this chapter provides the methods used for the analysis, the results, and their implications for the goal of the assessment.

B. Methods

1. Data Collection and Cleaning

Author citation and publication data were drawn from the Google Scholar database.[5] STPI notes that while publication data are partially curated by Google, the author profiles are the responsibility of the author and the level of information available varies. Authors can report a maximum of five interest areas in their profiles.

For the purposes of this pilot study, STPI conducted two independent analyses by selecting all authors who listed *genetics* and all authors who listed *artificial intelligence* as one of their academic disciplines. For each author, STPI obtained all listed publications, along with annual citation counts, co-authorship, and publication date.

STPI identified several data quality challenges during the data cleaning process, the majority of which were resolved, as shown in Table 7. One limitation that could not be resolved, however, is when individuals have the same name. Publications in Google Scholar may be listed jointly under either one or both profiles, unless manually curated by the authors. An assessment of the upper quartile of the well-cited and published authors did not find any definitive namesake authors, so these types of authors were allowed to remain in the sample as their influence was not likely to have an outsized impact on the subsequent analyses.

Table 7. Data Quality Issues with Google Scholar Database and Steps Taken to Mitigate Them

Challenge	Resolution
Listed publication years are often incorrect (e.g., contemporary scholars having publications that are purportedly hundreds of years old)	Remove all publications published prior to 1776
Some publications and citations are missing the year	Remove any publications or citations that are missing the year
Authors have citations listed prior to having any published works as well as publications much before they were ever cited	Remove any instances where citations are listed before any publication was recorded and any publications that were 5 or more years earlier than the first listed citation
Authors have different periods of time in the field	Truncate each author's publication and citation trajectory to 10 years

[5] The Google Scholar database includes a listing of published works such as academic journal articles, reports, and software, as well as a feature that allows authors to build their profile by including supplementary information such as top co-authors and listed academic discipline.

Challenge	Resolution
Author statistics are heavily skewed toward individuals with fewer citations and publications	Only include authors that have 100 or more citations and at least 1 publication

2. Cluster Analysis

The data from Google Scholar were used to identify several classes of authors that differ based on the properties of their academic careers as indicated by their publication and citation information. Namely, STPI hypothesized that researchers may differ based not only on the number of publications and citations (i.e., magnitude of the research output), but also on the trajectory of the number of publications and citations across time. Two rounds of cluster analysis were used to estimate (1) the model-implied categorization of the authors' publication and citation trajectories over time, and (2) the overall grouping based on the categories obtained in step 1 and the magnitude of the research output.

The first round of clustering was used to identify a concise categorization of the publication and citation trajectories of the sampled authors. The difference in trajectories was hypothesized to indicate the degree of career success and to assist in identifying outstanding scientists. To accomplish this, STPI used the *kmlShape* package (Genolini 2016) via the R software environment (R Core Team 2021). The kmlShape package facilitates clustering longitudinal data with respect to the shape of the trajectory over time, placing individuals with similar trends into the same group. Yet, as the underlying algorithm of kmlShape is computationally demanding, a data reduction step is taken prior to the clustering. Because of the large number of individuals in the sample and the relative comparability of many of the trajectories, STPI algorithmically selected 100 representative trajectories, referred to as *senators* in the kmlShape package, to stand in for the remaining trajectories in the sample during the clustering. The more senators chosen, the more representative the clustering results will be, as they will include more potential variations of trajectories. However, the more senators included exponentially increases the time it takes to run the clustering functions. The number 100 was chosen to optimize the variations included in the analysis and the time needed for the function to run. Based on the resulting categories computed on the 100 representative trajectories, each author was assigned to two categories, one for their citation trajectory and another for their publication trajectory.

The second round of clustering identified sub-groups within the sample of authors based on the number of publications produced and citations received, as well as the citation and publication trajectory categories identified in the previous step. From the second round of clustering, individuals are classified into one of three impact groups: *high*, *medium*, and *low impact*. Two candidate clustering methods were used to determine the best-fitting model for grouping authors: Partitioning Around Medoids (PAM; Kaufman and Rousseeuw 1990) and agglomerative hierarchical clustering (see Kaufman and Rousseeuw

2009) as implemented by the R package *cluster* (Maechler et al. 2021). STPI used the agglomerative coefficient (AC), which is an aggregate measure of how well all authors fit with the groups they were assigned, to identify the appropriate link function. A more in-depth explanation of each clustering method can be found in Appendix A.

3. Ranking Analysis Using Select Bibliometric Indicators

Following the cluster analysis, three authors from each of the three impact groups were randomly selected to assess whether they would be categorized into the same high, medium, or low impact clusters based on five commonly used publication and citation-based indicators: total publication count, fractional publication count, total citation count, fractional citation count, and mean citations per paper. STPI also assessed how each of the nine authors would be grouped based on their h-index and five h-index variations (i.e., m-index, A-index, AR-index, hc-index, and hg-index). The publication and citation histories of these nine authors were downloaded from WoS because Google Scholar does not provide annual citation rates at the publication-level.[6] The indicators, h-index, and variations were then calculated for each author using total publication and citation count at the 10-year mark (i.e., 10 years after their first recorded publication). To preserve anonymity, these nine authors were assigned letters A–I in this report. These authors were then regrouped by each metric using a standard competition ranking method[7] to show which impact group they would be grouped into.

C. Results

STPI performed the pilot study on two disciplines—genetics and AI—to observe how discipline dependent publication and citation patterns influence which individuals are characterized as *high impact.* The results of the clustering and ranking analyses are described below for genetics and AI, respectively.

1. Genetics

a. Cluster Analysis

STPI downloaded 15,045 author profiles from Google Scholar on November 9, 2021 that had listed *genetics* as a research interest area. After data cleaning, a total of 7,877 author profiles were used in the study. Our analysis focused on the first 10 years following an individual's first publication, and individuals were clustered into 3 groups representing

---

6 Please see Appendix G for further explanations about these data.

7 In standard competition ranking, individuals who have equal values receive the same ranking number. For example, in a competition, if there are two silver medalists with the same score, both are awarded a silver medal, but there is no bronze medalist.

3 tiers of research impact (i.e., low, medium, and high). The results of the initial clustering of publication and citation trajectories for the 100 senators, and the final clustering of all Google Scholar identified geneticists are detailed below.

1) Publication trajectories

From the initial clustering based on publication trajectories, 87 of the 7,877 author profiles were classified in the *high impact* cluster (blue); 1,051 in the *medium impact* cluster (green); and 6,739 in the *low impact* cluster (red; Figure 3).[8] The percentages in Figure 3 refer to the fraction of senators in each cluster group, with the bolded lines depicting the averages in that cluster group.

2) Citations trajectories

Based on clustering of citation trajectories, 20 researchers were classified in the *high impact* cluster (blue); 205 in the *medium impact* cluster (green); and 7,652 in the *low impact* cluster (red; Figure 4). As can be seen from Figure 4, those classified as *high impact* (blue) begin to separate from the rest of the sample early on in the first 10 years of their careers, and sustained this trajectory over the entire study period. Furthermore, both the publication and citation *high impact* trajectories are higher than those of the *medium* and *low impact* trajectories. STPI identified this group as *high impact*.

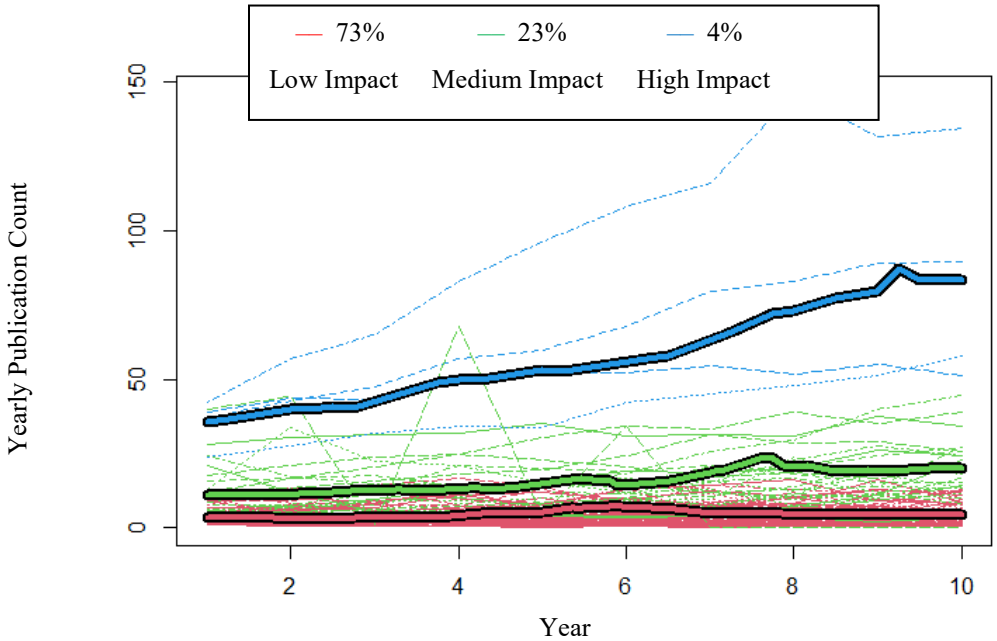


Figure 3. Clustering of Genetics Researchers' Publication Counts over 10 Years

[8] For all cluster graphs, the top bolded line represents the *high impact* cluster, the middle bolded line the *medium impact* cluster, and the bottom bolded line the *low impact* cluster.

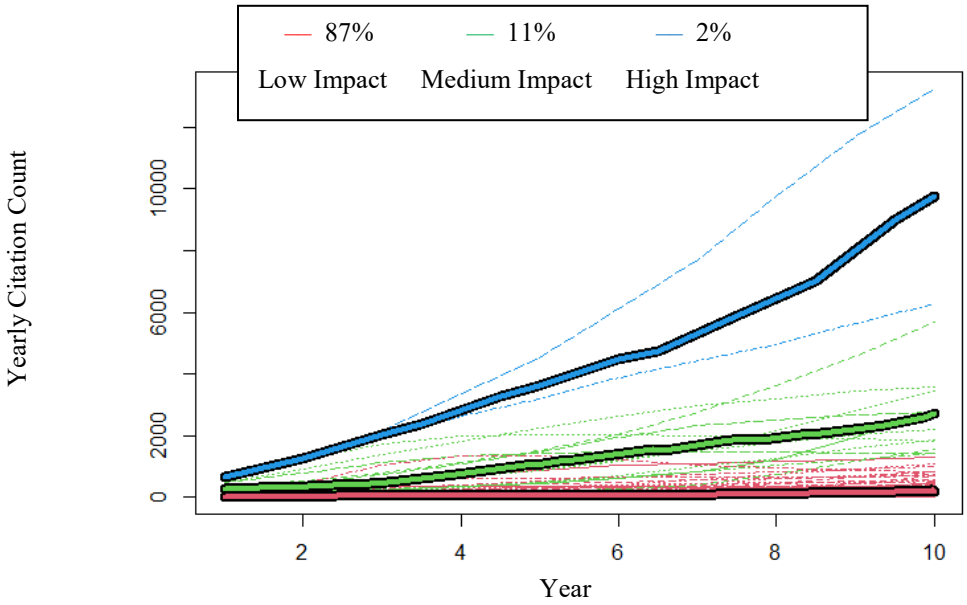


Figure 4. Clustering of Genetics Researchers' Citation Counts over 10 Years

3) Publication and citation trajectories

After the second round of clustering in which individual's publication and citation trajectories along with the total number of publications and citations were taken into consideration, 157 of the 7,877 author profiles (2%) were classified as *high impact*; 6,571 (83%) were classified as *medium impact*; and 1,149 (15%) were classified as *low impact* (Figure 5). The number of individuals in each cluster group can be found in Appendix F.

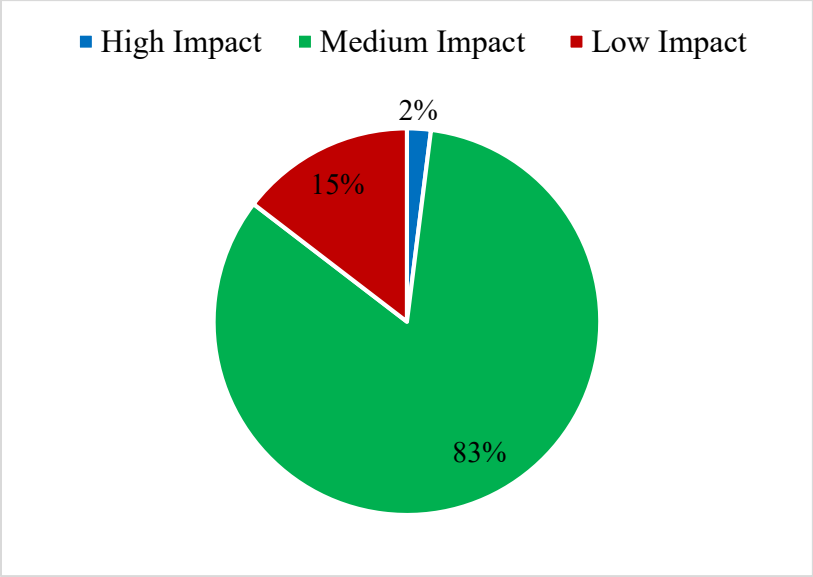


Figure 5. Clustering of Genetics Researchers by Citations and Publications

b. Ranking Analysis Using Select Bibliometric Indicators

Three scientists were randomly selected from each of the three clusters in the cluster analysis (Figure 6) to represent high, medium, and low impact groups. The results on total publication and fractional publication count, total and fractional citation count, and mean citations per paper are shown in Table 8 and represented pictorially in Figure 6. Researchers were categorized into different impact groups from the clustering analysis based on the bibliometric indicator used.

Table 8. Publication and Citation Indicator Values for Genetics Researchers

Author letter	Cluster group (impact)	Total publications	Fractional counting of publications	Total citations	Fractional counting of citations	Mean citations per paper
A	1 (high)	47	10.6	795	174.5	16.9
B	1 (high)	41	6.8	165	58.4	4.0
C	1 (high)	8	1.1	45	6.5	5.6
D	2 (medium)	5	1.2	109	18.5	21.8
E	2 (medium)	3	0.9	19	6.2	6.3
F	2 (medium)	7	1.0	133	19.6	19.0
G	3 (low)	23	2.4	1,215	119.9	52.8
H	3 (low)	126	20.9	809	142.8	6.4
I	3 (low)	42	5.5	598	64.8	14.2

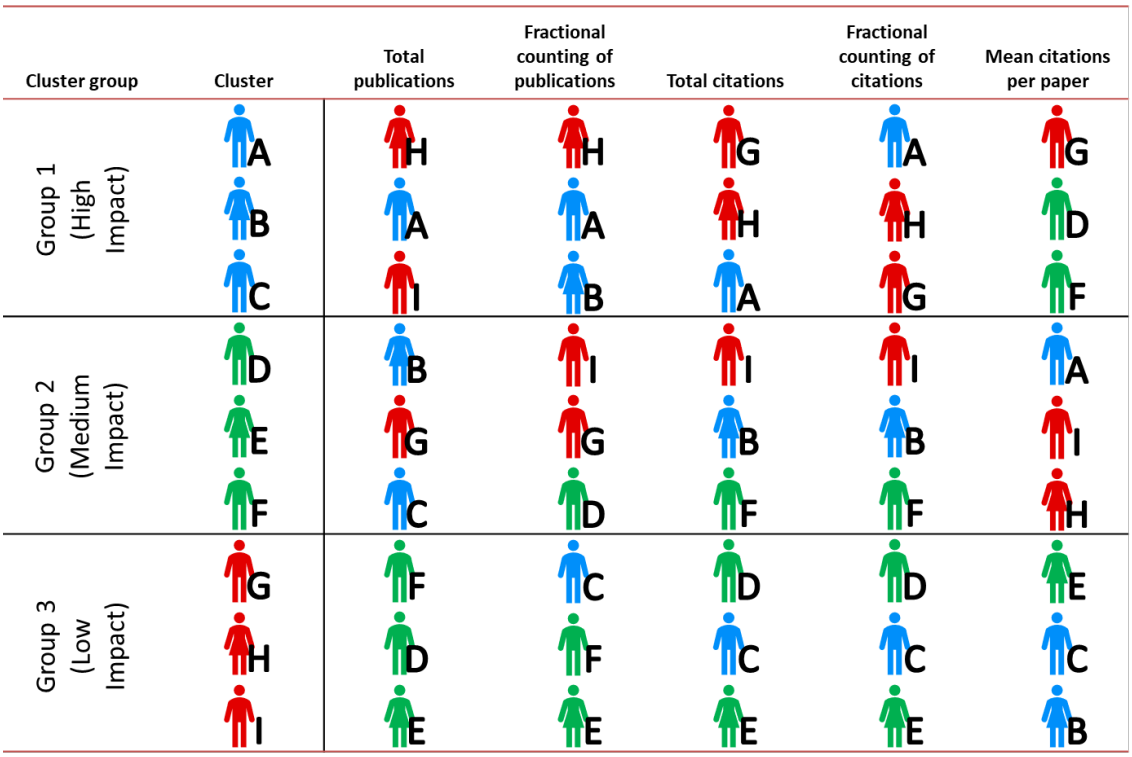


Figure 6. Ranking of Genetics Researchers Based on Select Publication and Citation Indicators

The ranking analysis was repeated using the h-index and five variations: m-index, A-index, AR-index, hc-index, and hg-index. Data are provided in Table 9 and represented pictorially in Figure 7. STPI determined that the three researchers in each influence group (column 2) received different, method-dependent rankings, as indicated by their redistribution by h-index variation (columns 3–7). Interestingly, the researchers in the low impact group were all placed in the high or medium impact groups, whereas the original high influence group was distributed across all three impact groups.

Table 9. H-index and Five Variation Values for Genetics Researchers

Author letter	Cluster group(impact)	h-index	m-index	A-index	AR-index	hc-index	hg-index
A	1 (high)	15	33	45	11	23	18.6
B	1 (high)	5	10	29	7	9	7.1
C	1 (high)	3	12	14	3	5	3.9
D	2 (medium)	4	24.5	27	4	5	4.0
E	2 (medium)	2	8.5	8	2	2	2.5
F	2 (medium)	5	15	27	5	6	5.9
G	3 (low)	16	56.5	74	15	19	19.2

Author letter	Cluster group(impact)	h-index	m-index	A-index	AR-index	hc-index	hg-index
H	3 (low)	15	27	39	10	16	19.8
I	3 (low)	14	22.5	33	11	16	17.9

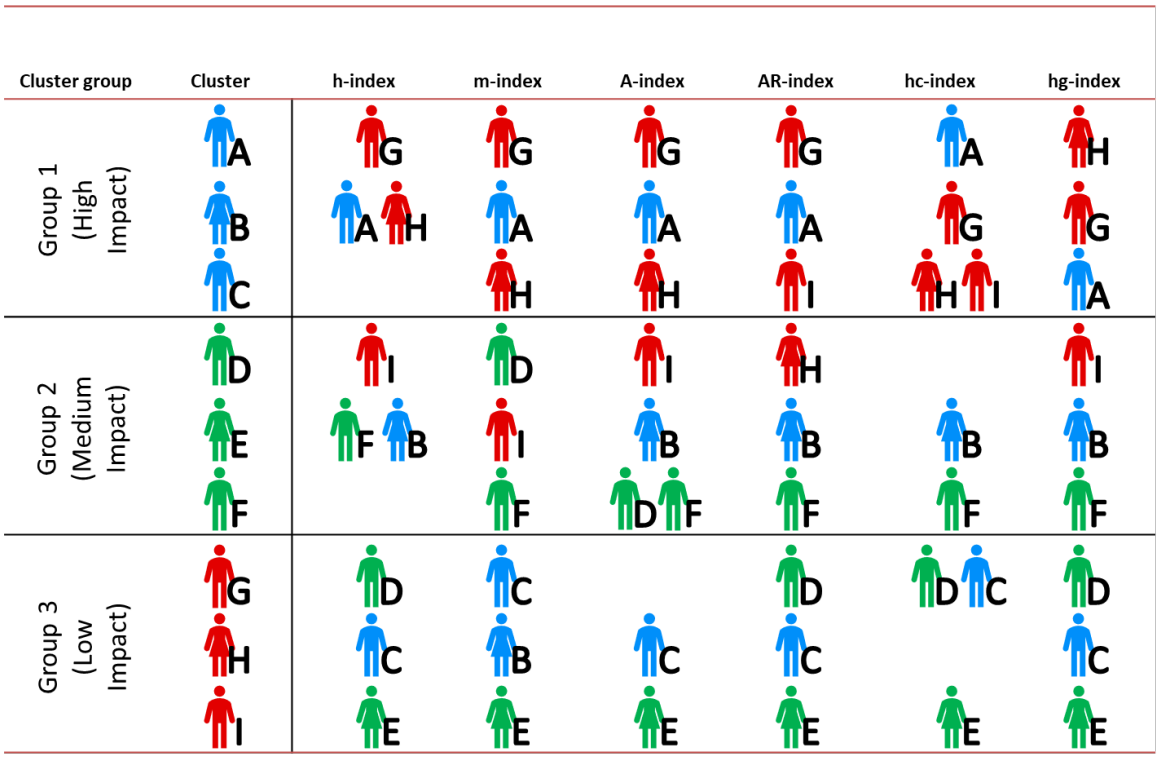


Figure 7. Ranking of Genetics Researchers Based on the H-index and Five Variations

2. Artificial Intelligence

a. Cluster Analysis

1) Publication trajectories

The cluster analysis was repeated for author profiles that listed *artificial intelligence* as a research interest area. A total of 11,650 author profiles were downloaded from Google Scholar on December 6, 2021. After data cleaning, a total of 9,505 author profiles were used as part of the pilot study on AI. From the first round of clustering based on publication trajectories, 51 of the 9,505 author profiles were grouped into the *high impact* cluster (blue); 919 were grouped into the *medium impact* cluster (green); and 8,535 were grouped into the *low impact* cluster (red; Figure 8).

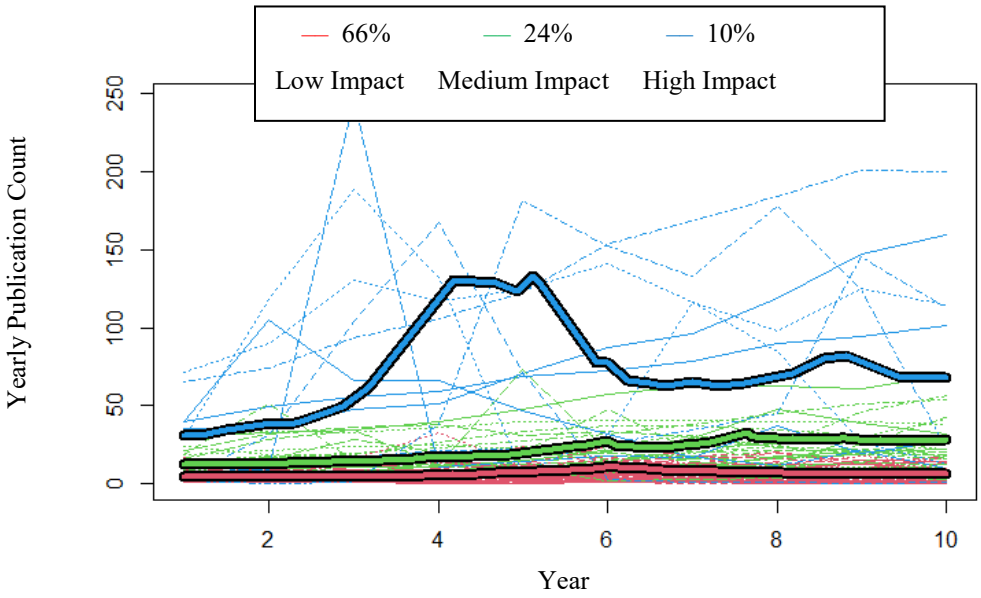


Figure 8. Clustering of Artificial Intelligence Researchers' Publication Counts over 10 Years

2) Citations trajectories

Based on clustering of citation trajectories, 14 of the 9,505 author profiles were grouped into the *high impact* cluster (blue); 54 in the *medium impact* cluster (green); and 9,437 in the *low impact* cluster (red; Figure 9). The clustering of AI researchers' citation counts over the 10 years yielded similar results to the citation count clustering of genetics researchers. Specifically, those in the *high impact* clusters for both AI and genetics separated from those in the other clusters early on and sustained a higher annual citation count over the entire 10 years.

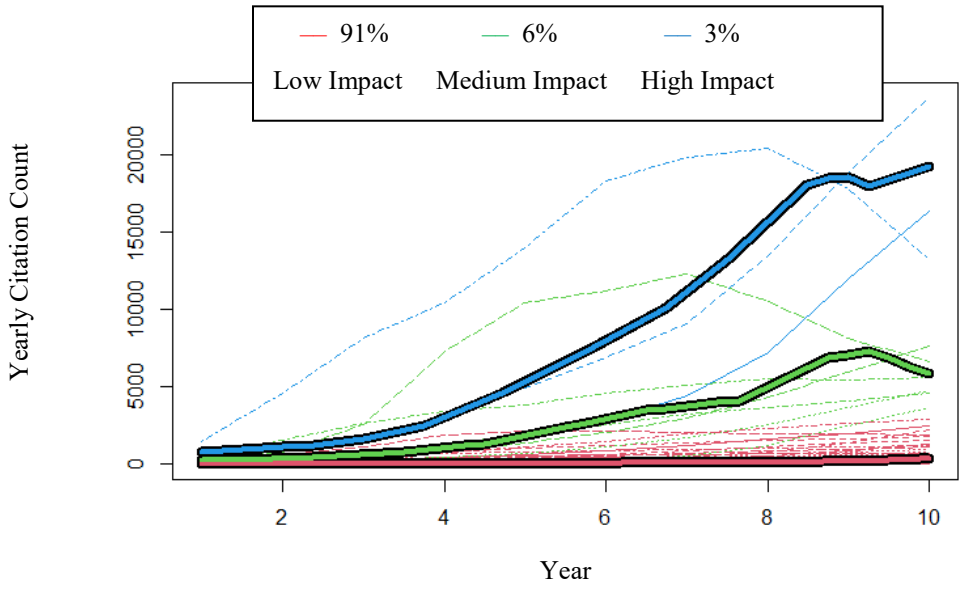


Figure 9. Clustering of Artificial Intelligence Researchers' Citation Counts over 10 Years

3) Publication and citation trajectories

After the final phase of clustering in which individual's publication and citation trajectories along with the total number of publications and citations were taken into consideration, 15 (0.16%) were grouped into the *high impact* cluster; 973 (10.24%) were grouped into the *medium impact* cluster; and 8,517 (89.61%) to the *low impact* cluster (Figure 10). Fewer researchers were selected for the *high impact* group in the AI analysis compared with geneticists, further highlighting how characteristics that make a researcher outstanding are discipline dependent and why comparisons should not be made across disciplines.

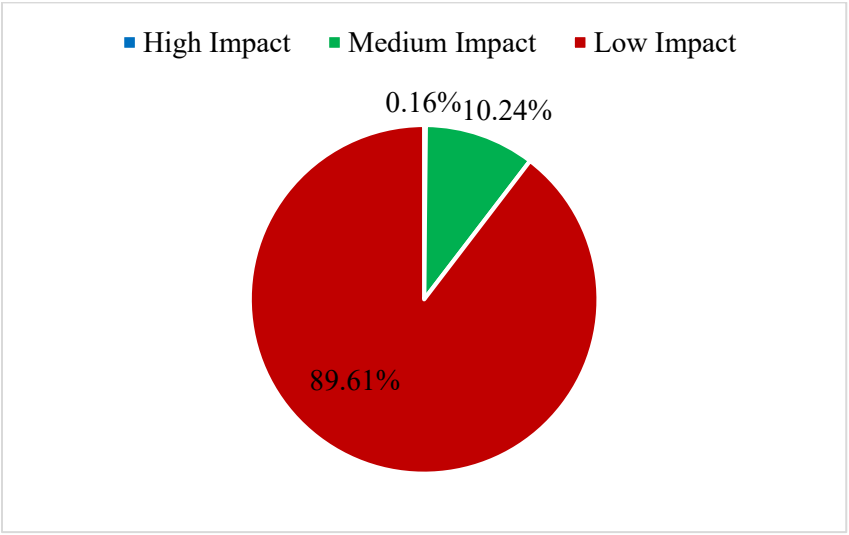


Figure 10. Clustering of Artificial Intelligence Researchers by Citations and Publications

b. Ranking Analysis

Three scientists were randomly selected from each of the three clusters in the cluster analysis to represent high, medium, and low impact groups. Data were assessed for selected publication and citation metrics and variations on the h-index.

Data for the analysis of total publication count, fractional publication count, total citation count, fractional citation count, and mean citations per paper are provided in Table 10 and represented pictorially in Figure 11. STPI determined that the three researchers in each impact group (column 2) receive different, method-dependent rankings, as indicated by their redistribution by bibliometric measure (columns 3–7).

Table 10. Publication and Citation Indicator Values for Artificial Intelligence Researchers

Author letter	Cluster group (impact)	Total publications	Fractional counting of publications	Total citations	Fractional counting of citations	Mean citations per paper
A	1 (high)	22	5.9	178	48.8	8.1
B	1 (high)	247	0.6	18,128	8.1	73.4
C	1 (high)	867	0.7	45,181	20.4	52.1
D	2 (medium)	54	16.8	229	52.7	4.2
E	2 (medium)	10	6	47	33.7	4.7
F	2 (medium)	107	23.5	338	74.2	3.2
G	3 (low)	11	3	48	13.2	4.4
H	3 (low)	3	1	34	15.9	11.3
I	3 (low)	11	3	74	22	6.7

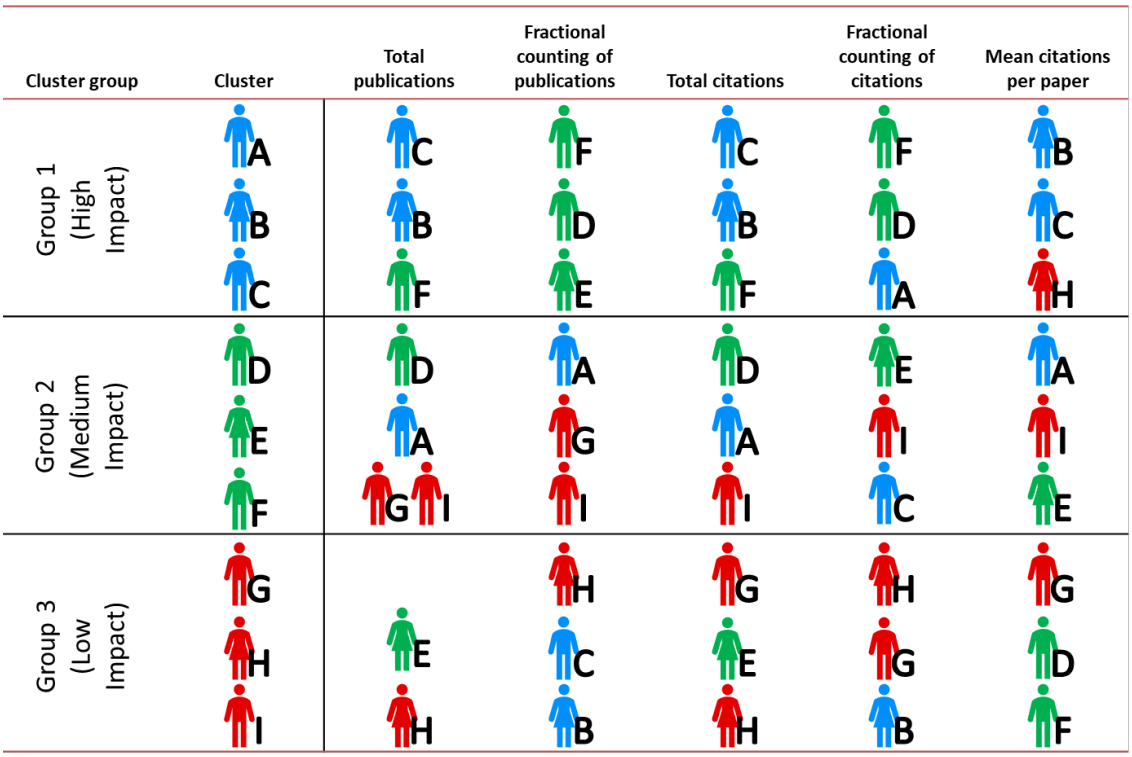


Figure 11. Ranking of Artificial Intelligence Researchers by Publication and Citation Indicators

The ranking analysis was repeated using the h-index and five variations: m-index, A-index, AR-index, hc-index, and hg-index. Data are provided in Table 11 and represented pictorially in Figure 12. STPI determined that the three researchers in each influence group (column 2) received different, method-dependent rankings, as indicated by their redistribution by h-index variation (columns 3–7). Interestingly, the researchers in the low impact group were all placed in the high or medium impact groups, whereas the original high influence group was distributed across all three impact groups.

Table 11. H-index and Five Variation Values for Artificial Intelligence Researchers

Author letter	Cluster group (impact)	h-index	m-index	A-index	AR-index	hc-index	hg-index
A	1 (high)	6	20.5	26	6	8	8.49
B	1 (high)	57	105	235	49	54	85.75
C	1 (high)	97	155	275	60	75	131.77
D	2 (medium)	9	13	17	5	11	9.95
E	2 (medium)	4	9.5	10	2	6	4.9
F	2 (medium)	11	18	18	5	14	11.96

Author letter	Cluster group (impact)	h-index	m-index	A-index	AR-index	hc-index	hg-index
G	3 (low)	4	12	10	2	6	4.47
H	3 (low)	2	16	16	4	3	2.45
I	3 (low)	6	9.5	11	4	7	6

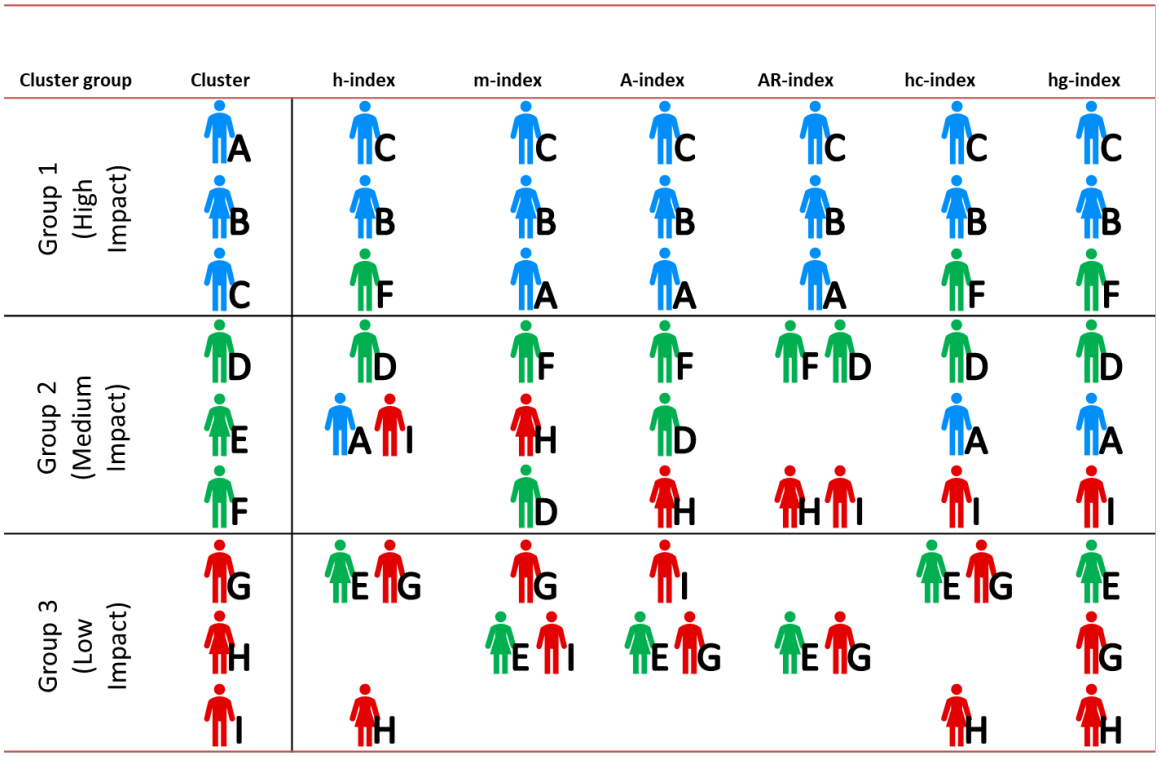


Figure 12. Ranking of Artificial Intelligence Researchers Based on the H-index and Five Variations

D. Conclusions from the Pilot Study

STPI conducted the pilot study to determine if publications, as an indicator of research productivity, and citations, as an indicator of research quality, could be used to identify patterns of research impact and whether the trajectory of publications and citations could, over time, distinguish high impact researchers from those with low and mid-level impact. The cluster analyses performed provided an alternative method to identify outstanding scientists using individuals' publication and citation trajectories.

STPI concluded that publication and citation patterns can be used to identify groups of scientists with high, medium, and low research impact. The analysis of AI as a representative physical science, and genetics, as a representative biological science, further suggest discipline-specific publications patterns. AI scientists grouped into the high impact

cluster had annual publication counts that peaked in the first 5 years of their careers and then decreased, whereas this measure continued to increase beyond the first 5 years for scientists in the genetics high impact cluster. An investigation into the early peak for AI scientist publications revealed it could be explained, in part, by publications that have very large number of authors, which may be more common in this field. These results suggest that identification of outstanding scientists may be most accurate within a scientific discipline.

STPI shows that this cluster analysis approach provides an alternative method to identify outstanding scientists. There was a high degree of discordance in the results between the cluster and ranking analysis, and within the ranking analysis when comparing bibliometric indicators. Individuals identified as outstanding, *high impact*, researchers in the cluster analysis were not strictly confirmed by the ranking analysis. The discordance between methods can be explained, in part, by use of different data sources: WoS for ranking analysis and Google Scholar for cluster analysis. To understand the differences and their implications for the pilot study, STPI compared the number of publications for the nine randomly selected AI researchers and found that the difference in publications ranged from 96% fewer to 147% more publications between their WoS profiles compared to their Google Scholar profiles. These differences in datasets preclude comparison of results between the two methods.

The pilot study also found discrepancies within the cluster and ranking results, that is, the identification of individuals as high impact scientists is dependent on the cluster or ranking method used in the evaluation. There is no baseline or widely accepted standard against which to evaluate and resolve the methodological differences. This challenge is well recognized and precludes identification of a single approach to evaluate high research impact with any degree of certainty. As a result, each bibliometric analysis should be considered only in the context for which it was performed.

There are two final considerations for the pilot study. The first addresses the STPI-chosen timeline of 10 years from first publication. Because of the high variability in the timing of a scientist's first publication, the 10-year span used for one individual is not the same 10 years for another individual. This is important because the scientific environment is constantly changing and may affect both the productivity and research impact an individual has in their first ten years. For instance, scientific collaborations have increased over time resulting in researchers having higher numbers of publications and citations than their counterparts from two or three decades ago when collaboration were not as common (Maher and Van Noorden 2021). Second, the dataset includes senior authors with long publication records. The cluster and ranking methods do not adjust for changes in scientific environment over time, such as electronic publications and increases in peer-reviewed publications that compete for citations, all factors with ramifications for the results (Kyvik 2003; Butler 2013).

Although we performed a pilot study on a limited population, other studies concur with our findings. As an example, a 2018 study compared 501 civil engineering awardees to their rankings based on the h-index and 10 variations for a total of 11 different ranking approaches (Raheel et al. 2018). The authors found that, for all 11 rankings, less than 50% of the top 10% of ranked scientists were also award winners. When rankings were divided into deciles, award winners were distributed across all 10 deciles. This report, and others, reinforces the challenge in identifying high impact scientists; however, the pilot study achieved its goals to identify patterns of research impact and whether the trajectory of publications and citations could, over time, distinguish high research impact researchers from those with low- and mid-level research impact.

5. Additional Data Sources for Consideration

As Chapters 3 and 4 have shown, the sole use of bibliometric indicators is not effective in identifying outstanding scientists. If one of the goals is to identify and retain outstanding foreign scientists who have studied or are currently working in the United States, then additional data are needed, often those involving personal identifiable information (PII), to delineate individuals based on citizenship status and country of origin. Data regarding patents, funding levels, awards and recognitions, etc., when combined with bibliometric data, can provide additional information and context when considering who is an outstanding scientist. A difficulty when combining different data sources is name disambiguation and making sure that an individual from one data source can be accurately matched to the same individual in another data source. In these instances, PII such as gender, race, ethnicity, and date of birth can help link individuals across different data sources.

In the section below, STPI identifies both existing data challenges and needs as well as additional data sources that, if combined with bibliometric indicators, may help identify outstanding scientists. The following is not a comprehensive list of additional data sources that may be linked to an individual researcher or their associated bibliometric data. Instead, it is a list of data sources that STPI identified during the course of this study that may be of interest when considering U.S. competitiveness in STEM talent broadly. The following list also showcases the variety of data that could be linked together to better assess the scientific, social, or economic impact of individual researchers, including those who are not U.S. citizens.

A. Existing Data Challenges and Needs

1. Accessing Personal Identifiable Information across Federal Agencies

Federal agencies such as the National Institutes of Health (NIH) and the National Science Foundation (NSF) collect PII such as gender, race, ethnicity, disability status, degrees and years in which they were awarded, and other information that can provide important context for bibliometric analyses or help perform investigations correlating researcher characteristics to scientific productivity and impact. However, public sources of

grant data, such as RePORTER[9] maintained by NIH, provide data only on funded applicants and contain no personal information other than their name and affiliation.

Access to individual PII for researchers external to the funding agency can be granted by undergoing a clearance process and requires a Federal sponsor to initiate the request. At the conclusion of the clearance process, an outside researcher can access internal grant databases, such as IMPACII at NIH or restricted portions of FASTLANE at NSF. However, sensitive PII, such as race and ethnicity, may further be protected by only being available to Federal employees or federally funded research and development centers (FFRDCs) such as STPI through a formal Data Use Agreement. Aggregated data on the composition of the research workforce by race, ethnicity, disability status, and degrees are available to the public (for example, from the NIH Data Book),[10] but these data cannot be linked to publications at the individual level.

One suggestion to overcome requesting PII access from each Federal agency is to establish a centralized and aggregated, interoperable Federal funding database containing a tiered system in which vetted researchers can access sensitive government data, not limited to PII, for statistical and research purposes.

2. No Centralized System to Access Federal Awards Data for Applicants and Awardees

Beyond academic publications, another important measure of a researcher's impact is the amount of research funding secured. STPI is not aware of the existence of a central repository of research funding information for academic or private sector awards, though several major Federal departments and agencies, such as NIH and NSF, maintain databases for public sector awards. However, these databases are not interoperable. There is no centralized system that can identify an individual across Federal awards databases, making the task of reliably tracing a researcher's Federal funding record difficult.

Alongside the individual department and agency award databases, the Department of Treasury hosts USAspending.gov, a publicly accessible website that tracks the majority of non-confidential Federal Government spending. While USAspending.gov contains a vast amount of information, including Federal research spending, data quality issues hamper the value of the insights drawn from these data. For instance, there are often missing or unreliable values, inconsistent naming, and different methods of data acquisition. In addition, the available data are often focused on institutions, rather than individuals, making the task of tracing funding to a given researcher difficult. Despite this, the data quality and consistency in USAspending.gov has improved since the website was first

---

[9] The NIH RePORTER site can be accessed at https://reporter.nih.gov/

[10] The NIH Data Book can be accessed at https://report.nih.gov/nihdatabook/

launched (Sage et al. 2021), and will likely continue to do so in the future. However, at the moment, USAspending.gov is not a dependable resource of Federal research funding for individual researchers.

In light of the challenges presented above, STPI recommends several additions to the practices of Federal awards databases. To mitigate the difficulty in identifying a researcher across databases, a unique Federal funding ID may be assigned to a researcher upon first submission of a funding proposal to a Federal agency.[11] This ID would follow the researcher across all Federal funding opportunities and would be entered in each agency or departmental database, making it possible to reliably trace an individual's research across databases. Federal agencies could also employ the use of ORCID IDs, discussed in more detail below. Furthermore, to facilitate data access, agencies and departments may consider a tiered system of access for certain types of information. For instance, certain types of less-sensitive information that is not currently publicly accessible may be made available to government contractors or individuals with special relationships to government organizations without the requirement of being credentialed.

B. Additional Data Sources

1. Federal Statistical Data

Federal statistical data is a robust, trustworthy source of information about individuals and businesses in the United States. The National Center for Science and Engineering Statistics (NCSES) is a Federal statistical agency located within the National Science Foundation that oversees the collection and dissemination of information on the U.S. STEM workforce and STEM degree recipients. We discuss two NCSES datasets of interest below.

Despite the promise that Federal statistical data hold, there are several laws to protect the privacy and confidentiality of individuals in the dataset that limit their usefulness. For example, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) states that "information acquired by an agency under a pledge of confidentiality for exclusively statistical purposes shall not be disclosed by an agency in identifiable form, for any use other than an exclusively statistical purpose, except with the informed consent of the respondent" (CIPSEA 2002). Essentially this statement means that Federal statistical data can only be used for activities involving describing, estimating or analyzing characteristics of groups, without identifying individuals, and "use of data in identifiable form for any purpose that is not a statistical purpose, including any administrative,

[11] National Security Presidential Memorandum 33 outlines the use and implementation for Federal agencies to use digital persistent identifiers to identify individual researchers (National Science and Technology Council 2022).

regulatory, law enforcement, adjudicatory, or other purpose that affects the rights, privileges, or benefits of a particular identifiable respondent" is prohibited without consent of the individual who provided the data (CIPSEA 2002).

a. Survey of Earned Doctorates

The Survey of Earned Doctorates (SED) is an annual census conducted since 1957 of all individuals receiving a research doctorate from an accredited U.S. institution in a given academic year (National Science Foundation 2021a). SED data are available in two formats—one for public use that is available through an interactive data tool from the NCSES,[12] and one that contains PII and requires restricted use data licensing approval by the NCSES. A list of key variables from the SED is provided below. Variables that are italicized are those that can only be accessed with restricted use data licensing approval from NCSES.

- Academic institution of doctorate
- Baccalaureate-origin institution (U.S. and foreign)
- *Birth year*
- *Citizenship status at graduation*
- *Country of birth and citizenship*
- Disability status
- Educational attainment of parents
- Educational history in college
- Field of each degree earned
- Graduate and undergraduate educational debt
- Marital status
- Postgraduation plans
- Primary and secondary work activities
- Source and type of financial support for postdoctoral study or research
- Type and location of employer
- Basic annual salary
- *Race and ethnicity*
- *Sex*
- Sources of financial support during graduate school
- Type of academic institution (e.g., historically Black colleges and universities, Carnegie codes, public or private) awarding the doctorate

Usefulness in future analyses: The SED contains information on citizenship status and country of origin for nearly every PhD recipient from U.S. institutions of higher education over the past 50+ years. If the SED data could be linked with publication records, we could assess bibliometric indicators for individuals with different citizenships or countries of

[12] The NCSES interactive data tool is available at https://ncsesdata.nsf.gov/home

origin. Additionally, we would know PhD graduation year, which would provide an alternate pathway to categorize early career researchers, and to assess their research output and research impact.

b. Survey of Doctoral Recipients

The Survey of Doctoral Recipients (SDR) is a biennial survey that has been conducted since 1973 and samples individuals[13] who have received a U.S. research doctoral degree in a science, engineering, or health (SEH) field (National Science Foundation 2021b). By using a fixed panel survey design in which the same individuals are asked to participate in the survey over time but also adding a sample of new doctoral graduates in each biennial survey cycle, the SDR provides information about the educational and occupational achievements as well as career movements of U.S.-trained doctoral scientists and engineers in both the United States and abroad. Similar to that of the SED, SDR data come in both public and restricted use formats. Public use SDR data from 1993 to present are available online, and access to restricted use data requires approval by NCSES (NCSES 2022). A list of key variables from the SDR is provided below. Variables that are italicized are those that can only be accessed with restricted use data licensing approval from NCSES.

- *Age*
- *Race*
- *Sex*
- *Ethnicity*
- *Citizenship*
- *Place of birth*
- Educational history
- Employment status
- Field of degree
- Occupation
- Job satisfaction
- Reason for changing employer or job
- Factors important in deciding to come to the United States

Usefulness in future analyses: Because the SDR tracks employment information of U.S. PhD recipients over time after graduation, we could identify and study particular cohorts of individuals that work in a particular sector (different types of academic institutions, private sector, and government). By tracking the bibliometric or other outputs (e.g., patents, awards, jobs) of these individuals by their post-PhD employment sector, we could better develop metrics of outstanding scientists outside of those in academia.

c. Similar International Statistical Datasets

A number of countries have their own efforts to track post-PhD careers, but we were unable to identify anything as well established and continuous as the NCSES surveys

[13] Individuals must be less than 76 years of age at time of survey to participate.

described above. Many countries do have centralized statistical systems and use their population surveys to capture some of these measurements, but many of these systems do not have the flexibility to answer the specific questions that SED and SDR allow. Here we highlight the broadest similar international data collection effort that we could identify.

In 2004, the Organisation for Economic Co-operation and Development (OECD), together with UNESCO and Eurostat, established the Careers of Doctorate Holders (CDH) initiative, which "set out to develop internationally comparable indicators on the careers and mobility of doctorate holders" (OECD 2019). The CDH would provide similar insights to the NCSES SDR. This effort was challenging in that it required an ability and understanding of how to combine and compare data from each data holder/country. Some countries fielded a dedicated CDH survey, while others derived the relevant information from their national labor force survey or from administrative data sources. The first CDH data collection was coordinated in 2006 with 25 countries participating, and these efforts continued until 2017, when they were paused for review by OECD to review resource prioritization. While the aggregate statistics provided in public release data files are useful to inform policymaking, as with the SED and SDR, to answer questions about research impact of individual researchers would require access to the individual level, likely PII, data. We were unable to determine if there are CDH restricted use data that are accessible to external researchers in a similar way that SED and SDR data are.

2. Private Third-Party Data

Many non-governmental organizations hold and curate databases of potential value to supplement our bibliometric analyses. Some of these organizations are for-profit entities, and others are non-profit organizations. Each entity listed here aggregates data from other sources, whether by web-scraping, partnering, or other means. In most cases, the curated datasets produced or accompanying analytical tools are available to researchers for particular analysis or insights for a fee.

a. IRIS

The Institute for Research on Innovation & Science (IRIS) at the University of Michigan has created a linked data infrastructure that enables research into the research impacts and outputs of U.S. research activities. The core dataset in the IRIS research dataset is administrative data (Federal and non-Federal sponsored projects, employment information on post-docs, graduate students, and undergraduate students, vendor spending, and subcontracts) from research universities that join IRIS as members. Though membership in IRIS continues to grow, in 2021, the IRIS dataset encompassed over 80 member campuses and represented over 40 percent of total U.S. R&D spending at universities. The administrative data can then be combined with other data sources, including NIH or NSF award data, Census data, dissertation data, data from Steppingblocks

(see below), and a new pilot is exploring possible connections to NCSES data, such as the SED.

Usefulness in future analyses: Based on IRIS's integrated data infrastructure, it could be possible to paint a much more complete picture of the factors statistically linked to a researcher becoming an outstanding scientist, early or late in their career. For example, there is the potential to show correlations between Federal funding levels, or demographic characteristics and research impact. This information could be useful, perhaps less in the identification of outstanding scientists, but in identifying areas where the U.S. Government could potentially improve the racial or gender diversity of outstanding scientists in a particular field, and thereby possibly create a wider base of domestic-grown outstanding scientists.

b. Steppingblocks

Steppingblocks collects, curates, and combines demographic, education, and workforce data to provide insights about education and workforce outcomes (Steppingblocks 2021). Data sources include online profiles, resumes, university websites, public databases, job postings, public filings, public salary databases, and government sources. The company uses machine learning and AI among other methods for data validation and data analyses. Steppingblocks has data on over 130 million individuals in the United States and has several research partnerships funded by Federal agencies. For instance, Steppingblocks, in conjunction with Clarivate, recently received an NSF grant to study the impact of foreign-born scientists and engineers. Select variables from Steppingblocks data include:

- Gender
- Age
- Job information (e.g., title, category)
- Contact information
- Employer information (e.g., employer name, industry)
- Salary
- Skills and certifications
- Education background (e.g., degree level, field, graduation year)
- Location

Usefulness in future analyses: Because Steppingblocks tracks individuals over their career progression, if this information was combined with publication records, we could develop greater insights into the factors associated with someone's rise as an outstanding scientist. For example, we could identify if individuals transitioned between roles in academia, or moved from academia into industry or government service. It is also possible that by tracking individuals who took these different career paths, we could better identify characteristics of outstanding scientists aside from bibliometric indicators.

c. Moody Analytics: Orbis Database

Orbis is a database containing corporate ownership, intellectual property, and financial information on more than 400 million companies worldwide (Orbis 2022). In addition, the database contains information approximately 350 million people globally including 162 million shareholders and 156 million owners. The database also links approximately 115 million patents to about 300 million companies and has information on about 13 million patent transactions. The Orbis database contains information on:

Companies	Patents
• Corporate structure	• Patent transactions
• Mergers and acquisitions	• Valuations
• Standardized financials	• Ownership timeline
• Industry codes	• Legal status updates
• Technology classifications	

Usefulness in future analyses: It is possible that we could use the Orbis database to identify outstanding scientists based on their economic impact, rather than their research impact. We could search Orbis for a group of scientists in a particular field and identify those with patent or company ownership, and trace the economic impacts (e.g., revenues) associated with that ownership.

d. ProQuest Dissertation and Thesis Database

The ProQuest Dissertation and Thesis Database (PQDT) is the largest global repository of dissertations, and was designated as the official dissertation repository by the U.S. Library of Congress (ProQuest n.d.). At present, it contains over 5 million dissertation citations and over 2.7 million full-text theses or dissertations, with records going back hundreds of years. While the largest number of entries in the database comes from U.S. institutions, there are over 100 countries and 3,100 institutions represented across the database. When searching PQDT, one can query fields including title, abstract, author, advisor, institution, keywords, and subject/department, among other entries.

Usefulness in future analyses: Because PQDT contains information on dissertation submission year, which generally coincides with PhD completion, if this information on graduation year could be linked with bibliometric information, we could conduct an analysis of early career outstanding scientists using graduation date (rather than first publication date) as a starting point for "early career." Using PQDT could also allow us to identify individuals whose PhD dissertation was in a particular research area (e.g., quantum information science or AI), which is not possible to assess from datasets such as the SED or SDR. Our current bibliometric study involves using the author's current research area

to group them with others in a similar field, but the author's current research area may differ from their PhD research area and there may be value in sorting by PhD research area, specifically if we are interested in better understanding early career researchers.

e. ORCID

ORCID is a not-for-profit organization aimed at increasing the connections between researchers; their contributions, funding, and other awards; and their affiliations through unique, persistent identifiers (ORCID 2021). Scientists around the world are able to create a free, unique, persistent identifier that can be used to clearly attribute themselves and contributors to scholarly research and outputs. ORCID enables funders to use their ORCID IDs in their funding workflows to connect grantees to the grants they have been awarded, publishers, and research organizations. ORCID's database can be used for name disambiguation and provides a clear, transparent history of a scientist's research and career trajectory. Key variables available through ORCID include:

- Unique, persistent identifiers for each individual
- Education background
- Institution affiliations through time
- Publications
- Peer review activity
- Society membership
- Funding
- Awards and distinctions

Usefulness in future analyses: If Federal agencies required the use of ORCID IDs as part of the application process, then applicants and awardees could be identified across different Federal awards databases. The continued use of ORCID IDs by both funders, publishers, and researchers would allow for the evaluation of researchers' careers through time. A researcher's ORCID profile can provide a longitudinal record of their educational, work, and funding history; professional accomplishments; and research outputs.

3. Citation Databases

a. Web of Science

WoS is a literature database by Clarivate that covers roughly 34,000 journals, books, and conference proceedings (Birkle et al. 2020). These documents are an accumulation of multiple databases including WoS's *Core Collection*; international databases from China, Russia, Latin America and Iberia, and Korea; and other specialized indexes such as Medline, Zoological Record, and patents. However, the *Core Collection* makes up roughly half of the WoS database. There are roughly 182 million records (journals, books, and proceedings) and over 99 million patents (Clarivate 2021a).

Access to WoS resources is divided into three tiers: the WoS platform tailored towards general use,[14] application programming interfaces (APIs), and raw data provided by Clarivate. In the WoS platform, which was used in the above pilot study, the search engine allows users to look up data using keyword fields or authors by first and last name. One of these fields includes "Web of Science Categories," which contains the 254 topic areas that publications can be sorted under. Publications can be categorized under multiple topic areas. Publication exports include various metadata, such as language, document type, and ORCIDs where applicable.

b. Scopus

Scopus is an abstract and citation database by Elsevier with over 77.8 million records as of January 2020, which include journals, books, conference proceedings, and trade publications. These records are derived from more than 25,000 active serial titles and 210,000 books (Scopus 2022). Similar to WoS, Scopus has a search engine that allows users to search by various keyword fields, author first and last name, and, additionally, affiliations. These metadata can also be downloaded in large batches into Excel documents. Furthermore, Scopus also automatically generates author profiles based on automatic name disambiguation and currently has 16 million author profiles.

c. Google Scholar

Google Scholar is another literature database that has open access to a wide range of literature. Unlike WoS, which only contains published materials, Google Scholar scrapes online resources to collect both published materials such as articles and books and non-published materials including conference proceedings, patents, software presentations, and white papers in both academic and non-academic areas. Because Google Scholar includes a wider range of potential documents with minimal manual review, the database often includes a greater number of publications and citations than WoS and Scopus. Because Google does not disclose how many documents Google Scholar contains, multiple studies have attempted to estimate the breadth of coverage. Two studies conducted in 2017 and reviewed by Delgado López-Cózar et al. estimated a range of 194 to 331 million articles, citations, and patents in Google Scholar (Delgado López-Cózar et al. 2018).

A comparison of the Web of Science, Scopus, and Google Scholar can be found in Appendix G.

[14] The general use WoS platform is available at https://www.webofscience.com/wos/woscc/basic-search

6. Summary and Next Steps

A. Summary

STPI addressed the following two research questions in this study:

1. *Can bibliometric indicators be used to accurately identify outstanding scientists within a scientific discipline through time, and what are the limitations of its use?*
2. *Can new bibliometric indicators or analytical approaches be developed to identify outstanding scientists within a scientific discipline through time?*

STPI's review of the literature showed that there is no universal set of bibliometric indicators to identify outstanding scientists. In addition, there are many limitations in the use of bibliometric indicators such as field dependence; inability to accurately predict the research impact of early career researchers and non-traditional academicians; and perpetuation of gender biases.

STPI's pilot study demonstrated that cluster analysis may provide an alternative approach to identify outstanding scientists. Cluster analysis identified three groups of scientists that, based on their publication and citation trajectories, could be categorized as having high, medium, and low research impact. The discipline areas investigated—AI and genetics—suggested interesting science discipline-specific patterns for research impact. further testing using additional data sources and larger sample sizes is needed to improve the approach.

B. Next Steps

STPI anticipates that a combination of strategies where bibliometrics is one component represents the most promising direction for identifying outstanding scientists with any degree of certainty.

1. Additional Bibliometric Analyses

a. Undercited "Sleeping Beauty" Research

When using citation and publication information as indicators of academic success, it is worth considering that these metrics do not always properly reflect the value of the work.

In particular, certain academic publications fall under the category of "sleeping beauty"[15] publication, a term that refers to *undercited* published works that accrued fewer citations and notable mentions in the early part of their existence than hindsight grounded in their later success would suggest they should have. Du and Wu (2018) describe two general categories of such works: (1) publications that are part of a cumulative process or a combination of several discoveries that are difficult to appreciate on their own, and (2) publications that disrupted established paradigms and took time for the field to catch up. One example of a transformative sleeping beauty publication is Gregor Mendel's *Experiments on Plant Hybridization*, which was met with skepticism or outright rejection by the scientific community at the time but is now considered a seminal work that transformed the field of genetics.

The concept of the sleeping beauty has been recorded in the literature under various names since the 1960s (e.g., Barber 1961 and Wyatt 1962), and has been described qualitatively by several authors (see Garfield 1980 and Costas et al. 2011 for notable examples). Van Raan (2004) was one of the first to provide an empirical definition of a sleeping beauty as a published work that had a long period of sleep after publication (e.g., 10 years) marked by a small number of annual citations, followed by a sudden period of resurgence. This resurgence is indicated by a stark increase in the number of citations, called the "awakening," which is often facilitated by a "prince"—a personification of an important citing article or other published work, such as a patent (van Raan 2015).

Several authors have suggested potential reasons for the occurrence of the sleeping beauty phenomenon—Wyatt (1962) reasoned that some discoveries have a delay in recognition due to technical limitations in the field that do not allow the discovery to be implemented. Cole (1970) suggested that certain published works may fall victim to the *Matthew's Effect*—for instance, the author's status in the social hierarchy of their field does not allow the work to proliferate in the same way as it would if it were published by an author of higher status. Alternatively, Cole suggested that the recognition may be initially lacking due to the published findings not agreeing with the commonly accepted notions of the field. Stent (1972) provided a closely related reasoning: although a particular finding may be well received, its full recognition is delayed due to the inability to initially find a connection between it and the canonical understanding in the field. Finally, Price (1976) argued that a sleeping beauty can occur when the finding is poorly communicated, such as when the explanation is overly technical or lacking in clarity.

Contemporary methods for identifying sleeping beauty published works mainly build on the work of van Raan (2004). Typically, these approaches weigh various factors—such

---

[15] This phenomenon has been discussed under several different names, such as: existed discovery (Barber 1961), premature discovery (Stent 1972), delayed recognition (Cole 1970, and Garfield 1970, 1980, 1989), Mendel syndrome (Costas et al. 2011).

as number of citations during the dormant period and the rise in the number of citations during and after the awakening period—and attempt to either categorize the paper as a sleeping beauty (van Raan 2004) or quantify the strength of the sleeping beauty signal for a particular published work (Ke et al. 2015). Although the approaches can detect sleeping beauties according to their respective definitions, they are only capable of doing so retrospectively. To STPI's knowledge, no article has outlined the conditions for the early detection of a sleeping beauty. At the same time, van Raan (2015, 2017, 2018) noted that patent citations can be an early indicator of a sleeping beauty and although the detection a patent citation may take fewer years than the traditional dormant-to-wakeful trajectory of a sleeping beauty, this approach still requires the potential sleeping beauty to rest for several years before it can be recognized. In addition, Hou and Zhang (2020) have successfully detected sleeping beauty articles via altmetrics, even going so far as to outperform traditional citation-based detection methods in some cases. Yet, the altmetrics approach is not broadly applicable as it is necessarily limited to publications that are hosted only on platforms that report altmetric scores as well as fields where altmetrics can gain traction as a method of scientific dissemination.

Despite the relative difficulty in detecting sleeping beauties, these types of published works require attention when considering the value of an academic career through the lens of publication and citation statistics. In fact, some of the most transformative scientific research has followed the trajectory of a sleeping beauty in the past and is likely to do so again in the future. The ability to detect and properly account for the value of a sleeping beauty is a crucial next step in evaluating researchers' contributions to science. More research is necessary to understand the common characteristics of sleeping beauties in the early part of their lifecycle, alternative methods of detection beyond citation trajectories, and any differences in sleeping beauty characteristics between fields.

b. RCR

The Relative Citation Ratio (RCR) was proposed in 2016 by staff at NIH (Hutchins et al. 2016). The goal of RCR was to correct the known limitations of existing measures of scientific productivity and impact, which do not account for differences in publishing norms across fields, undervalue collaborate work, and emphasize publication quantity over quality.

The innovation of RCR is that it is field- and time-normalized and is benchmarked to a typical NIH paper in the same publication year. RCR of an article is defined as the citation rate of the article (number of citations divided by years since publication) divided by the average impact factor (IF) of journals in which co-cited papers appeared. Co-cited articles are those cited by papers that also cite the article whose RCR is being calculated. If n articles from a given journal are co-cited, that journal's impact factor is included n times in computing the average IF. An RCR value of "1" means that the paper is cited as often

as expected based on the NIH norm, and the higher and lower values indicate that the paper was cited more or less than its "peers," respectively.

Some subsequent studies of RCR concluded that it was effective as a measure of research yield (e.g., Patel 2021), and in general the metric is considered a step forward in quantifying influence across fields (e.g., Ioannidis 2016). Further work extended RCR beyond the biomedical field to other subject areas (Purkayastha 2019). However, like other bibliometric indicators, RCR also has limitations. The original article proposing RCR acknowledged that it cannot adequately measure the influence of very recent papers. A 2017 study questioned the validity of the metric by pointing out the weaknesses of the normalization strategy (Janssens et al. 2017). Finally, a 2016 study found that RCR was not well correlated with expert assessment (Bornmann 2016).

Despite some skepticism of RCR, it is a measure worth exploring for identifying outstanding scientists. For example, additional studies correlating clusters with RCR values could help refine the clustering method.

2. Verification of Outstanding Scientist Status

Numerous methods have been used to identify outstanding scientists in their fields. Companies like Clarivate produce an annual list of the most highly cited researchers[16] based on a variety of factors including their citation count and the number of papers an individual has that is considered highly cited (Clarivate 2021b). The challenge, however, lies in deciding which method is producing the most accurate list of true outstanding scientists within a scientific field. Verification is a challenging task as there is no true list of individuals that are considered outstanding in their fields. In addition, because there is no universally agreed upon concept of what constitutes an outstanding scientist, one method valuing a certain set of factors in its selection criteria will generate a different, though likely overlapping, list of outstanding scientists than another method that has prioritized a different set of criteria. This does not mean that verification should not be attempted. Rather, it means that until a rigorous verification method is developed, there is no way of knowing or determining which method is better at identifying outstanding scientists. To address this, STPI suggests two strategies that could be pilot-tested to evaluate the accuracy of the method described in this report. As each has its own strengths and weaknesses (Table 12), it is desirable to use multiple methods in parallel to refine the algorithm.

Expert validation. Expert panels are commonly used to evaluate research quality and investigator's scientific potential; furthermore, RCR was validated using data from expert panels. The key weakness of expert panels is their subjectivity, conservatism, and potential

[16] The most recent list of highly cited researchers released by Clarivate was for 2021 and can be accessed at the following website: https://recognition.webofscience.com/awards/highly-cited/2021/.

for implicit and explicit bias. Reviewers can also be in conflict with researchers they are evaluating. However, expert opinion remains a cornerstone of evaluating scientists, and various safeguards are being introduced to reduce or eliminate biases and improve review fairness and robustness (Guthrie 2019). Therefore, a well-executed expert panel remains a powerful tool for assessment of scientific products and for validating bibliometric measures derived from these products. Typically, panels are asked to rate individual papers or portfolios of work using a numerical scheme established for this purpose. Crowd-sourcing the scientific community to identify the top researchers in their fields could be used as an alternative to rating candidates in a pre-determined sample. While free of some limitations inherent in the first option, this strategy may fail by not yielding a consensus group of individuals.

Using reputation indicators. Another option for validating bibliometric data is to use proxy measures of standing in the community, such as prizes, service on editorial and advisory boards, funding records, and invitations to give congressional testimony. STPI acknowledges that these indicators have limitations: they favor established over early career scientists, are not independent of each other (e.g., one prize is likely to lead to another), are field-specific (e.g., there is no Nobel Prize in mathematics), and few indicators are universally accepted as marking an exceptional researcher (e.g., Nobel Prize or Fields Medal). Nevertheless, this strategy has potential for validating a group of researchers identified through bibliometric analysis and an aggregate score based on a combination of several reputational indicators could improve its accuracy. STPI notes that these data could be labor intensive to collect for a large sample.

Table 12. The Pros and Cons of Two Validation Methods

Validation Method	Strengths	Weaknesses
Expert validation	Allows integration of various viewpoints to reach consensus, universally accepted, powerful if well implemented, easy to implement	Subjective, difficult to implement on a large scale, potential for conflicts of interest, may not reach consensus, dependent on the composition of the panel
Using reputational indicators	Relies on multiple measures, easy to implement, objective	Measures not independent of each other, limitations of individual indicators, labor intensive on a large scale, favors established scientists, field dependent

3. Identifying Outstanding Teams of Researchers

Identifying individual outstanding scientists underestimates the importance of collaboration and team contribution for scientific discovery. The full value of a researcher may not be fully realized without the element of collaboration (Dong et al. 2018).

Therefore, identifying outstanding teams of researchers (those who have produced high-quality scientific output collaboratively) represents another promising avenue of investigation.

Prior research on research collectives has often employed network analysis as a tool for understanding the relationships between researchers and quantifying their impact (Bales et al. 2008; Abbasi and Altmann 2011; Digiampietri and da Silva 2011; Hicks et al. 2019). Network analysis casts the data as a web of interconnected nodes held together by edges; in a hypothetical network graph of publications for a particular research topic, the publications form the nodes, and citations constitute the edges. The topography of the network can be used to compute a number of indicators that quantify the influence of a node within the network, demonstrating, for instance, how well connected a particular node is with respect to the remaining nodes.

In terms of the data used to identify impactful research groups, studies have often centered on publication-level information. For instance, a 2017 study used a network analysis of different publications (e.g., journal articles, clinical trial reports, patents, FDA medical reviews) as well as research award, grant, and authorship data as a proof-of-concept for a multi-measure evaluation for a research topic or trend (Keserci et al. 2017). The study demonstrated several measures for highlighting particularly important publications and impactful authors whose work indirectly led to the development of the therapeutics. This approach can be extended to identify impactful research groups by capturing the sub-networks around a particularly impactful author or all authors in the vicinity of an important publication. The impact of the authors captured in this manner can be further validated by cross-referencing the supplementary research award and grant data, along with any other related measures of success that are traceable at the author level.

However, when considering the complexity of a network of research products, it may be difficult to clearly delineate a research group based strictly on proximity to an impactful author or publication. An alternative approach is exemplified by the Map of Science tool developed by Center for Security and Emerging Technology (CSET), which was developed from a network of published products that were clustered based on citation linkages (Rahkovsky 2021). CSET has also provided an example case of identifying outstanding research products through the use of adjusted publication and citation metrics within the Map of Science (Acharya & Dunn 2022). Taken together, outstanding research groups may be distinguished by drawing on the Map of Science framework by first locating outstanding research products via adjusted metrics, determining the cluster of origin of the outstanding research product of interest, and identifying the researchers responsible for the published works within that cluster.

The network models listed above draw on large-scale datasets, involving a multitude of research areas; an analysis of talent from such a dataset may potentially result in a long list of outstanding researchers from various disciplines. A 2019 study suggested one

method to reduce the number of disciplines is to gauge the degree of emergence for each research topic in the sample (Porter et al. 2019). The study assigned "emergence scores" to research topics, measured by the novelty, momentum, growth, and community size for a given research topic algorithmically discovered in a series of research publication and patent abstracts. The emergence scores agreed with external validation metrics, and were shown to be representative of the state of the science for the chosen research topics. While it is a fairly new method, the emergence scoring algorithm shows promise as a tool for identifying research topics slated for growth, and as a supplementary tool for identifying outstanding research groups studying those topics.

Appendix A.
Context for Social, Economic, and National Security Impacts

Research impact is the demonstrable effect of an individual's scientific contributions as assessed through advances in a research field or to general scientific knowledge (research impact), and contributions to the general economic and social capital of the nation (economic impact, social impact), all of which have implications for the security of the nation (national security impact). Research impact was considered in the text of the report, and for completeness, here we review social, economic, and national security impact.

Social impact is the effect of research outputs on advancing policy decisions, public debate, and the general social capital of the nation. Social impact measures include number of references to the individual in public debate and public policy decisions, altmetrics, Academic Rigour and Relevance Index (AR2I; Phillips et al. 2017), social media presence, mass media mentions, non-science awards and honors, holding a government advisory position, soft influence (e.g., Sarah Gilbert, an Oxford University professor who helped lead the development of the Oxford/AstraZeneca coronavirus vaccine, having a Barbie doll modeled after her; Nuñez 2021), and participating in legislative planning meetings (Bornmann 2012). Researchers who hold a more recognizable public presence are more likely to have impacts beyond the academic sphere, and influence the general public understanding of science.

Economic impact is the effect of research outputs on advancing economic trends and the general economic capital of the nation. Economic impact measures include number of patents, number of industry projects, external funding relating to research cooperation with non-academic institutions, and start-up companies (Wilsdon et al. 2015).

National security impact is the effect of research outputs directly on science and technology that protect, or influence the protection of, the nation (Sarkesian et al. 2008) or indirectly through research that impacts the economy and social structure and capital of the nation. Dual use research—research that has potentially beneficial and detrimental uses—is a primary national security concern. The complexity of the direct and indirect impacts of research on national security makes identification of scientists with outstanding research impacts especially challenging.

Appendix B.
H-index Variation Tree

The following evolution tree is derived from a selection of the 85 variations highlighted by Bihari et al. (2021), which illustrates the multitude of variations that have been developed by scientometricians. Each evolution of variation is denoted in the green boxes while the mathematical motivation is denoted in the blue boxes.


h-index
Account for highly cited pubs
g-index
Account for pubs with average citations
h(2)-index
Require less strict citation count
w-index
Adds weight to least-cited pubs
f-index
Give less weight to least-cited pubs than f-index
t-index
Balance h- and g-index effect on high-hit papers
hg-index
Account for entire corpus & citations
tapered h-index
Account for authors w/ identical h-indices but diff pub & citation counts
A-index
Remove possibility of skewing by 1–2 influential pubs
R-index
Remove penalty for long h-cores
Rm-index
Account for research career age of scholars
m-quotient
Add weight to recent/early career pubs
AR-index
Add weight to citation of recent pubs
Contemporary h-index
Remove penalty for citing old pubs by measuring recent citations
Trend h-index
Account for change in h-index over time via size in core pubs
h_d-index
Use decade rather than arbitrary time cutoff
h_{pd}-index
Allow for varied time window
timed h-index
Adjust for pub timing within career via year-wise productivity
Career year h-index by publication
Account for scientific impact via citations
Career year h-index by publication year citation
Consider total # citations/year-wide impact
Career year h-index by citations
Account for co-authorship via authors' h-indices
$\bar{h}$-index
Remove penalty of coauthors with insufficient h-values
h_α-index
Penalize for self-citations
Discounted h-index
Adjust for collaboration in h-core
h_i-index
Adjust for skewing by a few articles with high co-authors
pure h-index
Adjust for variation in contribution
normalized h_i-index

Appendix C.
H-index and Variation Calculations

To calculate the h-index of a hypothetical Author A, the rank value of each of Author A's publications by their citation counts is assigned from biggest to smallest. A chart of hypothetical Author A's corpus of work is seen in Table C-1.

Table C-1. Author A's Start Data

Title	Publication Year	Rank	Citation Count	
Paper f	2012	1	200	
Paper q	2009	2	65	
Paper b	2016	3	24	h-core
Paper g	2004	4	12	
Paper u	2005	**5**	**7**	
Paper o	2007	6	3	
Paper t	2007	7	2	
Paper w	2019	8	1	
Paper k	2020	9	1	
Paper l	2021	10	1	
Paper e	2021	11	0	
Paper v	2005	12	0	

As a reminder, the h-index is defined as:

"A scientist has index h if h of his or her Np papers have at least h citations each and the other (Np – h) papers have ≤h citations each."

In other words, the h-index is the highest rank magnitude that is less than or equal to its number of citations. Another term to note is the h-core. The h-core is all papers that are "considered" in the h-index. In this case, the h-core are the publications ranked 1–5 since these are *within* the h-index calculation.

STPI selected the following five potential indices to predict outstanding scientists: A-index, m-index, AR-index, contemporary h-index, and hg-index.

A-index:

The A-index is the average of the h-core citations. Author A therefore has an A-index of 61.6.

m-index

The m-index is the median of the h-core citations, so Author A has m-index of 24.

AR-index:

The AR-index is an adjustment of the A-index that incorporates the age of a paper, where more recent papers are given more weight than older papers. The AR-index is defined as:

$$\sqrt{\sum \frac{citations\ of\ an\ hcore\ paper}{age\ of\ paper + 1}}$$

where the citation:age ratio is calculated for each h-core paper. The age is calculated as the current year minus the year of publication + 1.

For example, *Paper f* has a ratio of:

$$\frac{200}{(2021 + 1 - 2012)} = 20$$

This ratio calculation is repeated for all h-core papers (in this case *papers f, q, b, g,* and *u*), the ratios are summed, and then the square root is taken of the total sum. The AR-index of Author A is, therefore, 5.48.

The +1 to the current year avoids a divide-by-zero error. Table C-2 contains an age column and an AR-value column (pre-summing and pre-square root).

Table C-2. Author A's Data with AR-index Numbers

Title	Publication Year	Age	Rank	Citation Count	AR-value
Paper f	2012	10	1	200	20
Paper q	2009	13	2	65	5
Paper b	2016	6	3	24	4
Paper g	2004	18	4	12	0.666667
Paper u	2005	17	5	7	0.411765
Paper o	2007	15	6	3	0.2
Paper t	2007	15	7	2	0.133333
Paper w	2019	3	8	1	0.333333
Paper k	2020	2	9	1	0.5

Title	Publication Year	Age	Rank	Citation Count	AR-value
Paper l	2021	1	10	1	1
Paper e	2021	1	11	0	0
Paper v	2005	17	12	0	0

Contemporary h-index

The contemporary h-index (h^c) has a similar calculation to that of the AR-value but is calculated like the h-index.

First, each publication is assigned an S^c value, which is calculated exactly like the AR-value [citation/(age+1)] but multiplied by a scaling coefficient. In the original publication, the authors used a coefficient of 4 so that papers published during the current year are multiplied by 4, papers published 4 years ago have a weight multiplied by one, and papers published 6 years ago have a weight multiplied by 4/6. The selection of 4 as the coefficient appears arbitrary, so another value may be used instead (e.g., 5 or 10 since platforms like Google Scholar focus on publications in the last 5 or 10 years when calculating metrics).

The papers are then re-ranked based on their S^c value. Table C-3 contains the updated S^c column. The h^c is then calculated the same way as h-index where h^c is the largest rank magnitude that is less than or equal to S^c. In this case, [rank = S^c]. Therefore, Author A has h^c value of 4.

Table C-3. Author A's Data with Contemporary H-index Numbers

Title	Publication Year	Age	Rank	Citation Count	AR-value	S^c value (coef. 4)
Paper f	2012	10	1	200	20	80
Paper q	2009	13	2	65	5	20
Paper b	2016	6	3	24	4	16
Paper l	2021	1	4	1	1	4
Paper g	2004	18	5	12	0.666667	2.666667
Paper k	2020	2	6	1	0.5	2
Paper u	2005	17	7	7	0.411765	1.647059
Paper w	2019	3	8	1	0.333333	1.333333
Paper o	2007	15	9	3	0.2	0.8
Paper t	2007	15	10	2	0.133333	0.533333
Paper e	2021	1	11	0	0	0
Paper v	2005	17	12	0	0	0

hg-index

The hg-index is a combination of the h-index and the g-index. To calculate the g-index, first, the ranks of the publications are squared. Then the citations are summed. For example, paper b, which has rank = 3 and rank^2 = 9, has the sum of the citations from papers ranked 1, 2, and 3 (now ranked 1, 4, and 9, respectively).

Table C-4. Author A's Data with hg-index Numbers

Title	Publication Year	Age	Rank	Rank^2	Citation Count	Summed Citation Count	AR-value	S^c value (coef. 4)
Paper f	2012	10	1	1	200	200	20	80
Paper q	2009	13	2	4	65	265	5	20
Paper b	2016	6	3	9	24	289	4	16
Paper g	2004	18	4	16	12	301	0.7	2.7
Paper u	2005	17	5	25	7	308	0.4	1.6
Paper o	2007	15	6	36	3	311	0.2	0.8
Paper t	2007	15	7	49	2	313	0.1	0.5
Paper w	2019	3	8	64	1	314	0.3	1.3
Paper k	2020	2	9	81	1	315	0.5	2
Paper l	2021	1	10	100	1	316	1	4
Paper e	2021	1	11	121	0	316	0	0
Paper v	2005	17	12	144	0	316	0	0

The g-index is therefore the rank value found using the same method as the h-index. Because there is no inflection point (since [rank^2 > summed citation] is never achieved), Author A's g-index is 12. Note that the g-index is the rank and not the rank^2 value.

The hg-index is then the geometric mean of h and g, so the hg-index of Author A is:

$$\sqrt{5 * 12} = 7.7$$

The purpose of the hg-index is to balance the h-index's lack of credit to highly cited articles with the g-index's overinflated credit to highly cited articles.

Appendix D.
Assumptions and Rationale for Elements of the Task

This outline of assumptions and the rationale for this task follow the definitions used in the development of logic models. Assumptions are things that are accepted as true or as certain to happen, *without proof*. A rationale is a set of *reasons or a logical basis* for a course of action or a particular belief.

The purpose for conducting this exercise was to identify as many of assumptions and rationales for the concepts and methods integral to this task.

Assumptions for Research impact

- Outstanding scientists have outstanding research impact
- Outstanding scientists can be identified through their research impact
- Research impact can be measured in part, through the quality and quantity of scientific outputs and outcomes

Publications are a partial measure of the quantity of scientific outputs

- Citations are a partial measure of the quality of scientific outcomes
- Co-author networks are a partial measure of research impact and influence

Assumptions for the Use of Bibliometric Indicators to Identify Outstanding Scientists

- Outstanding scientists can be identified through measures of their research impact
- Research impact can be measured using bibliographic indicators of scientific outputs and outcomes
- Bibliometrics can identify outstanding academic scientists
- Bibliometrics can be used to assign values to scientists that correlate to research impact by measuring scientific output
- There are data available to calculate bibliometrics

Rationale for Identifying Outstanding Scientists

- Outstanding STEM scientists are critical to innovation and bolster U.S. competitiveness, national security, and the economy

- OSTP/the Federal Government wants to retain or attract outstanding STEM scientists to the United States
- Global competition for outstanding STEM scientists is increasing
- A significant share of the graduating body of PhDs from U.S. institutions are not U.S. citizens
  - 2017: 35% of S&E doctoral degrees were awarded to temporary visa holders (NSB 2019)
  - 2015: 75% of the temporary visa holders receiving doctoral degrees at a U.S. institution intended to stay in the United States and 20% to return to their home country (NCSES 2017)
- Federal grant funding process is cumbersome and time consuming; there are insufficient budgets to fund outstanding scientists

Rationale for the Use of Bibliometric Indicators to Identify Outstanding Scientists

- Bibliometrics have been historically used to measure published scientific output
- There is currently no robust and scalable method to identify outstanding scientists
- There is currently no robust and scalable method to predict who may become an outstanding scientist

Appendix E.
PAM and Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a bottom-up method, beginning with each author in their own group and subsequently combining pairs of groups based on their similarity (as determined by citation and publication information) until all groups, including groups with multiple authors, have been paired (see Figure E-1Figure 3 for an example). The pairwise combination of groups depends on a *link* function that determines how similarity is computed; as there are multiple link functions, it is first necessary to test which link function results in the best-fitting grouping structure. STPI used the agglomerative coefficient (AC), which is an aggregate measure of how well all authors fit with the groups they were assigned, to identify the appropriate link function.

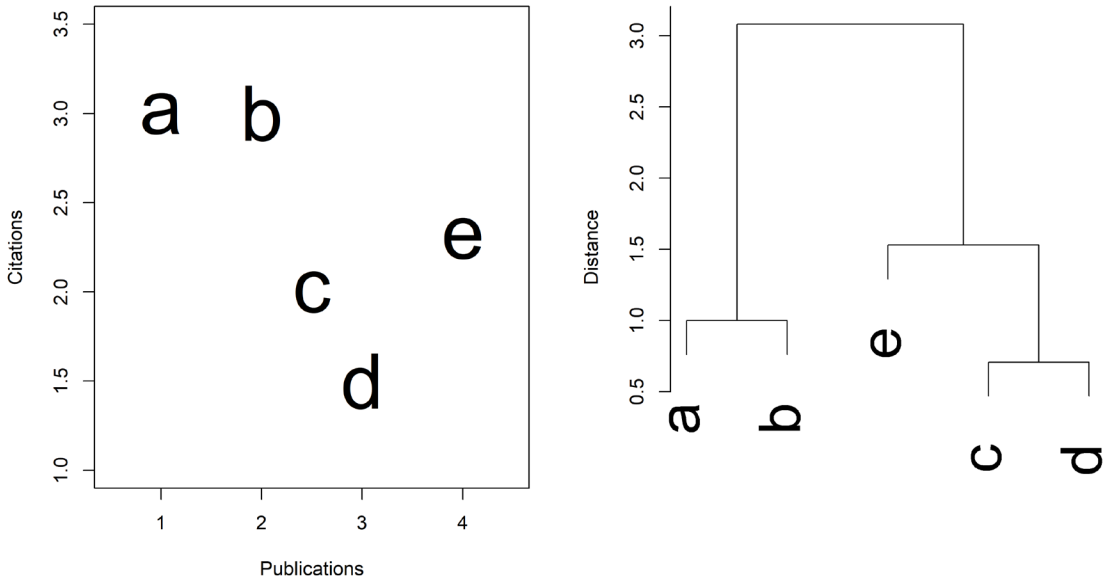


Figure E-1. Hierarchical Clustering Example of Authors a-e Based on Citations and Publications

PAM is an iterative algorithm that partitions the sample into several groups by first identifying *medoids,* in this case, authors in the sample that have a minimal dissimilarity with their other authors in their partition. Unlike the agglomerative hierarchical cluster, PAM requires users to identify the number of clusters prior to estimation, which is reflected

by the number of medoids used in the computation of groups—each author in the sample is assigned to the group represented by its most adjacent medoid.

Although neither method inherently recommends an optimal number of groups, average silhouette width can be used to compare the fit of several different clustering models with different numbers of groups. Average silhouette width provides a measure of cohesion of the clusters as compared with the distance between the clusters—a larger value for silhouette width suggests that the clusters are compact and distinct. Once the optimal number of groups has been identified for both PAM and agglomerative hierarchical clustering, the two methods can be compared for goodness of fit using internal measures of model fit such as the Dunn index and the within-cluster sum of squares. Both measures of fit favor better-fitting models and can be used to compare between PAM and agglomerative hierarchical clustering models.

Appendix F.
Final Clustering Tables

Table F-1. Distribution of Final Genetics Clustering

	High Impact	Medium Impact	Low Impact
Number of Profiles (%)	157 (1.99%)	6571 (83.42%)	1149 (14.59%)

Table F-2. Distribution of Final Artificial Intelligence Clustering

	High Impact	Medium Impact	Low Impact
Number of Profiles (%)	15 (0.16%)	973 (10.24%)	8517 (89.61%)

Appendix G.
Comparison of Web of Science, Scopus, and Google Scholar

A. Content

Both WoS and Scopus are marketed as highly curated publication databases that employ expert panels to determine titles to be included in their databases. There is particular focus on peer-reviewed or published documents including books and conference proceedings. Compared to Scopus, the WoS collection was found to have an overlap of 17.7 million documents—with Scopus containing 27 million documents and WoS containing 22.9 million (Visser et al. 2021). When broken down by discipline, documents in the life sciences had the greatest overlap of WoS documents with Scopus documents. Roughly less than half of Scopus documents in social sciences & humanities were covered by WoS. However, both Scopus and WoS had less than 20% of arts and humanities documents from Ulrich's Periodicals Directory, which contains a comprehensive database of periodicals (Mongeon and Paul-Hus 2016). This is indicative of chronic undercoverage of social sciences and humanities documents in databases. With regards to language, 90% of Scopus documents and 96% of WoS documents are in English, indicating significant undercoverage of non-English articles in both databases (Visser et al. 2021).

Google Scholar, on the other hand, had no active curation method. Instead, Google Scholar uses undisclosed web-scraping methods to pull non-peer reviewed and non-published materials in addition to the traditional journals, books, and conference proceedings. This makes Google Scholar's dataset significantly larger but less curated compared to WoS and Scopus.

B. Author Data

Both WoS and Scopus automatically create author profiles that are generated from publication metadata. Both platforms contain author profile search engines that allow users to search for authors by first and last name. Author profiles include a list of the authors' affiliations, their publications, and basic metrics such as the h-index, citation counts, and publication numbers. Both databases also allow users to export author-level citation histories, although Scopus exports are limited to 15-year time windows. However, STPI noted multiple instances of authors' profiles being combined when they shared identical or similar names within the free WoS Platform. There were also instances of individuals' profiles being split into different profiles because of differing metadata such as changes in

affiliation. Both WoS and Scopus offer purchasable APIs and raw data that include authenticated author profiles, which could contain more accurate name disambiguation.

Unlike WoS and Scopus, Google Scholar requires authors to create their profiles before they are made publicly available. Therefore, analyses that are dependent upon Google Scholar profile data have inherent selection bias; those with Google Scholar profiles may be more proactive in disseminating their research than the overall population of researchers. Moreover, the authors self-select up to five discipline fields that are free-form rather than limited to a pre-selected list. Because of this, authors can fail to select multiple fields if their research is interdisciplinary, can have varying levels of specificity when describing their discipline, or could be forced to exclude a discipline due to the limit of five areas.

C. Accessibility

It is much easier to export publication and author metadata from WoS and Scopus compared to Google Scholar. WoS and Scopus allow for robust exportation of search results or authors' publication histories. On the other hand, Google Scholar does not contain a built-in function to export either publication-level or author-level metadata. Data from Google Scholar has been cited as notoriously difficult to extract (Visser et al. 2021; Else 2018). As in the pilot study above, STPI had to apply a custom R web crawler to extract data for the study. Some documentation on Google Scholar content and curation processes is available but is relatively sparse compared to WoS and Scopus.[17] STPI recommends using WoS, Scopus, or other curated citation databases in any future analyses over Google Scholar because of their ease of use, accessibility, transparency, and validated data.

[17] Google's inclusion guidelines: https://scholar.google.com/intl/en/scholar/inclusion.html

References

Abbasi, Alireza, and Jorn Altmann. 2011. "On the correlation between research performance and social network analysis measures applied to research collaboration networks." In *2011 44th Hawaii International Conference on System Sciences*, pp. 1–10. IEEE.

Abramo, Giovanni, Andrea C. D'Angelo, and Gianluca Murgia. 2017. "The Relationship Among Research Productivity, Research Collaboration, and their Determinants." *Journal of Informetrics* 11 (4): 1016-1030. https://doi.org/10.1016/j.joi.2017.09.007.

Acharya, Ashwin, and Brian Dunn. 2022. *Comparing U.S. and Chinese Contributions to High-Impact AI Research*. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/comparing-u-s-and-chinese-contributions-to-high-impact-ai-research/

Aggarwal, C.C., 2018. "An introduction to cluster analysis." In Data clustering (pp. 1–28). Chapman and Hall/CRC.

Akella, Akhil Pandey, Hamed Alhoori, Pavan Ravikanth Kondamudi, Cole Freeman, and Haiming Zhou. 2021. "Early indicators of scientific impact: Predicting citations with altmetrics." *Journal of Informetrics* 15 (2): 101128.

Alonso, S., F. J. Cabrerizo, E. Herrera-Viedma, and F. Herrera. 2010. "hg-index: a new index to characterize the scientific output of researchers based on the h- and g-indices." *Scientometrics* 82 (2): 391–400. https://doi.org/10.1007/s11192-009-0047-5. https://doi.org/10.1007/s11192-009-0047-5.

Altmetric. 2022. "What are altmetrics? An introduction." https://www.altmetric.com/about-altmetrics/what-are-altmetrics/.

Bales, Michael E., Stephen B. Johnson, and Chunhua Weng. 2008. "Social network analysis of interdisciplinarity in obesity research." In *AMIA Annu Symp Proc*, vol. 870.

Barber, Bernard. 1961. "Resistance by scientists to scientific discovery." *Science* 134(3479): 596–602.

Batista, Pablo D., Mônica G. Campiteli, and Osame Kinouchi. 2006. "Is it possible to compare researchers with different scientific interests?" *Scientometrics* 68 (1): 179–189. https://doi.org/10.1007/s11192-006-0090-4. https://doi.org/10.1007/s11192-006-0090-4.

Bennett, D. and D. Taylor. 2003. "Unethical practices in authorship of scientific papers." *Emergency Medicine* 15: 263–270. https://doi.org/10.1046/j.1442-2026.2003.00432.x

Bihari, Anand, Sudhakar Tripathi, and Akshay Deepak. 2021. “A review on h-index and its alternative indices.” *Journal of Information Science* 0 (0): 01655515211014478. https://doi.org/10.1177/01655515211014478. https://journals.sagepub.com/doi/abs/10.1177/01655515211014478.

Birkle, Caroline, David A. Pendlebury, Joshua Schnell, and Jonathan Adams. 2020. “Web of Science as a Data Source for Research on Scientific and Scholarly Activity.” *Quantitative Science Studies* 1 (1): 363–376. https://doi.org/10.1162/qss_a_00018

Birnholtz, Jeremy. 2008. “When Authorship Isn't Enough: Lessons from CERN on the Implications of Formal and Informal Credit Attribution Mechanisms in Collaborative Research.” *Journal of Electronic Publishing* 11 (1): https://doi.org/10.3998/3336451.0011.105.

Bol, T., de Vaan, M., de Rijt, A. 2018. “The Matthew effect in science funding.” *PNAS* 115(19): 4887–4890.

Bornmann, Lutz. 2012. “What is Societal Impact of Research and How Can it be Assessed? A Literature Survey.” *Advances in Information Science* 64 (2): 217–233. https://doi.org/10.1002/asi.22803.

Bornmann, L. and Haunschild, R., 2017. “Relative Citation Ratio (RCR): an empirical attempt to study a new field-normalized bibliometric indicator.” *Journal of the Association for Information Science and Technology* 68(4): 1064–1067.

Bornmann, Lutz, and Robin Haunschild. 2018. “Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data.” *PloS one* 13 (5): e0197133.

Bornmann, Lutz, Rüdiger Mutz, and Hans-Dieter Daniel. 2008. “Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine.” *Journal of the American Society for Information Science and Technology* 59(5): 830–837.

Bu, Yi, Wei Lu, Yifei Wu, Hongkan Chen, and Yong Huang. 2021. “How Wide is the Citation Impact of Scientific Publications? A Cross-Discipline and Large Scale Analysis.” *Information Processing & Management* 58 (1): 102429.

Budd, J.M. and K.N. Stewart. 2015. “Is There Such a Thing as “Least Publishable Unit”? An Emperical Investigation.” *Libres* 25(2): 78–85.

Burrell, Quentin L. 2007. “Hirsch's h-index: A stochastic model.” *Journal of Informetrics* 1 (1): 16–25. https://doi.org/https://doi.org/10.1016/j.joi.2006.07.001.

Butler, Declan. 2013. “Investigating Journals: The Dark Side of Publishing.” *Nature* 495: 433–435. https://doi.org/10.1038/495433a

Cech, E.A., and Blair-Loy, M. 2019. “The changing career trajectories of new parents in STEM.” *Proceedings of the National Academy of Sciences of the United States of America* 116(10): 4182–4187.

Chu, J.S.G., and Evans, J.A. 2021. “Slowed canonical progress in large fields of science.” *PNAS* 118(41): e2021636118.

Clarivate. 2021. “Resources for librarians: Web of Science coverage details.” Accessed February 6, 2022. Available at: https://clarivate.libguides.com/librarianresources/coverage.

Clarivate. 2021b. “Highly Cited Researchers: Overview.” Accessed January 25, 2022. Available at: https://recognition.webofscience.com/awards/highly-cited/2021/methodology/.

Cole, Stephen. “Professional standing and the reception of scientific discoveries.” *American Journal of Sociology* 76, no. 2 (1970): 286–306.

Costas, Rodrigo, Thed N. van Leeuwen, and Anthony FJ van Raan. “The 'Mendel syndrome' in science: durability of scientific literature and its effects on bibliometric analysis of individual scientists.” *Scientometrics* 89, no. 1 (2011): 177–205.

Costas, Rodrigo, Zohreh Zahedi, and Paul Wouters. 2015. “Do 'altmetrics' correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective.” *Journal of the Association for Information Science and Technology* 66 (10): 2003–2019.

López-Cózar, Delgago, Orduna-Malea, Enrique, and Martín-Martín, Alberto. “Google Scholar as a Data Source for Research Assessment”. In *Springer Handbook of Science and Technology Indicators*, edited by Glänzel W., Moed H.F., Schmoch U., Thelwall M., Springer Handbooks.

Digiampietri, Luciano A., and Ernando E. da Silva. 2011. “A framework for social network of researchers analysis.” *Iberoamerican Journal of Applied Computing* 1(1): 1–24.

Dong, Yuxiao, Hao Ma, Jie Tang, and Kuansan Wang. 2018. “Collaboration diversity and scientific impact.” *arXiv* preprint arXiv:1806.03694

Du, Jian, and Yishan Wu. 2018. “A parameter-free index for identifying under-cited sleeping beauties in science.” *Scientometrics* 116, no. 2: 959–971.

Duma, N. 2020. “Gender differences in publication rates in oncology: looking at the past, present, and future.” *Cancer* 126(12): 2759–2761.

Durieux V, Gevenois PA. 2010. “Bibliometric indicators: quality measurements of scientific publication.” *Radiology,* 255(2): 342–51. doi: 10.1148/radiol.09090626.

Egghe, Leo. 2006. “An improvement of the H-index: The G-index.” *ISSI Newsletter* 2.

Egghe, Leo. 2008. “Mathematical theory of the h- and g-index in case of fractional counting of authorship.” *Journal of the American Society for Information Science and Technology* 59 (10): 1608–1616. https://doi.org/https://doi.org/10.1002/asi.20845. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.20845.

Else, Holly. “How I Scraped Data from Google Scholar.” Nature. April 11, 2018. https://www.nature.com/articles/d41586-018-04190-5

Ferrara, Emilio, and Alfonso E. Romero. 2013. “Scientific impact evaluation and the effect of self-citations: Mitigating the bias by discounting the h-index.” *Journal of*

the American Society for Information Science and Technology 64 (11): 2332–2339. https://doi.org/https://doi.org/10.1002/asi.22976. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.22976.

Fish, I.S. 2013. "Why do so many Chinese people share the same name?" *Foreign Policy*. Accessed February 6, 2022. Available at: https://foreignpolicy.com/2013/04/26/why-do-so-many-chinese-people-share-the-same-name/.

Garfield, E. 1972. "Citation Analysis as a Tool in Journal Evaluation: Journals can be Ranked by Frequency and Impact of Citations for Science Policy Studies." *Science* 178: 471–79. http://dx.doi.org/10.1126/science.178.4060.471

Garfield, E. 2007. "The Evolution of the Science Citation Index." *International Microbiology* 10: 65–69. https://doi.org/10.2436/20.1501.01.10

———. 1970. "Citation indexing for studying science." *Nature* 227, no. 5259: 669–671.

———. 1980. "Premature discovery or delayed recognition-why." *Current Contents* 21: 5–10.

———. 1989. "Creativity and science, Part 2. The process of scientific discovery." *Current Contents* 45, no. 6: 314–320.

Genolini, Christophe. 2016. "kmlShape: K-Means for Longitudinal Data using Shape-Respecting Distance." R package version 0.9.5. https://CRAN.R-project.org/package=kmlShape

Glänzel, W. 2006. "On the Opportunities and Limitations of the H-Index." *Science Focus* 1(1): 10–11. http://eprints.rclis.org/9378/1/H_Index_opprtunities.pdf

Green, Bob. 2019. "Intelligent metrics: The need for normalization of citation counts." Web of Science Group.

Guthrie, S. Rincon, D.R., McInroy, G., Ioppolo, B., Gunashekar, S. 2019. "Measuring bias, burden and conservatism in research funding processes." *F1000Research* 8(851): 851.

Harzing, A. 2016. "Publish or Perish." Accessed February 6, 2022. Available at https://harzing.com/resources/publish-or-perish.

He, Zhongyang, Zhen Lei, and Dashun Wang. 2018. "Modeling citation dynamics of "atypical" articles." *Journal of the Association for Information Science and Technology* 69(9): 1148–1160.

Hernandez-Alvarez, M. and J.M. Gomez. 2015. "Survey About Citation Context Analysis: Tasks, Techniques, and Resources." *Natural Language Engineering* 22(3): 327–349. https://doi.org/10.1017/S1351324915000388.

Hicks, Daniel J., David A. Coil, Carl G. Stahmer, and Jonathan A. Eisen. 2019. "Network analysis to evaluate the impact of research funding on research community consolidation." *PloS one* 14(6): e0218273.

Hirsch, J. E. 2005. "An index to quantify an individual's scientific research output." *Proceedings of the National Academy of Sciences of the United States of America*

102 (46): 16569–16572. https://doi.org/10.1073/pnas.0507655102. https://www.pnas.org/content/pnas/102/46/16569.full.pdf.

Hou, Jianhua, Hao Li, and Yang Zhang. 2020. “Identifying the princes base on Altmetrics: An awakening mechanism of sleeping beauties from the perspective of social media.” *Plos one* 15, no. 11: e0241772.

Huang, J., Gates, A.J., Sinatra, R., Barabasi, A. 2020. “Historical comparison of gender inequality in scientific careers across countries and disciplines.” *Proceedings of the National Academy of Sciences of the United States of America* 117(9): 4609–4616.

Hutchins, B.I., Yuan, X., Anderson, J.M. and Santangelo, G.M., 2016. “Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level.” *PLoS biology* 14(9): p.e1002541.

Ioannidis JPA, Boyack K, Wouters PF. 2016. “Citation Metrics: A Primer on How (Not) to Normalize.” *PLoS Biol* 14(9): e1002542. https://doi.org/10.1371/journal.pbio.1002542.

Janssens, A.C.J., Goodman, M., Powell, K.R. and Gwinn, M., 2017. “A critical evaluation of the algorithm behind the Relative Citation Ratio (RCR).” *PLoS biology* 15(10): p.e2002536.

Jin, BiHui, LiMing Liang, Ronald Rousseau, and Leo Egghe. 2007. “The R- and AR-indices: Complementing the h-index.” *Chinese Science Bulletin* 52 (6): 855–863. https://doi.org/10.1007/s11434-007-0145-9. https://doi.org/10.1007/s11434-007-0145-9.

Kaufman, Leonard, and Peter J. Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

———. 1990. “Partitioning around medoids (program pam).” *Finding groups in data: an introduction to cluster analysis* 344: 68–125.

Ke, Qing, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. 2015. “Defining and identifying sleeping beauties in science.” *Proceedings of the National Academy of Sciences* 112, no. 24: 7426–7431.

Keserci, Samet, Eric Livingston, Lingtian Wan, Alexander R. Pico, and George Chacko. 2017. “Research synergy and drug development: Bright stars in neighboring constellations.” *Heliyon* 3, no. 11: e00442.

Krist, W., 2013. *Globalization and America's Trade Agreements*. Woodrow Wilson Center Press.

Kyvik, Svein. 2003. “Changing Trends in Publishing Behavior Among University Faculty, 1980-2000.’ *Scientometrics* 58: 35–48. https://doi.org/10.1023/A:1025475423482

Lariviere, V., Ni, C., Gingras, Y., Cronin, B., Sugimoto, C.R. 2013. “Bibliometrics: global gender disparities in science.” *Nature* 504: 211–213.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. 2021. “cluster: Cluster Analysis Basics and Extensions. R package version 2.1.2.”

Maher, B., and Van Noorden, R. 2021. “How the COVID pandemic is changing global science collaborations.” *Nature* 594: 316–319.

Martin, Ben R., and John Irvine. 1983. “Assessing Basic research: Some Partial Indicators of Scientific Progress in Radio Astronomy.” *Research Policy* 12 (2):10.1016.

Mendel, Gregor. 1865. “Experiments in plant hybridization.” Verhandlungen des naturforschenden Vereins Brünn. Available at: www. mendelweb. org/Mendel. Html.

Mendoza, M., 2021. “Differences in citation patterns across areas, article types and age groups of researchers.’ *Publications* 9(4): 47.

Merton, R. 1968. “The Matthew effect in science: the reward and communication systems of science are considered.” *Science* 159(3810): 56–63.

Mihaljević, H., Tullney, M., Santamaría, L., & Steinfeldt, C. 2019. “Reflections on Gender Analyses of Bibliographic Corpora.” *Frontiers in big data*, 2, 29. https://doi.org/10.3389/fdata.2019.00029

Mihaljević-Brandt, H., Santamaría, L. and Tullney, M., 2016. “The effect of gender in the publication patterns in mathematics.” PLoS One, 11(10), p.e0165367.

Moravcsik, Michael J. 1977. “A Progress Report on the Quantification of Science. “ *Journal of Scientific and Industrial Research* 36 (5): 195–203.

Mostert, S.P., S.P.H. Ellenbroek, I. Meijer, G. van Ark and E.C. Klasen. 2010. “Societal output and use of research performed by health research groups.” *Health Research Policy and Systems* 8: 30. http://www.biomedcentral.com/content/pdf/1478-4505-8-30.pdf.

National Science and Technology Council. 2018. *Charting a course for success: America’s strategy for STEM education*. Accessed January 19, 2022. Available at: https://files.eric.ed.gov/fulltext/ED590474.pdf.

———. 2021. *Best practices for diversity and inclusion in STEM education and research: a guide by and for Federal agencies*. Accessed January 19, 2022. Available at: https://www.whitehouse.gov/wp-content/uploads/2021/09/091621-Best-Practices-for-Diversity-Inclusion-in-STEM.pdf?eType=EmailBlastContent&eId=83268b01-a660-4507-8e26-3120d3bdf70b.

———. 2022. *Guidance for implementing national security presidential memorandum 33 (NSPM-33) on national security strategy for United States Government-supported research and development*. Accessed February 10, 2022. Available at: https://www.whitehouse.gov/wp-content/uploads/2022/01/010422-NSPM-33-Implementation-Guidance.pdf.

National Science Foundation, National Center for Science and Engineering Statistics. 2017. “Doctorate Recipients from U.S. Universities: 2015.” *Special Report NSF 17-306.* Arlington, VA. https://www.nsf.gov/statistics/sed/2017/nsf17306/.

National Science Foundation. 2021a. "Survey of Earned Doctorates." Accessed January 19, 2022. Available at: https://www.nsf.gov/statistics/srvydoctorates/.

———. 2021b. "Survey of Doctorate Recipients." Accessed January 19, 2022. Available at: https://www.nsf.gov/statistics/srvydoctoratework/#:~:text=The%20Survey%20of%20Doctorate%20Recipients,or%20health%20(SEH)%20field.

NCSES. 2022. "Public Use Data Files Available for Download." Accessed January 19, 2022. Available at: https://ncsesdata.nsf.gov/datadownload/.

National Science Board, National Science Foundation. 2019. "Higher Education in Science and Engineering. Science and Engineering Indicators 2020." *NSB-2019-7.* Alexandria, VA. https://ncses.nsf.gov/pubs/nsb20197/.Nuñez, Xcaret. 2021. "Mattel's Barbie Turns Women of Science, Including COVID Vaccine Developer, Into Dolls." *NPR,* August 5, 2021. https://www.npr.org/2021/08/05/1024888880/mattels-barbie-turns-women-of-science-including-a-covid-vaccine-developer-into-d.

OECD. 2019. "OECD work on careers of doctorate holders" https://www.oecd.org/innovation/inno/careers-of-doctorate-holders.htm Last updated: July 3, 2019.

Okubu, Y. 1997. "Bibliometric indicators and analysis of research systems: Methods and examples." *OECD Science, Technology and Industry Working Papers* 1. https://doi:org/10.1787/208277770603

Orbis. 2022. "Orbis." Accessed January 19, 2022. Available at: https://www.bvdinfo.com/en-us/our-products/data/international/orbis?gclid=Cj0KCQiAip-PBhDVARIsAPP2xc1T7wdsH__9RVrosGiZkxR0SuyzVjqkZ9nWUdlFRHfI6nOsLCpV9QYaAqwIEALw_wcB.

ORCID. 2021. "About ORCID." Accessed January 21, 2022. Available at: https://info.orcid.org/what-is-orcid/.

Ortega, José-Luis. 2020. "Altmetrics data providers: A metaanalysis review of the coverage of metrics and publication." *El profesional de la información (EPI)* 29 (1).

Panaretos, J. and C. Malesios. 2009. "Assessing Scientific Research Performance and Impact with Single Indices." *Scientometrics* 81(3): 635–670. https://doi.org/10.1007/s11192-008-2174-9.

Patel, P.A., Gopali, R., Reddy, A. and Patel, K.K., 2021. "The relative citation ratio and the h-index among academic ophthalmologists: A retrospective cross-sectional analysis." *Annals of Medicine and Surgery* 71:103021.

Penfield, Teresa, Matthew J. Baker, Rosa Scoble, and Michael C. Wykes. 2014. "Assessment, Evaluations, and Definitions of Research Impact: A Review." *Research Evaluation* 23 (1): 21–32. https://doi.org/10.1093/reseval/rvt021

Phillips, Paul, Luiz Moutinho, and Pedro Godinho. 2017. "Developing and testing a method to measure academic societial impact." *Higher Education Quarterly* 72 (2): 12154.

Porter, Michael E. 1990. "The competitive advantage of nations." *Competitive Intelligence Review* 1(1): 14–14.

ProQuest. n.d. "ProQuest Dissertations & Theses Global" https://about.proquest.com/en/products-services/pqdtglobal/ Accessed January 31, 2022.

Purkayastha, A., Palmaro, E., Falk-Krzesinski, H.J. and Baas, J., 2019. "Comparison of two article-level, field-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR)." *Journal of Informetrics* 13(2): 635–642.

Rahkovsky, Ilya, Autumn Toney, Kevin W. Boyack, Richard Klavans, and Dewey A. Murdick. 2021. "AI Research Funding Portfolios and Extreme Growth." *Frontiers in Research Metrics and Analytics* 6: 11.

Ravenscroft, James, Maria Liakata, Amanda Clare, and Daniel Duma. 2017. "Measuring Research impact Beyond Academia: An Assessment of Existing Impact Metrics and Proposed Improvements." *PLoS ONE* 12(3): https://doi.org/10.1371/journal.pone.0173152.

R Core Team. 2021. "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ringelhan, Stefanie, Jutta Wollersheim, and Isabell M Welpe. 2015. "I like, I cite? Do Facebook likes predict the impact of scientific work?" *PloS one* 10 (8): e0134389.

Sage, Michelle, Paula Rascona, and Vijay D'Souza. 2021. "Federal Spending Transparency: Opportunities Exist to Further Improve the Information Available on USAspending.gov." Available at: https://www.gao.gov/assets/gao-22-104702.pdf.

Sarkesian, Sam C., John A. Williams, and Stephen J. Cimbala. 2008. *U.S. National Security: Policymakers, Processes & Politics*, Boulder, CO: Rienner.

Scopus. 2022. "Content coverage guide." Accessed February 6, 2022. Available at: https://www.elsevier.com/__data/assets/pdf_file/0007/69451/Scopus_ContentCoverage_Guide_WEB.pdf.

Sidiropoulos, Antonis, Dimitrios Katsaros, and Yannis Manolopoulos. 2007. "Generalized Hirsch h-index for disclosing latent facts in citation networks." *Scientometrics* 72 (2): 253–280.

Simko, I., 2015. "Analysis of bibliometric indicators to determine citation bias." *Palgrave Communications* 1(1): 1–9.

Siudem, Grzegorz, Barbara Zogala-Siudem, Anna Cena, and Marek Gagolweski. 2020. "Three Dimensions of Research impact." *Proceedings of the National Academy of Sciences of the U.S.* 117 (25): 13896–13900. https://doi.org/10.1073/pnas.2001064117.

Small, H. 2018. “Citation Indexing Revisited: Garfield’s Early Vision and Its Implications for the Future.” *Front.Res.Metr.Anal.* https://doi.org/10.3389/frma.2018.00008.

Smith, D.R. 2007. “Historical Development of the Journal Impact Factor and Its Relevance for Occupational Health.” *Industrial Health* 45: 730–742. https://doi.org/10.2486/indhealth.45.730

———. 2012. “Impact Factors, Scientometrics and the History of Citation-Basd Research.” *Scientometrics* 92: 419–427. https://doi.org/10.1007/s11192-012-0685-x

Stent, Gunther S. 1972. “Prematurity and uniqueness in scientific discovery.” *Scientific American* 227, no. 6: 84–93.

Steppingblocks. 2021. “Our data.” Accessed January 19, 2022. Available at: https://www.steppingblocks.com/privacy-policy.

Sugiyama, K., T. Kumar, M.-Y. Kan, and R.C. Tripathi. 2010. “Identifying citing sentences in research papers using supervised learning.” In 2010 International Conference on Information Retrieval and Knowledge Management (CAMP), Shah Alam, Selangor, Malaysia, pp. 67–72. http://dx.doi.org/10.1109/INFRKM.2010.5466945.

Tahamtan, I. and L. Bornmann. 2019. “What Do Citation Counts Measure? An Updated Review of Studies on Citations in Scientific Documents Published Between 2006-2018.” *Scientometrics* 121: 1635–1684. https://doi.org/10.1007/s11192-019-03243-4.

Tan, P.N., Steinbach, M. and Kumar, V., 2013. “Data mining cluster analysis: basic concepts and algorithms.” *Introduction to data mining* 487: 533.

Thelwall, Michael. 2020. “The pros and cons of the use of altmetrics in research assessment.” *Scholarly Assessment Reports* 2(1), p.2. DOI: http://doi.org/10.29024/sar.10.

Thelwall, Mike, Stefanie Haustein, Vincent Larivière, and Cassidy R Sugimoto. 2013. “Do altmetrics work? Twitter and ten other social web services.” *PloS one* 8 (5): e64841.

Thelwall, Mike, and Tamara Nevill. 2018. “Could scientists use Altmetric. com scores to predict longer term citation counts?” *Journal of informetrics* 12(1): 237–248.

Vaidya, Jayant S. 2005. “V-index: a fairer index to quantify an individual's research output capacity.” *BMJ*.

van Raan, Anthony FJ. 2015. “Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations.” *PloS one* 10(10): e0139786.

———. 2017. “Sleeping beauties cited in patents: Is there also a dormitory of inventions?” *Scientometrics* 110(3): 1123–1156.

van Raan, Anthony FJ, and Jos J. Winnink. 2018. “Do younger Sleeping Beauties prefer a technological prince? “ *Scientometrics* 114(2): 701–717.

Viglione, G. 2020. “Are women publishing less during the pandemic? Here's what the data say.” *Nature* 581: 365–366.

Visser, Martijn, Nees Jan van Eck, and Ludo Waltman. 2021. “Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic.” *Qualitative Science Studies*. 2(1): 20–41. https://doi.org/10.1162/qss_a_00112

Wang, J., Veugelers, R. and Stephan, P. 2017. “Bias against novelty in science: A cautionary tale for users of bibliometric indicators.” *Research Policy* 46(8): 1416–1436.

Wildgaard, L., J.W. Schneider, and B. Larsen. 2014. “A Review of the Characteristics of 108 Author-Level Bibliometric Indicators.” *Scientometrics* 101: 125–158. https://doi.org/10.1007/s11192-014-1423-3.

Wildgaard, L., 2015. A comparison of 17 author-level bibliometric indicators for researchers in Astronomy, Environmental Science, Philosophy and Public Health in Web of Science and Google Scholar. *Scientometrics*, 104(3), pp.873–906.

Wilsdon, James & Allen, Liz & Belfiore, Eleonora & Campbell, Philip & Curry, Stephen & Hill, Steven & Jones, Richard & Kain, Roger & Kerridge, Simon & Thelwall, Mike & Tinkler, Jane & Viney, Ian & Wouters, Paul & Hill, Jude & Johnson, Ben. 2015. “The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.” 10.13140/RG.2.1.4929.1363.

Wohlin, Claes. 2009. “A new index for the citation curve of researchers.” *Scientometrics* 81 (2): 521–533. https://doi.org/10.1007/s11192-008-2155-z. https://akjournals.com/view/journals/11192/81/2/article-p521.xml.

Zhang, G., Y. Ding, and S. Milojević. 2013. “Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content.” *Journal of the American Society for Information Science and Technology* 64(7): 1490–1503. https://doi.org/10.1002/asi.22850.

Abbreviations

AC	agglomerative coefficient
AI	artificial intelligence
altmetrics	alternative metrics
APIs	application programming interfaces
CDH	Careers of Doctorate Holders
CIPSEA	Confidential Information Protection and Statistical Efficiency Act
CSET	Center for Security and Emerging Technology
DOIs	digital object identifiers
FFRDCs	federally funded research and development centers
ID	identification
IF	impact factor
IRIS	Institute for Research on Innovation & Science
NCSES	National Center for Science and Engineering Statistics
NIH	National Institutes of Health
NSB	National Science Board
NSF	National Science Foundation
OECD	Organisation for Economic Co-operation and Development
ORCID	Open Researcher and Contributor ID
OSTP	Office of Science and Technology Policy
PAM	Partitioning Around Medoids
PhD	Doctor of Philosophy
PII	personal identifiable information
PoP	Publish or Perish
PQDT	ProQuest Dissertation and Thesis Database
R&E	research and development
RCR	Relative Citation Ratio
SCI	Science Citation Index
SDR	Survey of Doctoral Recipients
SED	Survey of Earned Doctorates
SEH	science, engineering, or health
STEM	science, technology, engineering, and mathematics
STPI	Science and Technology Policy Institute
UNESCO	United Nations Educational, Scientific and Cultural Organization
WoS	Web of Science

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-03-2022	FINAL	

4. TITLE AND SUBTITLE

Identifying Outstanding Scientists Using Bibliometric Indicators

5a. CONTRACT NUMBER

NSFOIA-0408601

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Han, Xueying
Tinkle, Sally S.
Panko, Pavel
Dinh, Nathan N. L.
Hazan, Gabriella G.
Wong, Gifford J.
Goldman, Abby R.
Katz, Luba

5d. PROJECT NUMBER

TP-20-1005.DD

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

IDA Science and Technology Policy Institute
1899 Pennsylvania Avenue, NW, Suite 520
Washington, DC 20006-3602

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document D-33048

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Science and Technology Policy
Executive Office of the President
Eisenhower Executive Office Building
1650 Pennsylvania Avenue, NW
Washington, DC 20504

10. SPONSOR/MONITOR'S ACRONYM(S)

OSTP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (24 May 2022).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The White House Office of Science and Technology Policy (OSTP) asked the IDA Science and Technology Policy Institute (STPI) to evaluate whether outstanding scientists can be identified using existing bibliometric indicators and to assess the limitations of using bibliometric indicators for this purpose. STPI was also asked to develop a new analytical approach that takes advantage of bibliometric indicators to determine which individuals are outstanding scientists. Specifically, the two research questions addressed in this study are: (1) Can bibliometric indicators be used to accurately identify outstanding scientists within a scientific discipline through time, and what are the limitations of their use? and (2) Can new analytical approaches be developed to identify outstanding scientists within a scientific discipline through time? Regarding the first question, STPI found that there is not an agreed upon set of bibliometric indicators that can or should be used to identify outstanding scientists. To the second question, STPI identified the inability to verify outstanding scientists as a general challenge because there is no consensus on a set of criteria that characterizes an outstanding scientist, but offered potential next steps for OSTP to consider that could refine the identification process.

15. SUBJECT TERMS

bibliometrics; citation-based indicators; outstanding scientists; publication-based indicators; research impact

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Hebner, Gregory
Unclassified	Unclassified	Unclassified	U/L	82	19b. TELEPHONE NUMBER *(Include area code)* 202-456-6129

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18