Analysis of citation dynamics reveals that you do not receive enough recognition for your influential science

Salsabil Arabi[1], Chaoqun Ni[1], and B. Ian Hutchins[1*]

[1] Information School, University of Wisconsin-Madison, Madison, WI

* Correspondence to: bihutchins@wisc.edu

Abstract

During career advancement and funding allocation decisions in biomedicine, reviewers have traditionally depended on journal-level measures of scientific influence like the impact factor. Prestigious journals are thought to pursue a reputation of exclusivity by rejecting large quantities of papers, many of which may be meritorious. It is possible that this process could create a system whereby some influential articles are prospectively identified and recognized by journal brands but most influential articles are overlooked. Here, we measure the degree to which journal prestige hierarchies capture or overlook influential science. We quantify the fraction of scientists' articles that would receive recognition because (a) they are published in journals above a chosen impact factor threshold, or (b) they are at least as well-cited as articles appearing in such journals. We find that the number of papers cited as frequently as those published in high impact factor journals vastly exceeds the number of papers these prestigious journals publish. At the investigator level, this phenomenon extends across gender, racial, and career stage groupings of scientists. We also find that approximately half of researchers never publish in a venue with an impact factor above 15, which under journal-level evaluation regimes may exclude them from consideration for opportunities. Many of these researchers publish equally influential work, however, raising the possibility that the traditionally chosen journal-level measures that are routinely considered under decision-making norms, policy, or law, may recognize as little as 10-20% of the work that warrants recognition.

Introduction

Biomedical hiring and promotion committees use journal-level heuristics because the quantity of documentation that researchers provide in their applications exceeds the attentional capacity that reviewers can provide (1). Ideally, scientific experts would read and apply their scientific expertise to each article under consideration for a given decision about hiring, promotion, or resource allocation (1-3). In some settings, this ideal may be possible, however, in many settings it is not feasible. For example, tenure-track job advertisements often attract hundreds of applicants, each with dozens of relevant publications to assess. In such a case, the biomedical research community recommended the inclusion of quantitative indicators, but ones that reflect information specific to the articles under consideration, not the venues in which they appear (1, 3).

One sensibility that often resonates with members of the scientific community is that if we recognize papers during personnel advancement based in large part on the citation rate of their venues, we should also recognize papers cited equally well, even if appearing in less prestigious venues (4). In other words, if we recognize a paper published in Cell, we should also recognize papers appearing in journals with lower impact factors that are equally well-cited and valued by practitioners. However, journal-level metrics fail to identify up to 90% of the equally well-cited papers in the biomedical research literature, controlling for field and year of publication (5). In other words, the recall of the biomedical community's chosen measure of meritorious papers is on the order of a dismal 10%.

Here, we measure the degree of misallocation of recognition based on the exclusive use of journal-level metrics at the level of individual articles and investigators. We use a common approach, impact factor thresholding (6), to identify papers published in prestigious venues ("Journal Elite" papers) as well as those that are equally well-cited but published in any journal ("Citation Elite" papers). We use a journal citation rate (hereafter referred to as "impact factor") of 15 citations per article per year as our main threshold, but test 10 and 20 citations per paper per year as well for robustness checks. We find that the number of papers that would receive recognition using article-level metrics instead of journal-level metrics is several-fold. Notably, half of biomedical researchers have never published in an elite journal, but a large fraction of these have published equally well-cited papers.

Results

Journal-level metrics overlook the most influential science

In a previous study, we asked how many highly influential papers are actually published in the most prestigious journals, such as *Science, Nature*, and *Cell* (5), using the Relative Citation Ratio (RCR), an article-level citation indicator that accounts for field and year of publication (4, 5, 7-9). We showed that there are nearly ten times more papers that achieve citation levels comparable to those in prestigious venues such as *Cell, Nature,* and *Science*, yet appear in less prestigious venues. This builds on previous work showing that there has been a decreasing correlation over the decades between journal impact factors and individual article citation rates (10). Number of publications can also influence career advancement (van Dijk et al., 2014), but because we fixed the number of publications and compared only the article-versus journal-level metrics for each paper, this is controlled for in our analysis. There is a weak correlation between the number of publications per year and citation statistics, but it may not rise to the level of practical significance (see Supplement). These results raise the alarming possibility that, in identifying influential papers exclusively based on the prestige of the journal in which they appear, research assessment now overlooks the vast majority of equally influential articles. However, the extent of this oversight may vary depending on the precise journal citation rate threshold chosen to identify "prestigious" journals.

To test the robustness of our findings, we varied the impact factor threshold used to classify a journal as "prestigious" or not. We determined the median RCR of papers published in journals matching or exceeding the chosen impact factor threshold (impact factor >= 15, "Journal Elites"). Articles with an RCR higher than the median of those published in journals above the impact factor thresholds were labeled as article-level "Citation Elites" (Figure 1A-B). Throughout this paper, we assess the robustness of findings to different choices of impact factor thresholds

and find them to be generally invariant (Supplemental Materials). We find that most influential papers are published in journals below the impact factor threshold defining "prestigious" (Figure 1C). This result raises a question: where are these other influential papers published? The modal response depends slightly on the impact factor threshold used to define "prestigious", but in general, such highly influential papers are published in journals with lower impact factors (Figure 1C). When testing the lower bounds on thresholds, we found that for Journal Elite papers to outnumber Citation Elite papers, the threshold for 'prestigious' journals would need to be as low as an impact factor of 3, which is unlikely. Furthermore, the more restrictive the impact factor threshold one uses, the fewer Citation Elite papers are recognized using journal-level measures (Figure 1D).

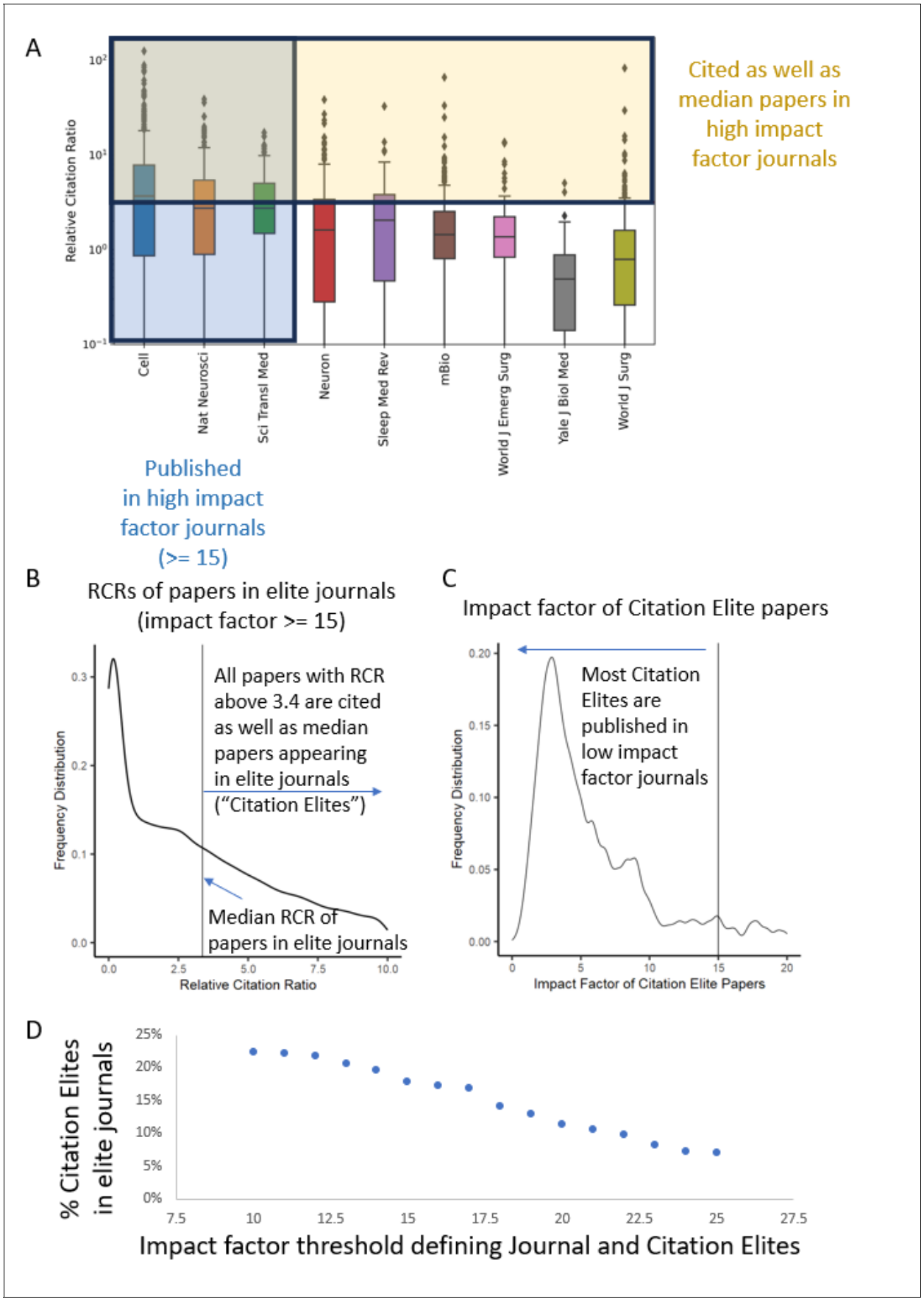


Figure 1. Most influential papers are published in lower-impact factor journals. (A) Schematic for determining which papers are as highly cited as those in prestigious journals. Box and whisker graphs of the RCR of papers from three journals above the selected impact factor threshold are shown (left, blue shading, "Journal Elites") alongside similar graphs for journals with subthreshold (< 15 citations per paper per year) are shown to the right. Articles with RCR's above the line (yellow shading) are considered Citation Elites. All of these journals published multiple Citation Elites. (B) Distribution of article-level Relative Citation Ratios for papers published in elite journals (impact factor >= 15). (C) Impact factor of journals where Citation Elite papers were published. Most papers that are as well cited as those in impact factor >= 15 journals are, in fact, published in lower-impact factor journals. (D) Generalization of the findings in (C) across different impact factor thresholds. The more stringent the threshold to define an elite journal, the more Citation Elites are published among lower-impact factor journals.

Incorporating citation elite papers into evaluation benefits the vast majority of authors

To evaluate how these article-level results generalize to investigator-level indicators, we turned to publicly available, author-level profiles published by the National Center for Biotechnology Information (NCBI) (11). Biomedical investigators have the opportunity to generate an opt-in public publication profile that can also be used to populate their most relevant publications on NIH bio-sketches while they apply for funding. We downloaded 50422 publicly available publication profiles from NCBI in 2022 that matched NIH-funded investigators, ranging from graduate fellows to late-career investigators (Figure 2). We focused on researchers who published at least one article in recent years (i.e., 2010-2019) to exclude artifacts from investigators who left the biomedical research workforce prior to 2010.

The use of journal-level metrics is historically pervasive and entrenched in biomedical research hiring and promotion decisions (1, 2). How might this change if the community also recognized Citation Elite papers? The fraction of scientists that have a higher number of citation elite papers vs. Journal Elite papers is nearly an order of magnitude higher (Figure 2). Specifically, 65.19% of scientists would be better recognized for their research based on article-level metrics, while about 4.56% would be based on journal-level metrics. This suggests a substantial improvement in recognition for a large segment of the biomedical research workforce by including article-level indicators as a way of recognizing research.

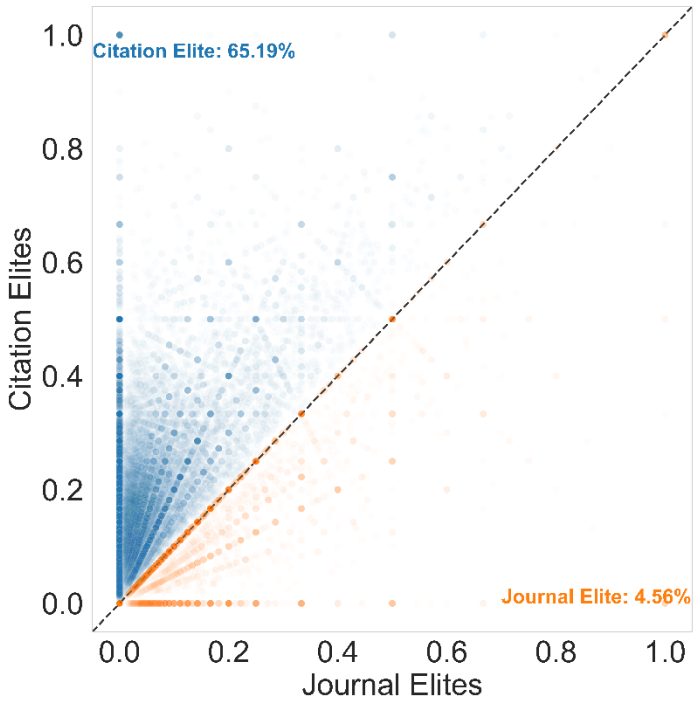


Figure 2. Scatter plot of the fraction of each researchers' publications (each dot is one researcher) that are above the journal impact factor threshold (x-axis) vs. those that are above the corresponding article-level citation threshold (y-axis). Researchers above the line would receive more recognition using article-level metrics (hereafter labeled as Citation Elite) while those below the line would receive more recognition using journal-level metrics (hereafter referred to as Journal Elite). Numbers do not add to 100% because some scientists publish no papers that would fall into the Journal Elite or Citation Elite categories.

Who benefits from article-level recognition?

Structural barriers to equality in the biomedical research workforce have traditionally favored scientists who are more senior, white, or men (12-14). We asked whether historically underrepresented groups likewise receive more recognition using article- vs. journal-level metrics. We therefore examined the distribution of demographic characteristics of those scientists who receive more recognition using journals vs. articles (see Methods). For each author, we collated the fraction of papers in their entire profile that met the definition of citation elite or comparatively journal elite. Because the number of papers in each author's profile is fixed between the two measures, the number of publications is controlled for. Overall, 65.19% of researchers would receive more recognition using article-level metrics, while a vastly smaller 4.56% would use journal-level metrics. Of those who receive more recognition with journal-level metrics, 69.53% are men, while 30.47% are women (Figure 3). By contrast, of those who would receive more recognition with article-level metrics 62.28% were men while 37.71% were women (Figure 3), much closer to the overall distribution of men and women in this dataset (38.87% female) (Figure 3). By comparing metrics for men and women in absolute, rather than relative terms, we observe that 18.43-fold more women and 13.34-fold more men receive more recognition with article-level metrics (Figure 3 and Supplemental Table 6).

Figure 3. Demographic analysis of scientists who receive more recognition with journal- vs. article-level measures (RCR). (Top, within box) Breakdown of how many scientists filtered by those who have more Citation Elite (these scientists are labeled as Citation Elites) papers than Journal Elite papers (labeled as Journal Elites) are male vs. female. Black line, the actual number of male vs. female scientists in this dataset. (Top, outside box) Breakdown of how many scientists filtered by gender receive more recognition with article-level citations (Citation Elite) vs. journal impact factor (Journal Elite). (Middle) Breakdown filtered by race (Binomial p value < 0.0001 across all racial groups). (Bottom) Breakdown filtered by seniority.

We next examined the same effects stratified by race. The vast majority of scientists, regardless of race, would receive more recognition using article-level metrics (Figure 3). Statistical analysis confirmed that this is a highly significant effect (Binomial p-value < 0.0001, Supplemental Table 1) for each racial category (Asian, Black, White, Hispanic). The phenomenon of receiving more recognition with article-level measures appears to be broadly shared across racial groups. Recent work suggests that applying thresholding to imputed racial predictions as we do here could, in principle, amplify class imbalances inherent in the data (15-17).

We, therefore, confirmed our results on continuous racial prediction scores and found similar results (Supplemental Figure 1). Accordingly, if raw article citation rates are used as a substitute for RCR as a robustness check, nearly identical results are observed (Supplemental Figure 2). These results hold even when subsetting to those top 10,000 researchers ranked by either sorting method, article- or journal-level metrics (see Supplement).

Finally, we examined the relationship of career stage and the differences between researcher recognition with journal- and article-level metrics (Figure 3 and Supplemental Table 7). As we observed with racial demographics, improvements in recognition are broadly shared across early, mid, and senior career stages.

Reexamination under strict zero-sum conditions

If article-level measures were considered instead or in addition to journal-level measures, would any career advancement benefit be gained? After all, competition for resources that enable research is often framed as zero-sum, whereby researchers are ranked by correlates of merit and some percentage selected. Under a zero-sum framework, it would be expected that there should be no proportional differences in the recognition of researchers on the basis of article-level metrics vs. journal-level metrics, as each person better-recognized under one measure would be replaced by someone better-recognized under the other. Surprisingly, we still observe large differences between researchers labeled as Citation Elites vs. Journal Elites, even when using rank-ordered metrics that are ostensibly zero-sum. Many more researchers published rank-ordered Citation Elite papers than published rank-ordered Journal Elite papers (47.1% vs. 27.8%).

Such a result should not be possible for ranked continuous data, as the proportions should theoretically be equal at 50%. To investigate this phenomenon further, we examined the probability distributions of both raw and rank-ordered statistical distributions. (Figure 4). While the distribution scores are skewed for both journal-level and article-level metrics (Figure 4A-B), the Citation Elite frequency distribution has a fatter tail, indicating more researchers publish such highly cited papers - consistent with the previous results. Rank-ordering the data should produce uniform statistical distributions, aligning with a zero-sum competition framework. For much of the probability distribution (Figure 4C-D), we see that this is the case. However, in both datasets, we observe that there is zero-inflation. In other words, for each indicator, some researchers never publish Journal or Citation Elite papers. However, the degree of zero-inflation differs greatly between Journal and Citation Elites(Figure 4C-D): Vastly more scientists receive recognition for their influential work when article-level indicators are considered than journal-level indicators.

Even using rank-order data that aligns with a zero-sum competition framework, we observe that, in general, researchers are much more frequently recognized using article-level indicators (Figure 5). We observe a similar gender difference favoring article-level metrics using the rank-ordered data as we observed using the continuous data. Likewise, we observed that each gender, racial group, and career stage in general would receive more recognition with article-level indicators (Figure 5). Thus, even under a nominally zero-sum framework, diverse researchers would receive more recognition using article-level indicators. A large minority of researchers , who would be overlooked by journal-level metrics, would gain recognition if article-level metrics were used (Figure 6). Therefore, the artificial scarcity of recognition in a journal-level framework impacts a sizeable minority, if not the outright majority, of biomedical researchers.

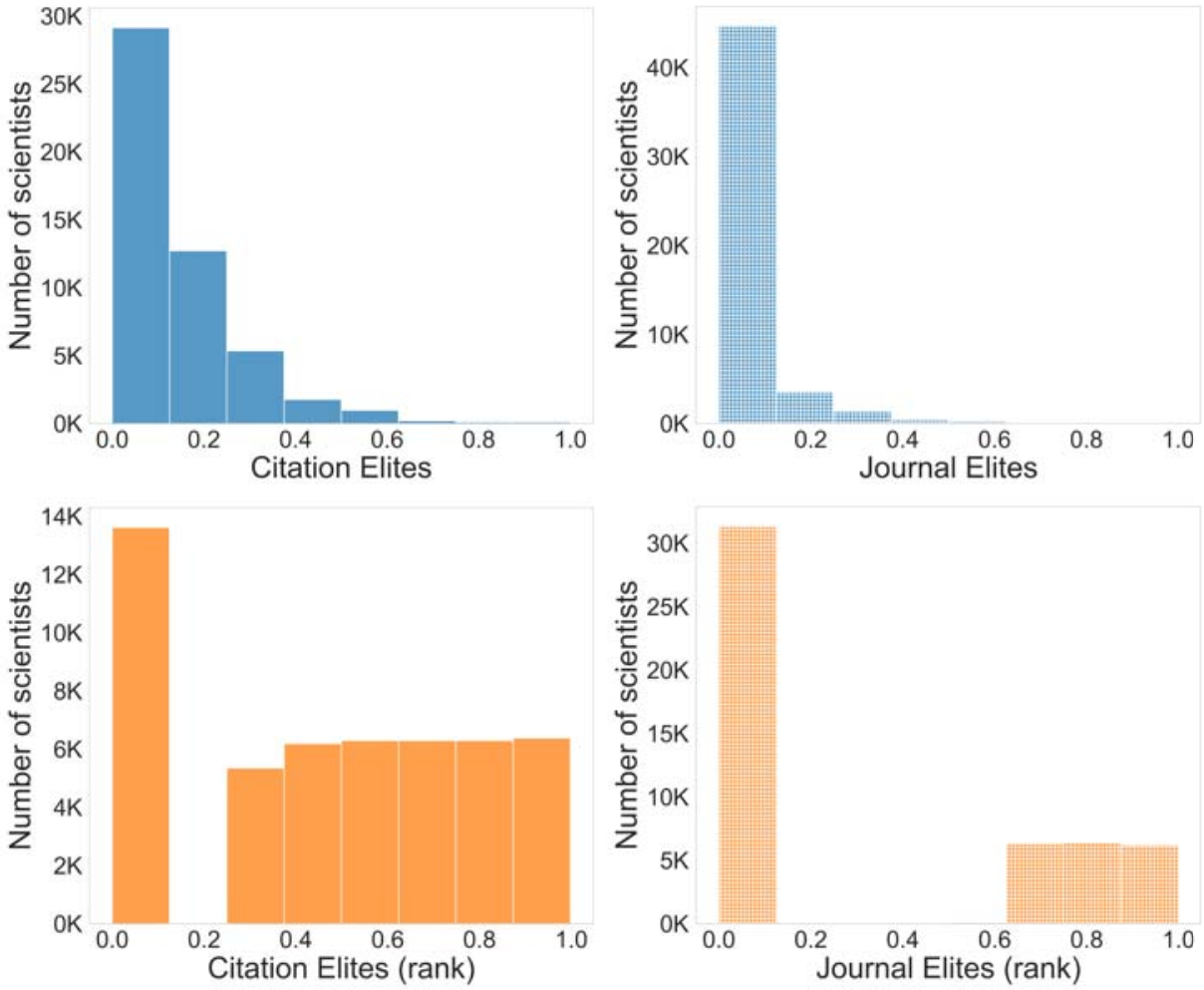


Figure 4. Distributions of researcher recognition. (A, B) Frequency distributions of the fraction of each researcher's publications that are recognized under article-level (A, "Citation Elites") vs. journal-level metrics (B, "Journal Elites"). (C, D) Frequency distributions after rank-ordering to align the scores with a zero-sum framework using article-level (C) vs. journal-level metrics (D). In principle, rank-ordering should flatten these into uniform distributions, but there is a high degree of zero-inflation. Approximately half of researchers have zero publications in prestigious journals (D, left). By contrast, only a quarter of researchers do not have articles that are cited as well as those published in prestigious journals (C, left).

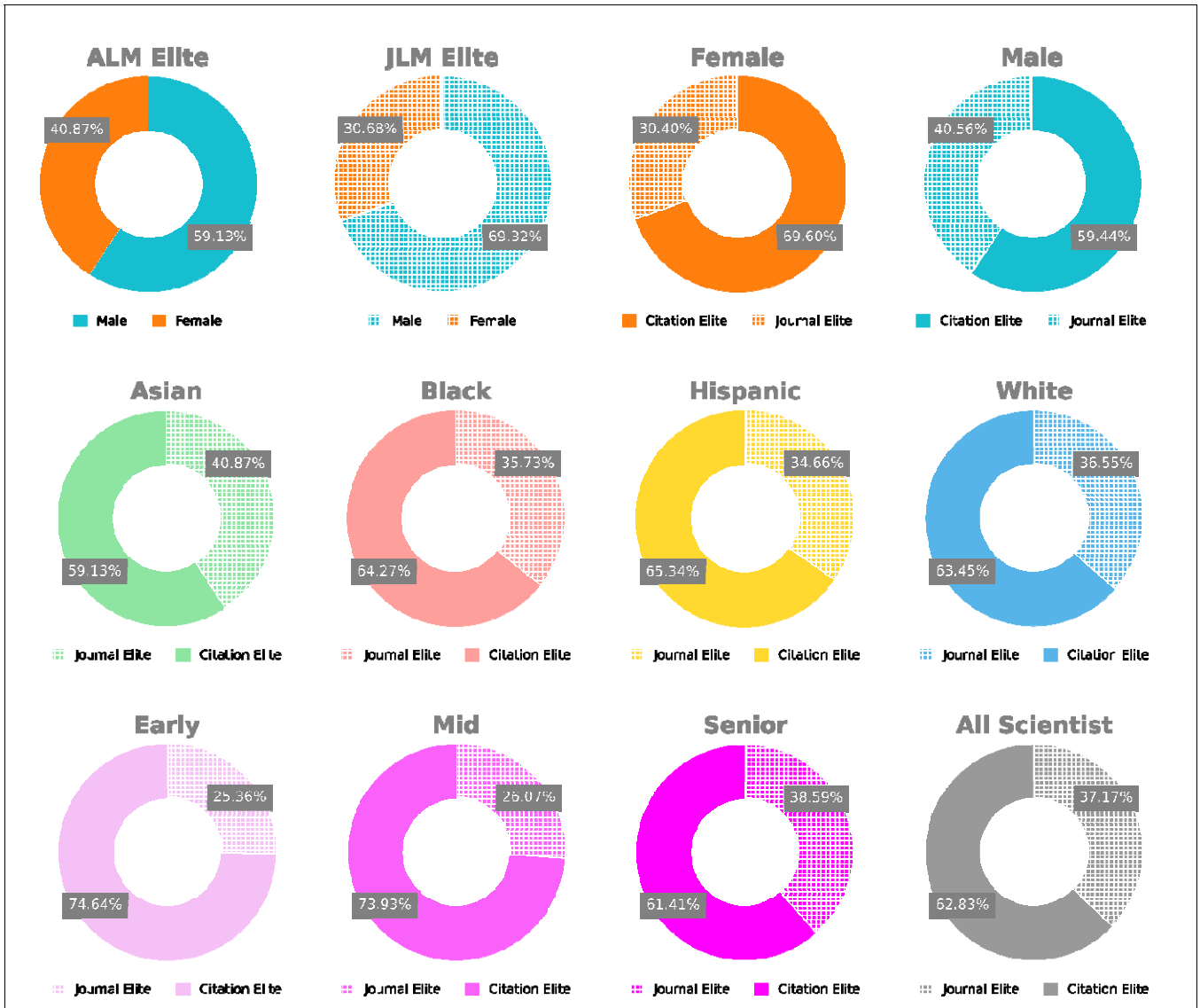
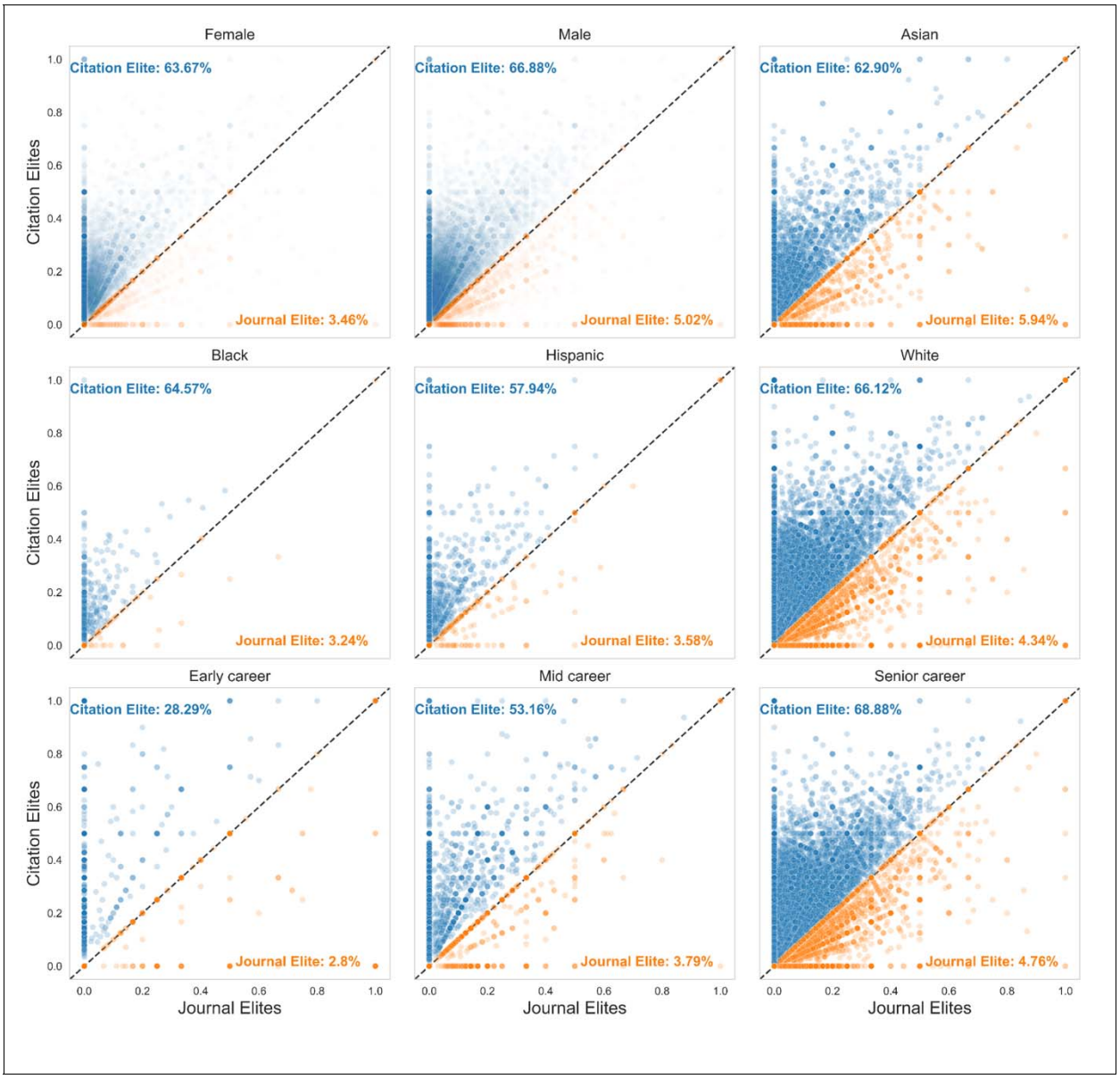


Figure 5. Demographic analysis of scientists who receive more recognition with journal- vs. article-level (RCR) measures using zero-sum rankings. (Top) Breakdown by gender. (Middle) Breakdown by race. (Bottom) Breakdown by seniority.



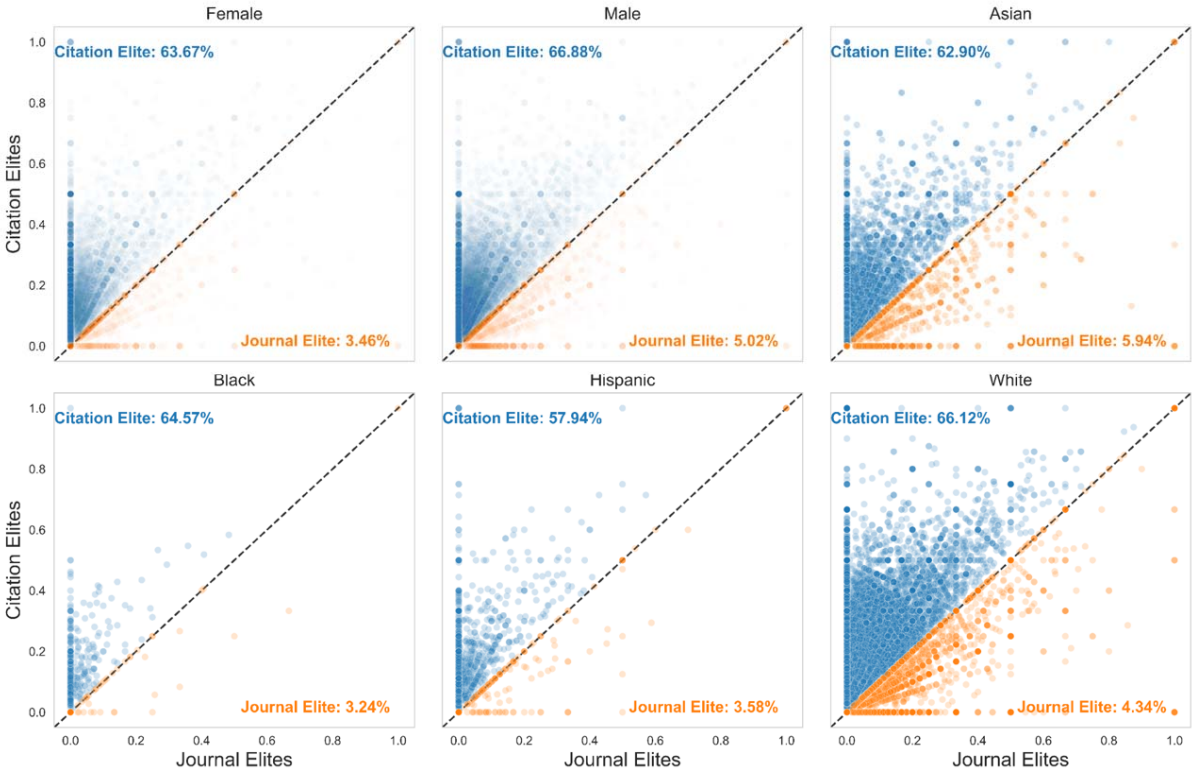


Figure 6. Scatter plots of the fraction of researchers' papers are Journal Elite (x-axis) vs. Citation Elite (y-axis), stratified by gender or race.

Discussion

This study investigated the degree of misallocation of recognition based on the exclusive use of journal-level metrics in assessing individual investigators, contrasting it with the use of article-level indicators as recommended by the scientific community (1, 3). We find that, using a variety of impact factor thresholds, researchers overwhelmingly receive more recognition with article-level indicators. This advantage cuts across racial, gender, and career groups. Even under zero-sum conditions, we observe that many more researchers would receive recognition using article-level indicators. This is because most researchers never publish in a "prestigious" journal as operationalized by impact factor thresholds, but many published papers in lower impact factor journals are as highly cited as those appearing in these journals.

In 2013, signatories to the San Francisco Declaration on Research Assessment acknowledged the pervasive harm that the continued use of journal-level heuristics in funding, hiring, and promotion decisions causes to scientists (1, 3). This mental heuristic is akin to judging a paper by the company it keeps (the average citation rate of all the other papers published in the journal) rather than by the article's own downstream influence. Signatories acknowledged that, while the use of scientific judgment about a paper is the gold standard, in many decision-making contexts where there are potentially thousands or millions of scientific articles to compare, it is

necessary to augment expert judgment with metrics. These should be article-level to ensure that the information is specific to the article(s) in question, rather than by the company they keep.

Our results echo and underscore the importance of considering article-level metrics in assessing scholars and scholarships. While citations are a form of recognition by the community, journals names are more visible to reviewers, who often use journal-level metrics such as journal impact factors as a shorthand for scientific impact. To assess article-level impact, reviewers would need to investigate each paper's citation record, which is rarely done in practice unless these metrics are included in curricula vitae. This is not to diminish by any means the value of journal-level metrics, but to underscore that article-level metrics provide important insights, especially for researchers who publish impactful work outside of prestigious venues for various reasons.

These are not as visible to reviewers of applicants as the journals listed in an applicant's files (and thereby reviewers' perceptions of article influence via their well-known impact factors). This is because in order to assess citation-based measures, reviewers only need to be familiar with each journal to assess its collective citation influence, while reviewers would have to look up the citation influence for each individual paper in a curriculum vitae to understand each paper's influence on its community unless article-level metrics were normalized to be included in curriculum vitae. This is not to say that journal-level measures hold no value, but rather that article level measures convey unique information that is valuable. This is particularly advantageous to those researchers currently being assessed on journal level measures without respect to the influence their individual papers have had, and who may not be recognized based on their venues of publication.

A dynamic, sustainable, and diverse scientific workforce is vital for a country's economic growth, global competitiveness, and ability to address pressing societal challenges. Yet, the sustainability of the U.S. biomedical research workforce has been an ongoing concern, particularly given its skewness towards older researchers (13, 18), and underrepresentation of African American/Black (AA/B) researchers (14, 19, 20). Furthermore, the U.S. scientific workforce shows historically male-dominated structure (21-23). The gender disparities may be attributed to various factors, among which is female researchers' limited access to resources and opportunities for publishing in esteemed journals (21,22,28,31). Consequently, the barriers faced by women in publishing in these journals result in unfavorable evaluations based on journal-level metrics. Given the consequence of such evaluation, female researchers would be further disadvantaged, as suggested by our results. Other underrepresented groups in the biomedical scientific workforce would also suffer for similar reasons.

Such trends exacerbate worries about the workforce's future. These topics have been discussed at congressional appropriations hearings (24), motivated specific appropriations for the National Institutes of Health (NIH) in the 21st Century Cures Act (13), and have been the subject of written testimony from NIH leadership to Congress (25). NIH Initiatives such as the Next Generation Research Initiative and Faculty Institutional Recruitment for Sustainable Transformation initiative have attempted to address some of these workforce problems (26, 27). However, the scientific community itself bears responsibility for structural issues that persist. The inappropriate use of journal-level metrics, as highlighted by DORA, remains a key challenge, as it amplifies recognition for established, often white, male researchers while marginalizing others (19, 20), potentially amplifying structural inequalities in the workforce. Results from this study offer further evidence that the utilization of article-level metrics may

contribute to addressing the broader structural imbalance in our scientific workforce and help building a competitive scientific workforce.

Despite widespread criticism, journal-level metric such as journal impact factor remains a key criterion in review, promotion, and tenure (RPT) decisions in many institutions (43-44). Several studies reported its inclusion as an important criterion for hiring, tenure, performance evaluation, and promotion and this trend persists across countries and faculty positions (45-46), despite being well-critiqued as an evaluator of the quality of publications. Additionally, a decent percentage of research-intensive as well as Master's institutions refer to journal impact factor or related terms and encourage their usage in review, promotion, and tenure documents (47). Moreover, this metric is often considered correlated with the quality, significance and impact of research by the institutions and review boards. Therefore, many researchers in the biomedical community feel that they are unfairly judged in research assessment, arguing that they do not receive enough credit given the influence on the research community of their published work (32). Our analysis supports this sentiment, showing that a significantly larger number of researchers, particularly women and underrepresented groups, would gain greater recognition through article-level metrics.

In conclusion, the current system, which prioritizes journal prestige, systematically overlooks the contributions of many influential researchers. The artificial scarcity of recognition enforced by journal prestige not only diminishes the visibility of influential work but systematically excludes a large number of highly impactful scientists from receiving the recognition they deserve. Our study highlights the need for broader adoption of article-level metrics to foster a more equitable and accurate recognition system in science in order to build a competitive scientific workforce.

Like many other studies, our research is not immune to limitations. One caveat to this work is that we do not examine the extent to which past influence predicts future influence, though research along these lines has already been conducted. Furthermore, while we focus on citation impact of scholars and scholarships, the ultimate beneficiaries of biomedical research are patients, and there is growing interest in evaluating research based on societal outcomes. Future research could explore how article-level measures align with societal impact, particularly in clinical settings.

Methods

Publication profiles

Publicly available data on funded NIH grants were downloaded in bulk from NIH ExPORTER (33). From these, names of principal investigators (who could be graduate students on training grants through senior investigators funded on long-term research project grants) were extracted. From these names, those investigators who opted in to posting a public publication profile on MyNCBI (11) were crawled using the MyNCBI URL structure that incorporates the name representation found in NIH ExPORTER data. PubMed (34) identifier numbers for Medline papers were extracted from those public sites to construct a profile of each researcher's publications. The publicly available WRU (35) (add Imai et al 2022 in Science Advances ref) and gendeR (36) R packages were used for race and gender imputation. We define the early career researchers as those who have a career age of five years or less and mid-career researchers as

those who have a career age of ten years or less but more than five (37). The rest are categorized as senior researchers.

Identifying prestigious articles using journal- and article-level metrics

Biomedicine has long recognized papers published in prestigious venues as influential. Many have advanced the view that papers that are as highly cited as those appearing in such prestigious journals should also be recognized as influential. We used a common heuristic, impact factor thresholding (6), to identify papers published in prestigious journals (i.e. using an impact factor threshold of 15, papers published in journals with an impact factor of 15 or more would be recognized as influential in a journal-level recognition regime). We used the journal citation rate in the iCite database as our measurement of journal impact factors (5, 7, 38-42).

Under the sensibility that papers that are as highly cited as those appearing in such prestigious journals should also be recognized as influential, we used the NIH Relative Citation Ratio (RCR) (5, 7), which measures influence as the number of citations per year an article has received, accounting for its field of research and publication year. Only articles tagged as research articles, which excludes article types such as reviews and editorials, were considered.

For a given impact factor threshold, articles were flagged as influential using journal-level metrics ("Journal Elites") if they were published in a journal whose impact factor exceeded the chosen threshold in the year of the article's publication.

For article level metrics, the following procedure was used to identify papers recognized as influential ("Citation Elites") as those in a prestigious journal selected by the given impact factor threshold:

1. Identify papers published in a journal whose impact factor exceeded the chosen threshold
2. Measure the median RCR of papers published in these high impact factor journals (RCR_{med})
3. Flag as influential any paper exceeding RCR_{med}

The same procedure was used for analyses using citations/year, substituting that measure for RCR as a robustness check (see Supplemental Figure 2).

Data Collection

For this analysis, we pulled 50422 publicly available author profiles from National Center for Biotechnology Information (NCBI) in 2021 that matched NIH-funded investigators. For each author, we extracted the following features to conduct our analysis:

start year(int64): The year with the earliest publication of the author as per the database record

end year(int64): The year with the latest publication of the author as per the database record

decade(int64): The decade of the earliest publication of the author

article level metric(float64): fraction of papers of the author that has rcr equal to or higher than the median rcr of papers published above the specific JIF threshold

journal level metric(float64): fraction of papers of the author that meet or exceed the specific impact factor threshold

article level metric_rank(float64): percentile rank of article level metric

journal level metric_rank(float64): percentile rank of journal level metric

rcr_median(float64): median rcr of the papers with

appl(int64): application id associated with the name from the database

surname(object): the surname of the author

forename(object): the forename of the author

forename_longest(object): for multiple forenames, the one with the most characters

male(float64): proportion of male names based on U.S. census data with the forename_longest of the author using R's gendeR package

female(float64): proportion of female names based on U.S. census data with the forename_longest of the author using R's gendeR package

gender(categorical): defined based on the largest of male and female

white(float64): the probability of being white inferred from surname using R's wru package

black(float64): the probability of being black inferred from surname using R's wru package

his(float64): the probability of being Hispanic inferred from surname using R's wru package

asi(float64): the probability of being Asian inferred from surname using R's wru package

oth(float64): the probability of being anything other than white, black, hispanic, and asian inferred from surname using R's wru package

gender(categorical): defined based on the largest of the race probabilities

art_better(bool): Binary class for whether a researcher appears more competitive under article level metric compared to journal level metric

journ_better(bool): Binary class for whether a researcher appears more competitive under journal level metric compared to article level metric

art_journ_same(bool): Binary class for whether journal level metric equals to article level metric

Data availability

Article-level data used in this paper are available at Figshare (https://doi.org/10.35092/yhjc.c.4586573). Anonymized author-level derivative data are available at Figshare (https://figshare.com/s/fbe6dd959928a28367c3).

Supplemental Materials

Statistical Analysis

Threshold	Race	P value
	Asian	< 0.0001

Threshold 10	Black	< 0.0001
	White	< 0.0001
	Hispanic	< 0.0001
	Asian	< 0.0001
Threshold 15	Black	< 0.0001
	White	< 0.0001
	Hispanic	< 0.0001
	Asian	< 0.0001
Threshold 20	Black	< 0.0001
	White	< 0.0001
	Hispanic	< 0.0001

Supplement Table 1: Binomial test on race: We conducted 2-sided binomial test and used binom_test function in the stats module from scipy library and python 3.8.8 version for the binomial test. All 4 racial categories showed significance across all three thresholds (impact factor >= 10, 15, or 20, respectively).

Threshold	chi sq statistic	p-value
10	61.77	< 0.0001
15	40.57	< 0.001
20	9.87	0.0016

Supplemental Table 2: Chi sq test on gender (all 3 thresholds showed significance): We used python's stats module from scipy library and python 3.8.8 version for the Chi-square test.

Threshold	chi sq statistic	p-value
10	96.81	< 0.0001
15	50.026	< 0.0001
20	30.8199	< 0.0001

Supplemental Table 3: Chi-square test on race: We used python's stats module from scipy library and python 3.8.8 version for the Chi-square test. We classified each scientist in a particular race if the probability of the researcher being in that racial group is greater than 0.50.

Career Stage	P-value
Early	< 0.0001
Mid	0.0
Senior	0.0

Supplemental Table 4: Paired sample t-test on Citation Elites vs Journal Elites based on career stage: We conducted a matched-pair t-test for each career stage across all the thresholds. We used python's stats module from the scipy library and the python 3.8.8 version for the test. We define early career researchers as those who have a career age of five years or less and mid-career researchers as those who have a career age of ten years or less but more than five. The rest are categorized as senior researchers.

	Asian	Black	White	Hispanic
Journal Elite	495	16	1687	65
Citation Elite	5244	319	25713	1051

Supplemental Table 5. Investigators stratified by race whose papers are most frequently recognized in the Journal Elite vs. Citation Elite categories. Chi sq table for race at an impact factor threshold of 15. Chi-sq: 50.026, p value: < 0.0001

	Female	Male
Journal Elite	592	1351
Citation Elite	10909	18017

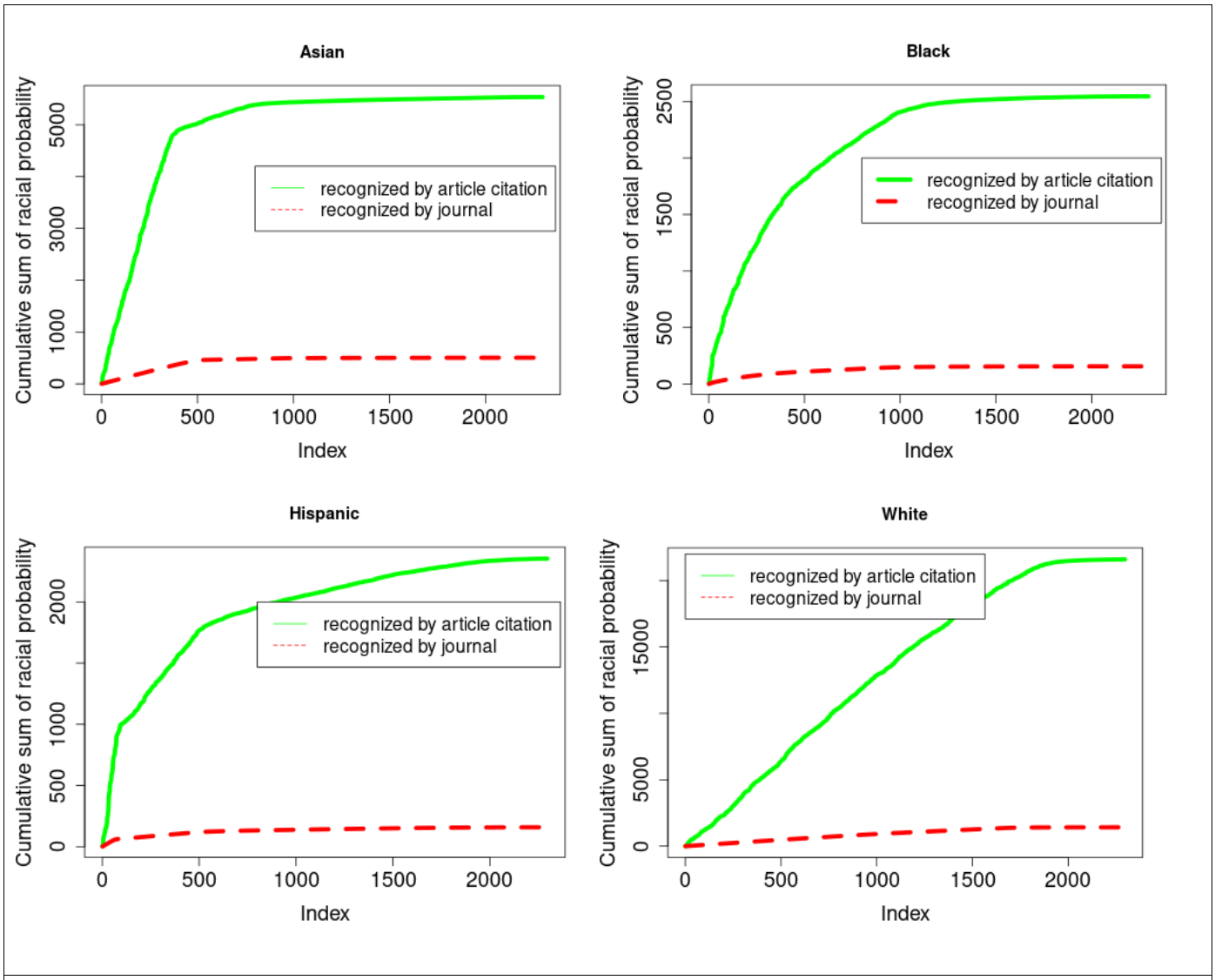
Supplemental Table 6. Investigators stratified by gender whose papers are most frequently recognized in the Journal Elite vs. Citation Elite categories. Chi sq: 40.575, p value: < 0.0001

	Early	Mid	Senior
Journal Elite	71	201	2026
Citation Elite	717	2813	29341

Supplemental Table 7: Investigators stratified by career stage whose papers are most frequently recognized in the Journal Elite vs. Citation Elite categories. Chi-sq: 8.289, p-value: 0.0158.

Threshold (Impact Factor 15+)	p-value
Asian	< 2.2e-16
Black	< 2.2e-16
Hispanic	< 2.2e-16
White	< 2.2e-16

Supplemental Table 8: KS Test on Racial Probability:
We conducted 2-sided k-s test on each racial probability score for scientists who receive more recognition under article vs journal-level metrics (also see Supplemental Figure 1). The p-value of the k-s test is significant across all three thresholds (impact factor >= 10, 15, or 20, respectively).



Supplemental Figure 1: Cumulative probability distribution of racial score of scientists who are more recognized as Citation Elites vs. Journal Elites. p-values shown in Supplemental Table 8.

Race	**p-value**
Asian	< 0.0001
Black	0.19
White	< 0.0001
Hispanic	0.33

Supplement Table 9: Proportion-z test on race: We conducted 2-sided proportion z test across all racial categories using proportions_ztest function in the stats module from statsmodels library and python 3.8.8 version for the test. We get similar significance across all thresholds.

All racial groups benefit from article-level metrics. Many scientists would receive more recognition under the article-level evaluation metric and this trend persists across all the racial groups. We conducted a 2-sided k-s test on each racial probability score for scientists who receive more recognition under article-level RCR vs journal-level metrics. The p-value of the k-s test is significant across all three thresholds (impact factor >= 10, 15, or 20, respectively). We also conducted 2-sided binomial tests on each racial group for both article-level RCR and article citation rate. The p-value of the test is significant across ass three thresholds (impact factor >= 10, 15, or 20, respectively) suggesting that a higher proportion of people will benefit from ALM for each racial group.

Though our analysis suggests that each racial group will get more recognition under the ALM, this magnitude of benefit could be different for different racial groups (Supplement Table 9). For Blacks and Hispanics, there is no significant difference in the proportion of the particular racial group in the citation elite vs journal elite group across different thresholds (impact factor >= 10, 15, 20) for both article-level metrics - RCR and article-citation rate. For Asians, the proportion of Asians in the citation elite is significantly lower than the proportion of Asians in the journal elite group across different thresholds (impact factor >= 10, 15, 20) for RCR. For whites, the proportion of whites in the citation elite is significantly higher than the proportion of whites in the journal elite group for all thresholds (10, 15, 20) for RCR. However, we see mixed results for ACR. Interestingly, it suggests that the benefit of ALM is distributed across different racial groups, but favors different groups differently.

Supplement Figure 2: Demographic analysis of scientists who receive more recognition with journal- vs. article-level measures (Article-citation rate). (Top, within box) Breakdown of how many scientists filtered by those who have more Citation Elite (these scientists are labeled as Citation Elites) papers than Journal Elite papers (labeled as Journal Elites) are male vs. female. (Top, outside box) Breakdown of how many scientists filtered by gender receive more recognition with article-level citations (Citation Elite) vs. journal impact factor (Journal Elite). (Middle) Breakdown filtered by race. (Bottom) Breakdown filtered by seniority.

Correlation Analysis

We analyzed the correlation between the metric and publication per year. We observe a weak but significant correlation between article-level metric and the publication per year. The Pearson correlation coefficient between article-level metric and publication per year is 0.08 with a p-value < 0.001. However, there is no significant correlation between journal-level metric and publication per year (Pearson correlation coefficient 0.008, p-value = 0.058).

Subgroup Analysis

Category	Metric	male	female
Top 10k profile (ALM)	Citation Elite Journal Elite	64.33 69.44	35.67 30.54
Top 10k profile (JIF)	Citation Elite Journal Elite	70.55 69.38	29.45 30.61
Bottom 10k profile (JIF)	Citation Elite Journal Elite	69.12 70.19	30.33 28.85

Supplemental Table 10: Percentage of male and female scientists in citation elite and journal elite.

Category	Gender Group	Citation Elite	Journal Elite
Top 10k profile (ALM)	Female Male	97.41 96.76	2.589 3.24
Top 10k profile (JIF)	Female Male	76.53 77.51	23.47 22.49
Bottom 10k profile (JIF)	Female Male	98.77 98.67	1.22 1.33

Supplemental Table 11: Percentage of citation elite and journal elite across gender groups.

Category	Racial Group	Citation Elite	Journal Elite
Top 10k profile (ALM)	Asian Black White Hispanic	95.06 96.15 97.07 97.40	4.94 3.85 2.93 2.59

Top 10k profile (JIF)	Asian Black White Hispanic	72.27 81.36 77.70 77.64	27.72 18.64 22.29 22.36
Bottom 10k profile (JIF)	Asian Black White Hispanic	98.55 98.91 98.69 99.14	1.45 1.09 1.31 0.855

Supplemental Table 12: Percentage of citation elite and journal elite across racial groups.

Category	Career Group	Citation Elite	Journal Elite
Top 10k profile (ALM)	Early Mid Senior	96.43 97.13 96.75	3.56 2.86 3.25
Top 10k profile (JIF)	Early Mid Senior	56.79 72.67 77.51	43.21 27.32 22.49
Bottom 10k profile (JIF)	Early Mid Senior	100 98.73 98.66	0.00 1.27 1.34

Supplemental Table 13: Percentage of citation elite and journal elite across career groups.

Category	Citation Elite	Journal Elite
Top 10k profile (ALM)	96.78	3.22
Top 10k profile (JIF)	76.69	23.31
Bottom 10k profile (JIF)	98.66	1.3

Supplemental Table 14: Percentage of citation elite and journal elite across the entire dataset.

Threshold 15

Test	Hits	Acr
Binomial test	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)
Ks test	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)

Supplemental Table 15: Statistical tests on threshold 15.

Threshold 10

Test	Hits	Acr

Binomial test	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)
Ks test	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)

Supplemental Table 16: Statistical tests on threshold 10.

Threshold 20

Test	Hits	Acr
Binomial test	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (<	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-

	0.0001)	sided p (< 0.0001)
Ks test	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)	Asian (hits greater - sig) - 2-sided p (< 0.0001) Black (hits greater - sig) - 2-sided p (< 0.0001) White (hits greater - sig) - 2-sided p (< 0.0001) Hispanic (hits greater - sig) - 2-sided p (< 0.0001)

Supplemental Table 17: Statistical tests on threshold 20.

Acknowledgements

Support for this work was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation.

References

1. Bertuzzi S, Drubin DG. No shortcuts for research assessment. Molecular Biology of the Cell. 2013;24(10):1505-6.
2. Alberts B, Kirschner MW, Tilghman S, Varmus H. Rescuing US biomedical research from its systemic flaws. Proceedings of the National Academy of Sciences. 2014;111(16):5773-7.
3. Larivière V, Sugimoto CR. The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects. Springer Handbook of Science and Technology Indicators. Springer Handbooks2019. p. 3-24.
4. Santangelo GM. Article-level assessment of influence and translation in biomedical research. Mol Biol Cell. 2017;28(11):1401-8.
5. Hutchins BI, Yuan X, Anderson JM, Santangelo GM. Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. PLoS Biol. 2016;14(9):e1002541.
6. Johnston M. We have met the enemy, and it is us. Genetics. 2013;194(4):791-2.
7. Hutchins BI, Hoppe TA, Meseroll RA, Anderson JM, Santangelo GM. Additional support for RCR: A validated article-level measure of scientific influence. PLoS Biol. 2017;15(10):e2003552.
8. Yu H, Willis KA, Litovitz A, Harriman RM, Davis MT, Meyer P, et al. The effect of mentee and mentor gender on scientific productivity of applicants for NIH training fellowships. bioRxiv. 2021.
9. Hoppe TA, Arabi S, Hutchins BI. Predicting substantive biomedical citations without full text. Proceedings of the National Academy of Sciences. 2023;120(30).
10. Lozano GA, Larivière V, Gingras Y. The weakening relationship between the impact factor and papers' citations in the digital age. Journal of the American Society for Information Science and Technology. 2012;63(11):2140-5.
11. National Center for Biotechnology Information: National Library of Medicine; [Available from: https://www.ncbi.nlm.nih.gov/.

12. Basson I, Ni C, Badia G, Tufenkji N, Sugimoto CR, Larivière V. Gender differences in submission behavior exacerbate publication disparities in elite journals. Elife. 2023;12:RP90049.
13. Harris A. Young, Brilliant, and Underfunded. New York Times. 2014.
14. Hoppe TA, Litovitz A, Willis KA, Meseroll RA, Perkins MJ, Hutchins BI, et al. Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. Sci Adv. 2019;5(10):eaaw7238.
15. Bornmann L, Kozlowski D, Murray DS, Bell A, Hulsey W, Larivière V, et al. Avoiding bias when inferring race using name-based approaches. Plos One. 2022;17(3).
16. Lockhart JW, King MM, Munsch C. Name-based demographic inference and the unequal distribution of misrecognition. Nature Human Behaviour. 2023;7(7):1084-95.
17. Lockhart JW, King MM, Munsch CL. Computer algorithms infer gender, race and ethnicity. Here's how to avoid their pitfalls. Nature. 2023.
18. Lauer M. Extramural Nexus [Internet]: National Institutes of Health. 2021. Available from: https://nexus.od.nih.gov/all/2021/11/18/long-term-trends-in-the-age-of-principal-investigators-supported-for-the-first-time-on-nih-r01-awards/.
19. Heidt A. Racial inequalities in journals highlighted in giant study. Nature. 2023.
20. Liu F, Rahwan T, AlShebli B. Non-White scientists appear on fewer editorial boards, spend more time under review, and receive fewer citations. Proceedings of the National Academy of Sciences. 2023;120(13).
21. Lozano S, Bendels MHK, Müller R, Brueggmann D, Groneberg DA. Gender disparities in high-quality research revealed by Nature Index journals. Plos One. 2018;13(1).
22. Ceci SJ, Ginther DK, Kahn S, Williams WM. Women in Academic Science. Psychological Science in the Public Interest. 2014;15(3):75-141.
23. Larivière V, Ni C, Gingras Y, Cronin B, Sugimoto CR. Bibliometrics: Global gender disparities in science. Nature. 2013;504(7479):211-3.
24. Hearings before a subcommittee of the Committee on Appropriations: Hearing before the Subcommittee on Labor, Health and Human Services, Education, and Related Agencies, House of Representatives, One Hundred Fifteenth Congress, First Session Sess. (2017).
25. Collins FS. Testimony on the Implementation of the 21st Century Cures Act: Progress and the Path Forward for Medical Innovation. In: Health NIo, editor. 2017.
26. Faculty Institutional Recruitment for Sustainable Transformation (FIRST) NIH Common Fund2021 [Available from: https://www.commonfund.nih.gov/first.
27. Lauer MS, Roychowdhury D. Inequalities in the distribution of National Institutes of Health research project grant funding. Elife. 2021;10.
28. Squazzoni F, Bravo G, Farjam M, Marusic A, Mehmani B, Willis M, et al. Peer review and gender bias: A study on 145 scholarly journals. Science Advances. 2021;7(2).
29. Casad BJ, Franks JE, Garasky CE, Kittleman MM, Roesler AC, Hall DY, et al. Gender inequality in academia: Problems and solutions for women faculty in STEM. J Neurosci Res. 2021;99(1):13-23.
30. Day AE, Corbett P, Boyle J. Is there a gender gap in chemical sciences scholarly communication? Chem Sci. 2020;11(8):2277-301.
31. Holman L, Stuart-Fox D, Hauser CE. The gender gap in science: How long until women are equally represented? Plos Biology. 2018;16(4).
32. Berenbaum MR. Impact factor impacts on early-career scientist careers. Proceedings of the National Academy of Sciences. 2019;116(34):16659-62.
33. Health NIo. ExPORTER 2021 [Available from: https://exporter.nih.gov/.
34. Medicine NLo. Download MEDLINE/PubMed Data 2020 [Available from: https://www.nlm.nih.gov/databases/download/pubmed_medline.html.
35. Imai K, Khanna K. Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. Political Analysis. 2017;24(2):263-72.

36. Blevins C, Mullen LA. Jane, John ... Leslie? A Historical Method for Algorithmic Gender Prediction. Digit Humanit Q. 2015;9.
37. Boothby C, Milojevic S, Larivière V, Radicchi F, Sugimoto CR, editors. Consistent churn of early career researchers: an analysis of turnover and replacement in the scientific workforce2022.
38. iCite: National Institutes of Health; 2015 [Available from: https://icite.od.nih.gov/.
39. Hutchins BI. A tipping point for open citation data. Quantitative Science Studies. 2021:1-5.
40. Hutchins BI, Baker KL, Davis MT, Diwersy MA, Haque E, Harriman RM, et al. The NIH Open Citation Collection: A public access, broad coverage resource. PLoS Biol. 2019;17(10):e3000385.
41. Hutchins BI, Davis MT, Meseroll RA, Santangelo GM. Predicting translational progress in biomedical research. PLoS Biol. 2019;17(10):e3000416.
42. iCite, Hutchins BI, Santangelo GM. iCite Database Snapshots (NIH Open Citation Collection). In: Health NIo, editor. Figshare2019.
43. Schimanski, L. A, & Alperin, J. P. (2018). The evaluation of scholarship in academic promotion and tenure processes: Past, present, and future. F1000Research, 7, 1605. https://doi.org/10.12688/f1000research.16493.1
44. Rice DB, Raffoul H, Ioannidis J P A, Moher D. Academic criteria for promotion and tenure in biomedical sciences faculties: cross sectional analysis of international sample of universities BMJ 2020; 369 :m2081 doi:10.1136/bmj.m2081
45. Walker, R. L., Sykes, L., Hemmelgarn, B. R., & Quan, H. (2010). Authors' opinions on publication in relation to annual performance assessment. BMC medical education, 10, 21. https://doi.org/10.1186/1472-6920-10-21
46. Danielle B. R., Raffoul H, Ioannidis J P A, and Moher D. 2021. Academic criteria for promotion and tenure in faculties of medicine: a cross-sectional study of the Canadian U15 universities. FACETS. 6(): 58-70. https://doi.org/10.1139/facets-2020-0044
47. McKiernan EC, Schimanski LA, Muñoz Nieves C, Matthias L, Niles MT, Alperin JP. Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. Elife. 2019 Jul 31;8:e47338. doi: 10.7554/eLife.47338. PMID: 31364991; PMCID: PMC6668985.