F1000Prime recommended articles and their citations: an exploratory study of four journals

Peiling Wang¹ · Joshua Williams¹ · Nan Zhang² · Qiang Wu²

Received: 9 April 2019 / Published online: 30 November 2019
© The Author(s) 2019

Abstract

This study examined F1000Prime recommended research and review articles published in Cell, JAMA: The Journal of the American Medical Association, The Lancet, and The New England Journal of Medicine (NEJM) in 2010. The analyses included (1) the classifications assigned to the articles; (2) differences in Web of Science (WoS) citation counts over 9 years between the articles with F1000Prime recommendations and the other articles of the same journal; (3) correlations between the F1000Prime rating scores and WoS citation counts; (4) scaled graphic comparisons of the two measures; (5) content analysis of the top 5 WoS cited and top 5 F1000Prime scored NEJM articles. The results show that most of the recommended articles were classified as New Finding, Clinical Trial, Conformation, Interesting Hypothesis, and Technical Advance. The top classifications differed between the medical journals (JAMA, The Lancet, and NEJM) and the biology journal (Cell); for the latter, both New Finding and Interesting Hypothesis occurred more frequently than the three medical journals. The articles recommended by F1000 Faculty members were cited significantly more than other articles of the same journal for the three medical journals, but no significance was found between the two sets of articles in Cell. The correlations between the F1000Prime rating scores and WoS citation counts of the articles in the same journal were significant for the two medical journals (The Lancet and NEJM) and the biology journal (Cell). NEJM showed significances in both the upper quantile (top 50%), and the upper quartile (top 25%) sets. One of the medical journals, JAMA, did not show any significant correlation between the two measures. Despite the significant correlations of the three journals, Min–Max scaled graphic comparisons of the two measures did not reveal any patterns for predicting citation trends by F1000Prime rating scores. The peak citation year of the articles ranged from 2 to 8 years after the publication year for NEJM. Content analysis of the top-cited and top-scored NEJM articles found that highly commendable papers with comments such as “exceptional,” “landmark study,” or “paradigm shift” received varied rating scores. In comparison, some of the results corroborate with previous studies. Further studies are suggested to include additional journals and different years as well as alternative methods. Studies are needed to understand how F1000 Faculty assign ratings and what criteria they use. In addition, it is also worth investigating how F1000Prime users perceive the meanings of the ratings.

Keywords F1000Prime rating scores · F1000 Faculty · Citation analysis · Post-publication peer recommendation · Expert recommendation · Scaled comparisons · Content analysis of F1000Prime comments

Introduction

A challenge facing scientists in the Big Science era is the information explosion, a phenomenon predicted more than 50 years ago (de Solla Price 1961). Bornmann and Mutz (2015) report that the growth of global publication output is at a rate of approximately 3% yearly. Researchers use criteria beyond topical relevance to select relevant and useful information to read or cite. These criteria include high impact journals, known authors, and being cited by known researchers (Dorch 2012; Small 2004; Tressoldi et al. 2013; Wang and White 1999). As scientific literature continues to proliferate, users of bibliographic databases tend to get more results than they are willing to go through. Researchers at all levels feel the need for filtering tools simply because senior researchers no longer have the time to read every journal in their field, and junior researchers lack the required knowledge or expertise to judge the quality of the publications (Pontis et al. 2015). To satisfy this need, F1000 invented a new service to provide post-publication peer recommendations. In 2009, F1000 merged its two earlier services, F1000 Biology (inaugurated in 2002) and F1000 Medicine (launched in 2006), into one single database, F1000Prime. Access to this database requires a subscription. Recommendations are submitted by F1000Prime Faculty members who were nominated by peers as experts in the fields. These faculty members (FMs) can have their designated associates (AFMs) to co-review an article (Fig. 1). For each recommendation, the FM (maybe co-authored with the AFM) assigns a rating score (1 star—Good, 2 stars—Very Good, or 3 stars—Exceptional), classifies the article from a predefined list of classifications (e.g., New Finding, Controversial, Good for Teaching, etc.), and writes a comment. An article may receive multiple recommendations from which scores are aggregated. In addition to recommendations, F1000 faculty members can submit a comment as Dissent without a rating score.

The three main reasons for using F1000Prime may include (1) to identify key papers rapidly; (2) to locate important papers published in journals which the scientist does not usually read; (3) to share peer opinions (Wets et al. 2003). For new researchers, F1000Prime recommendations can help them recognize significant publications. Post-publication peer recommendation can be considered an extension of the pre-publication peer review for acceptance or rejection (Benos et al. 2007); peer review comments in closed-referee systems are generally not available to readers (Hirschauer 2015). F1000Prime provides a platform for sharing experts' evaluations of published research (Vardell and Swogger 2014). F1000Prime also has the potential as an alternative to traditional metrics for assessing the quality of research outputs (Bornmann and Leydesdorff 2013) or to recognize significant research results as soon as they are published (Mohammadi and Thelwall 2013).

The purpose of this study is to understand what the F1000 Faculty recommend and how the F1000Prime recommended articles are being cited. Furthermore, the study also takes a closer look at the relationship between the two measures, F1000Prime rating scores and citation counts. To conduct an in-depth analysis of the two sets of articles from the same journal to shed light upon experts' ratings and citation patterns, our data collection focused on research and review articles published in *Cell*, *JAMA: The Journal of the American Medical Association (JAMA)*, *The Lancet*, and the *New England*

Fig. 1 A F1000Prime recommended NEJM article with the highest total stars

Journal of Medicine (NEJM) in 2010. These journals were selected based on several factors. Starting from the Top 100 and Top 10 journals in biology and medicine by the Biomedical and Life Sciences Division of the Special Library Association in 2008 and 2009 (<http://dbiosla.org/publications/resources/dbio100.html>; <http://dbiosla.org/publications/resources/topten.html>), we considered the journal's impact factor (JIF), readership or library subscriptions, subject coverage, journal longevity, and F1000Prime recommendations. The publication year 2010 was chosen because F1000 merged two previous databases in 2009 into F1000Prime, and faculty members most likely read and recommend publications in current issues. Additionally, choosing 2010 instead of later or multiple publication years allowed a longer citation window of 8 years. Finally, Focusing on research and review articles should minimize the influence of article types.

The specific research questions related to the data collected from the four journals (Cell, JAMA, The Lancet, and NEJM) in 2010 are as follows:

- RQ1 What classifications do the F1000 Faculty assign to the articles they recommend?
- RQ2 What are the differences in Web of Science (WoS) citation counts from 2010 to 2018 between F1000 Faculty recommended articles and other articles?
- RQ3 What is the relationship between the F1000Prime rating and the WoS citation count? Which percentile-based sets show significant correlations between the two measures?

- RQ4 What are the patterns of the scaled F1000Prime rating scores and WoS citation counts?
- RQ5 What comments do highly-recommended or highly-cited articles receive? How are the FM comments mapped to the rating scores?

Literature review

Due to space limits, this literature review is not intended to be comprehensive; rather, it will cover major studies from which the concepts or methods are being adopted and results compared. In one of the early studies of F1000Prime, Allen et al. (2009) found a strong positive association between F1000 Faculty rating scores and citations based on a set of 687 research papers associated with Wellcome Trust, with the exception that highly rated papers were not highly cited during the first 3 years after publication. They caution that research assessments relying on single indicators might miss important work. Wardle (2010) examined 1530 articles published in 7 major ecological journals in 2005 to compare F1000Prime ratings with citation counts. He found that F1000Prime articles performed poorly in predicting subsequent citation counts because some highly cited publications were not recommended, although approximately one-third of the recommended articles were cited less frequently.

Bornmann and Leydesdorff (2013) studied whether F1000prime scores, as the paper's quality proxies, correlate with citation-based metrics. The authors analyzed a set of 125 F1000Prime recommended papers in cell biology or immunology published in 2008. They compared these articles' FFa scores and other indicators such as journal impact factor and citation counts. They found that percentile in the subject area as a normalized indicator had the highest correlation with FFa scores. (FFa, calculated using a complex formula to aggregate the rating scores, 1 for Recommended, 2 for Must read, 3 for Exceptional, by individual FMs and updated whenever a new recommendation is submitted; FFa is no longer listed alongside the paper by F1000. Two of the three labels for ratings also changed.) In another study, Bornmann and Leydesdorff (2015) examined the relative influences of F1000Prime scores on citation counts over 10 years. They analyzed citations to 9898 F1000Prime recommended papers published between 2000 and 2004 and cited until 2013. Contrary to their expectations, their results suggested that F1000Prime rating scores as quality proxies of the papers played less of a role in the later citation counts than journal impact factor. Also, as a quality proxy, Bornmann and Haunschild (2018) found that the factor (citation and reader impact) was strongly correlated to paper quality using the negative binomial regression between F1000Prime scores and other citation-based metrics. They acknowledged the limitation of using F1000Prime data as a proxy for paper's quality.

The effect of classifications by F1000 Faculty on FFa scores and citation counts studied by Mohammadi and Thelwall (2013) showed that citation counts and FFa scores were significantly different for two classifications: New Finding and Changes Clinical Practice. They suggest that F1000Prime is a useful tool to perform quick evaluations and to recognize research outputs that may not be acknowledged by citations assuming that the majority of the articles F1000 Faculty review are in the same publication year.

Waltman and Costas (2014) conducted a large-scale study of F1000Prime rating scores with citations using a dataset from F1000 in 2013. The rating scores in the 132,622 recommendations related to 102,360 publications had a distribution of 58.6% for 1 star, 34.6% for 2 stars, and 6.9% for 3 stars. Using Web of Science, the researchers collected a second

dataset of citations from the 1.7 million publications published from 2006 to 2009 using a 3-year citation window to compare the F1000Prime recommended publications with the publications without F1000Prime recommendations in the same period. They found a relatively weak correlation between F1000Prime scores and citation counts and suggested further research to identify the differences between the two measures.

Du et al. (2015) examined the effect of research level or article type on F1000Prime recommendations and citations. Levels of research articles were classified as basic research, clinical, or mixed. The type of articles was either evidence-based or transformative. The researchers collected 47,446 F1000Prime recommended articles (set 1) published between 1999 and 2013 in three biology-medicine pairs of subject areas. Using a citation window of 3–5 years after publication, they collected citations to 28,254 articles from Scopus (set 2). Their findings suggested that research level had little impact, but article type affected correlations: (1) evidence-based research publications were cited more than transformative research publications; (2) transformative research publications were recommended more than evidence-based research publications.

In a study of the datasets from F1000Prime and Mendeley, Bornmann and Haunschild (2015) found that the majority of the recommendations were classified as New Finding (67%) and rated as Good (1 star at 46%) or Very Good (2 stars at 44%). Based on what records Mendeley users save, they found that all user groups except for the lecturer group were interested in articles classified as New Finding. Surprisingly, Good for Teaching was of interest to the lecturer group and the non-academic research group.

Although F1000Prime is a relatively new phenomenon in scientific communication, researchers have developed many strategies to deal with methodological challenges in studying its impact. In addition to testing the entire dataset, Bornmann et al. (2013) developed strategies to test subsets based on percentiles. A subset of data can be defined by rank-ordered measures such as top 5%, upper quantile (50%), etc. For normal distributions, Mean Normalized Citation Score (MNCS) has mostly been used for large datasets and disciplinary analysis as the cited-side indicators (Bornmann and Marx 2015). In a large-scale study, Bornmann and Haunschild (2016) investigated the correlation between relative citation ratio (RCR) and F1000Prime scores of 16,521 articles. RCR was one of the four measures derived from raw citation counts in the study. They found relatively low correlations between F1000Prime scores and RCR. A new indicator, Mantel–Haenszel quotient (MHq), was recently proposed and tested by the same researchers (Haunschild and Bornmann 2018) to deal with the sparse data problem. Li et al. (2011) filled in the nonexistent items in a dataset with zero values to match items with another dataset in which they ran comparisons. Bornmann (2015) tested F1000Prime ratings assigned by different FMs to the same articles and found a low level of inter-rater's agreement.

To examine the citation time window, Wang (2013) collected citation data from 14 disciplines over 31 years. Wang's analyses found that in biology, citations of journal articles started rising very late and lasted very long while in clinical medicine, citations of journal articles started rising quickly and lasted very long. Wang also found citation windows differed by document types. Although editorials and letters had similar early rising, editorials lasted much shorter than letters. The implication is that choosing a citation window should consider the disciplinary field and document type.

Although the studies of F1000Prime have answered many interesting questions about classifications of recommended articles, patterns of ratings by FMs, and relationships between rating scores and citation counts, there are still many unanswered questions for researchers in both information science and scientometrics. While the large datasets achieved by many studies provided a big picture and statistically validated results,

smaller and more focused datasets should also be considered because the advantage of in-depth analysis of the outliers can provide additional understanding of a very complex phenomenon.

Research methods

In this study, we collected data from five bibliographic sources: F1000Prime.com, Clarivate Web of Science (WoS), Elsevier Scopus, Elsevier ScienceDirect, and each journal's Web portal. In January 2019, we collected two sets of data from four journals: Cell, JAMA, The Lancet, and NEJM. The rationales for the journal selection are provided in the Introduction. (to briefly recap: We intended to get relatively homogeneous datasets from single high-impact journals in the year following the merging of two F1000 databases into F1000Prime). Table 1 summarizes the data.

Search strategies and collected data

F1000Prime is a Web database that provides advanced search functions. Each journal was searched and limited to the publication year 2010 for all recommended articles using the advanced search (e.g., Journal: N Engl J Med AND Year published: 2010–2010). The scraped data include DOI, URL, Title, Authors, Publication Date; for each recommendation, the data include Date (Fig. 1), rating (in words and stars), Classified as (multi-values), and a comment.

WoS provides an advanced search function that was used as a starting point to retrieve all the articles published in each journal in 2010; the output included annual citation counts from 2010 to 2018. In the example in Fig. 2, the searches were further limited to research or review articles and excluded Letters, Editorials, Corrections, etc. The output included 232 records, and each had the fields: DOI, Title, Authors, Publication Date, Total Cites, Cites for each year (2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018). Scopus provides functions similar to WoS except for not including DOI field in its citation output. A

Table 1 Journals with recommended articles in F1000Prime

Journals	JIF (2010)	Holding libraries (OCLC data)	Readership or Online
Cell (1974-)	32.406	1713	> 817,000 ^a
Journal of the American Medical Association (1883-)	30.011	3752	> 1,500,000 ^b
The Lancet (1923-)	33.633	3548	> 12,000,000 ^c
New England Journal of Medicine ^a (1812-)	53.486	4190	> 600,000 ^d

^aMonthly online visit in a media kit because the journal does not share readership data (<https://media-kit-2019.elsevierdigitaledition.com/>)

^bData provided by the journal at <https://jamanetwork.com/journals/jama/pages/for-authors#fa-why-publish> (286,000 printer recipients; 1.3 million e-recipients of ToC)

^cThe Lancet reports 84 million full-text downloads since 2012. (https://info.thelancet.com/request-access?utm_campaign=jls-rtl&utm_source=site-footer&utm_medium=website)

^dData provided by the journal at <https://www.nejm.org/media-center/fact-sheet>

# 1	232	(SO=JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION AND PY=2010) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Review) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=2010-2018
-----	-----	---

Fig. 2 Advanced search in Web of Science

comparison of the search results of WoS with those of Scopus revealed that the trend of citedness of the articles was similar, although the numbers were slightly different. With the focus on trends and patterns, by applying normalization or rescaling to the data, the differences in numbers from the two citation databases did not affect results.

In order to exclude non-research articles such as Correspondences, Images, Editorial, Letters, etc. from the datasets, different strategies were developed for the four journals: (1) for Cell and The Lancet, we searched ScienceDirect database; (2) for JAMA, we used its web portal's Advanced Search; (3) for NEJM, we used its web portal's browsing by document type (Article or Review). NEJM has special object types such as case records, and clinical images and videos that are not peer-reviewed but assigned unique DOIs. WoS search filter for articles could not exclude these types of documents from the search output. We used NEJM's DOI codes to clean the dataset (e.g., <https://doi.org/10.1056/nejmicm0810147> is a pdf of two images in clinical medicine; <https://doi.org/10.1056/nejmra0808281> is a review article; <https://doi.org/10.1056/nejmoa1005361> is an original research article). Of the 166 F1000Prime NEJM articles, 9 were removed; 61 of the 354 WoS NEJM articles were also removed. In addition to removing non research articles from datasets, missing articles due to search functions or problematic records were identified and added. Specifically, when an F1000Prime recommended article did not have a citation record in cited reference search results from WoS, searching this article using a basic search often identified variations of the article's record. Therefore, 3 records were added to the 232 JAMA results from the initial WoS search (Fig. 2), resulting in 235 articles in the final dataset (Table 3).

Datasets

Both the WoS search outputs and scraped F1000Prime data were imported to build the database for analysis. The field DOI was the identifier to match and link the records from the two databases. Specific challenges in building the database included: (1) falsedrops in the search results due to database errors; (2) missing DOI in records (e.g., 1.2% for NEJM); (3) name variations due to diacritical marks (e.g., Römer in F1000Prime is Roemer in WoS); (4) messy data (e.g., no delimiter or space between two classifications); (5) date format variation (e.g., in F1000Prime 2010 July 3 or 16 July 2010, and in WoS July 14 2010). For problem (1), DOI codes were used to remove the misclassified articles, or the journal was checked for the correct article type. For problem (2), author or title searches filled in the missing DOIs. For problems (3) to (5), data were cleaned and transformed.

The final database includes dataset (1) F1000Prime data as summarized in Table 2 and dataset (2) citation data of the four journals between 2010 and 2018 from WoS as summarized in Table 3. Not included in significance or correlation tests: (1) F1000Prime recommended but are not research or review articles; (2) published journal articles that have no WoS citations. Each journal article belongs to one of the two sets: F1000Prime recommended (set 1) and the other articles of the same journal in the same year (set 2). Set 1 has two measures: F1000Prime rating scores and WoS citation counts.

Table 2 F1000Prime recommended articles published in 2010

Journal	Research and review articles	F1000Prime recommended	Total stars		Total recommendations			Skewness ^c
			Median	Max	Median	Max	Total	
Cell	288 ^a	193 ^b	3	15	2	7	406	1.315
JAMA	247 ^a	55	2	8	1	5	92	1.466
The Lancet	272	82 ^b	2	16	1	7	127	2.945
NEJM	293	166 ^b	3	23	2	13	370	2.339
Total	1100	496	–	–	–	–	995	–

^aThese numbers include additional articles added to the original search results based on F1000Prime results as explained in the previous section

^bThese numbers include non-research or review articles; they are excluded in set-based comparisons (Table 3)

^cSkewness (or *adjusted Fisher–Pearson standardized moment coefficient*) from SPSS descriptive statistics output measures a distribution's shape. According to Doane and Seward (2011), the following formula was adopted by SPSS $G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$ where n is the sample size, x_i is the i th data point in the sample, $\bar{x}$ is the sample mean, and s is the standard deviation. It is reasonable to interpret that a normal distribution of a dataset will have a Skewness value 0 ± 0.5 ; a highly skewed distribution will have a value > 1 (positively skewed) or a value < -1 (negatively skewed)

Table 3 Articles in two datasets by grouping factor F1000Prime recommendation

Journal	Recommended in F1000Prime?			Citations in WoS				
				Total	Median ^c	Min	Max	Skewness ^d
Cell	Y	188 ^b	65.28%	53,417	184	18	3951	6.136
	N	100	34.72%	31,812	155	9	2984	3.465
Total		288		85,229	168			4.765
JAMA ^a	Y	55	23.40%	19,077	257	5	4035	6.091
	N	180	76.60%	33,319	123	1	1936	4.042
Total		235 ^a		52,396	139			7.375
The Lancet ^b	Y	80 ^b	29.41%	36,090	356	7	2489	2.311
	N	192	70.59%	56,265	205	5	1737	2.242
Total		272		92,355	244			2.399
NEJM ^b	Y	157 ^b	53.58%	96,494	381	32	6640	2.958
	N	136	46.42%	39,793	186	18	2124	3.988
Total		293		136,287	285			4.706
Total Articles: 1088		F1000Prime: 480		366,267				

^aJAMA's 12 articles without F1000Prime recommendations or WoS citations not included in comparisons of two measures

^bFor these three journals, F1000Prime's 16 recommended non-research/review articles or retracted articles were excluded in comparisons

^cDue to skewed distributions, the median is used instead of mean; numbers were rounded up for this table only

^dThe *adjusted Fisher–Pearson standardized moment coefficient* (See Table 2 note c)

Analyses and results

The classifications assigned by the FMs were analyzed to depict the distribution. Non-parametric statistical analyses tested: (1) the differences in citation counts between F1000Prime recommended articles (set 1) and the other articles from the same journal (set 2) over 9 years; (2) correlations between the two measures of the articles in set 1; (3) scaled comparisons of two measures of the same articles in set 1. For correlational analyses, tests were run on the sets, including all articles and rank-based quantile and quartile subsets. The use of Min–Max scaling instead of Mean Normalized Citation Score (MNCS) to compare the two measures of the articles in set 1 was to consider the highly skewed data distributions for both measures. Analysis of the one superstar F1000Prime-recommended article from each journal and the one most-cited WoS article from each journal compared their ranks by the two measures. For NEJM, indepth analysis focused on the top 5 F1000Prime recommended articles and the top 5 cited articles. The two ranking criteria resulted in a set of 8 articles because 2 articles were ranked top 5 by both criteria, and 6 articles were ranked top 5 by only one criterion. Sentiment analysis focused on the comments to these articles by FMs. All statistical analyses used SPSS; the sentiment analysis used QDA Miner. The rest of the section presents results corresponding to the research questions (RQs).

Classifications of the recommended articles by F1000 faculty (RQ1)

Assigning classifications is optional in F1000Prime. A total of 995 recommendations were analyzed to observe the FMs assigned classifications. All articles had at least 1 classification by 1 FM (Fig. 1). When the same classification was assigned by more than 1 FM to the same article, this classification only counted once. On average, there were 2 classifications per article, ranging between 1 and 6 (for Cell: 1–5; JAMA: 1–5; The Lancet: 1–6; NEJM: 1–5). Figure 3 depicts the classification distribution for all the journals. The top 5 classifications were New Finding (31.4%), Clinical Trial (18.7%), Confirmation (13.6%), Interesting Hypothesis (11.6%), and Technical Advance (7.5%). For individual journals, the top classifications varied (Table 4). For example, New Finding was ranked the 1st in Cell, the 2nd in both The Lancet and NEJM, but the 3rd in JAMA; Clinical Trial was ranked the 1st

Fig. 3 F1000Prime recommended articles in Cell, JAMA, The Lancet, and NEJM

in JAMA, The Lancet and NEJM, but was nonoccurring in Cell. Good for Teaching, however, occurred only in JAMA and NEJM a few times (1.1%) (Table 4).

Differences in citations between F1000Prime articles and other articles (RQ2)

For all the journals, the citation data were highly skewed, measured by the *adjusted Fisher–Pearson standardized moment coefficient* (Table 3), with a few outliers that were highly cited (Fig. 4). The differences between the two sets of articles are visible in Fig. 4. The F1000Prime recommended articles (set 1) had a greater median, more most-cited articles, and a longer tail than the set of other articles (set 2). Therefore, non-parametric tests were performed to examine the statistical differences in citations between the two sets (Table 5). Results were significant at $p < .05$ or $p < .01$ for JAMA and NEJM for total citations and each of the 9 years. The Lancet datasets showed significant differences for total citations and most of the individual years except for 2010, which is the publication year. No significance was shown for Cell.

Table 4 Articles as classified by F1000 Faculty in 4 journals

Cell	Count (%)	JAMA	Count (%)
New Finding	142 (42.1)	Clinical Trial	22 (23.4)
Interesting Hypothesis	75 (22.3)	Confirmation	22 (23.4)
Technical Advance	44 (13.1)	New Finding	21 (22.3)
Novel Drug Target	32 (9.5)	Controversial	9 (9.6)
Confirmation	27 (8.0)	Review/Commentary	6 (6.4)
Controversial	12 (3.7)	Changes Clinical Practice	5 (5.3)
Refutation	3 (0.9)	Refutation	3 (3.2)
Review/Commentary	2 (0.6)	Technical Advance	2 (2.1)
		Interesting Hypothesis	2 (2.1)
		Systematic Review/Meta-review	1 (1.1)
		Good for Teaching	1 (1.1)
The Lancet	Count (%)	NEJM	Count (%)
Clinical Trial	48 (33.3)	Clinical Trial	103 (29.5)
New Finding	34 (23.6)	New Finding	93 (26.6)
Confirmation	30 (20.8)	Confirmation	47 (13.5)
Changes Clinical Practice	10 (6.9)	Interesting Hypothesis	25 (7.2)
Technical Advance	6 (4.2)	Changes Clinical Practice	22 (6.3)
Interesting Hypothesis	5 (3.5)	Technical Advance	17 (4.9)
Systematic Review/Meta-	4 (2.8)	Controversial	17 (4.9)
Review/Commentary	3 (2.1)	Review/Commentary	9 (2.6)
Controversial	3 (2.1)	Novel Drug Target	6 (1.7)
Refutation	1 (0.7)	Refutation	4 (1.1)
		Good for Teaching	4 (1.1)
		Systematic Review/Meta-review	1 (0.3)
		Changes Clinical Trial	1 (0.3)

Fig. 4 Citations to research and review articles published in NEJM 2010 ($N=293$)

Correlations between F1000Prime scores and WoS citations (RQ3)

Non-parametric correlational tests between F1000Prime scores and WoS citation counts are summarized in Table 6. For all the articles in set 1, three of the four journals, Cell, The Lancet and NEJM were significantly correlated at $p < .01$ between F1000Prime scores and WoS citation counts. The test of the articles in the upper quantile ($\geq$ median) was ranked by F1000Prime scores and by WoS citation counts respectively, the above three journals still showed significant correlations at $p < .05$. Towards the upper quartile (75th percentile or the top 25%), the only journal with a significant correlation at $p < .05$ was NEJM. Zooming into each journal, JAMA showed no significant correlation between the two measures for set 1 and the two subsets, while NEJM showed significant correlations for set 1 and the two subsets. The percentage of the articles that made it to the final quantile set or quartile set was always below the defined percentage indicating that the two measures ranked the articles differently.

Patterns of F1000Prime scores and citation counts by scaling (RQ4)

The results in Table 6 prompted further analyses to rescale the two measures in order to plot the trends of each measure. Because of the skewed distributions, Min–Max scaling, instead of Z-score standardization, was applied to F1000Prime scores and WoS citation counts using the following formula:

$$y' = (y - \min(y)) / (\max(y) - \min(y))$$

Along the x -axis in Fig. 5, the 188 Cell's 2010 articles were arranged in the order of F1000Prime scores descending and then by WoS citation counts also descending. The y -axis depicted the two data series using the solid blue line for the F1000Prime scaled data and the dashed orange line for the WoS citation scaled data. If the two data points

Table 5 Non-Parametric tests of differences in citation counts of the F1000Prime recommended articles versus other articles in the same journal published in 2010

All years	@2010	@2011	@2012	@2013	@2014	@2015	@2016	@2017	@2018
Cell 2010 Articles Cited between 2010 and 2018									
Citation Median	172.00	3.00	19.00	25.00	23.00	22.00	20.00	18.00	16.50
Mann–Whitney U	8413.500	8290.500	8679.000	8275.500	8902.500	8688.500	8247.500	8363.000	8478.000
Wilcoxon W	13,463.500	13,340.500	13,729.000	13,325.500	13,952.500	13,738.500	13,297.500	13,413.000	13,528.000
Asymp. Sig. (2-tailed)	.143	.097	.284	.095	.460	.231	.087	.123	.170
JAMA 2010 Articles Cited between 2010 and 2018									
Citation Median	139.00	2.00	15.00	21.00	19.00	17.00	17.00	15.00	12.00
Mann–Whitney U	2914.500	3943.000	2910.000	2836.000	2962.000	3029.500	3094.500	3026.500	3128.000
Wilcoxon W	19,204.500	20,233.000	19,200.000	19,126.000	19,252.000	19,319.500	19,384.500	19,316.500	19,418.000
Asymp. Sig. (2-tailed)	.000	.021	.000	.000	.000	.000	.000	.000	.000
The Lancet 2010 Articles Cited between 2010 and 2018									
Citation Median	243.50	5.00	23.50	31.00	29.00	30.00	29.00	25.00	21.50
Mann–Whitney U	5272.000	6685.500	5329.000	5148.500	5092.500	5656.000	5728.500	5986.500	5777.000
Wilcoxon W	23,800.000	25,213.500	23,857.000	23,676.500	23,620.500	24,184.000	24,256.500	24,514.500	24,305.000
Asymp. Sig. (2-tailed)	.000	.091	.000	.000	.000	.001	.001	.004	.001
NEJM 2010 Articles Cited between 2010 and 2018									
Citation Median	285.00	5.00	32.00	40.00	37.00	32.00	28.00	28.00	23.00
Mann–Whitney U	6635.500	8519.500	6788.500	6697.000	6678.500	6683.500	6469.000	6759.000	6847.500
Wilcoxon W	15,951.500	17,835.500	16,104.500	16,013.000	15,994.500	15,999.500	15,785.000	16,075.000	16,163.500
Asymp. Sig. (2-tailed)	.000	.003	.000	.000	.000	.000	.000	.000	.000

Cell: F1000Prime recommended articles = 188 (set 1); Other articles = 100 (set 2)

JAMA: F1000Prime recommended articles = 55 (set 1); Other articles = 180 (set 2)

The Lancet: F1000Prime recommended articles = 80 (set 1); Other articles = 192 (set 2)

NEJM: F1000Prime recommended articles = 157 (set 1); Other articles = 136 (set 2)

Table 6 Correlation between F1000Prime scores and WoS citation counts

Journal	Articles ^a		
	Correlation (Kendall's tau_b test)		
	Whole dataset	Upper quantile ^b	Upper quartile ^b
Cell	<i>n</i> = 188 .206* (<i>p</i> = .000)	<i>n</i> = 63 (33.5%) <i>.199**</i> (<i>p</i> = .030)	<i>n</i> = 36 (19.1%) <i>.077***</i> (<i>p</i> = .077)
JAMA	<i>n</i> = 55 <i>.102***</i> (<i>p</i> = .317)	<i>n</i> = 23 (41.8%) <i>.018***</i> (<i>p</i> = .913)	<i>n</i> = 7 (12.7%) <i>-.117***</i> (<i>p</i> = .734)
The Lancet	<i>n</i> = 80 .294* (<i>p</i> = .000)	<i>n</i> = 28 (35.0%) <i>.328**</i> (<i>p</i> = .025)	<i>n</i> = 14 (17.5%) <i>.024***</i> (<i>p</i> = .910)
NEJM	<i>n</i> = 157 .241* (<i>p</i> = .000)	<i>n</i> = 65 (41.4%) <i>.196**</i> (<i>p</i> = .029)	<i>n</i> = 20 (12.7%) <i>.354**</i> (<i>p</i> = .033)

^aThe articles in each set (*n*) have both F1000Prime scores and WoS citation counts

^bQuantile ($\geq$ median); Quartile (75th percentile); for F1000Prime articles with tied stars, recommenders used for the tiebreaker

*Bold numbers indicate that the Kendall's tau_b test correlation is significant at 0.01 level (2-tailed)

**Italicized numbers indicate that the Kendall's tau_b test correlation is significant at 0.05 level (2-tailed)

***Kendall's tau_b test correlation is not significant

overlapped, both measures ranked the article the same. For each article, the distance between the two values can be seen as a vertical line (up or down bar). Figures 6, 7 and 8 omitted these lines. When the orange data point is below the blue data dot, the distance is negative, suggesting a highly recommended article did not get the same rank by citation counts (vice versa), and the larger the distance, the bigger the difference. All the journals in Figs. 5, 6, 7 and 8 had both positive and negative zones, but NEJM had the most interpolating zones, and JAMA had only one large negative and one small positive zone. These plots showed that the highly scored F1000Prime articles were on the most left part of the figures with larger negative distances; the most cited articles tended to be on the right part of the figures with smaller positive zones where the articles' F1000Prime scores were lower. In addition, the spikes indicate superstar articles by citation counts with lower F1000Prime ratings. For Cell and JAMA (Table 7), the most cited 2010 article (total citations of 3951

Fig. 5 Comparisons of F1000Prime scores and WoS citations for journal Cell

Fig. 6 Comparisons of F1000Prime scores and WoS citations for JAMA

Fig. 7 Comparisons of F1000Prime scores and WoS citations for The Lancet

Fig. 8 Comparisons of F1000Prime scores and WoS citations for NEJM

Table 7 The most-cited article and its F1000Prime recommendations

Journal	WoS			F1000Prime			
	Cites	Peak year	Range	Rank/dataset	Score	FM	Most recent
Cell	3951	2017	43–699	158th (ties)/188	1	1	2011
JAMA	4035	2012	154–924	34th (ties)/55	1	1	2012
The Lancet	2489	2017	1–388	5th/80	6	2	2014
NEJM	6640	2016	60–1101	67th/157	3	2	2010

F1000Prime articles have more ties in which all the following articles in the group share the same rank by F1000Prime star rating and number of FMs. To break the ties, WoS citation count served as the third criterion. WoS has fewer ties; F1000Prime two measures served as the tie-breaker

Table 8 The most-scored F1000Prime article and its citation counts

Journal	F1000Prime		WoS			
	Score	FM	Citations	Rank	Peak year	Range
Cell	15	7	110	214th/288	2012	6–19
JAMA	8	4	720	6th/235	2018	1–138
The Lancet	16	7	1090	15th/272	2016	17–150
NEJM	23	13	898	31st/293	2016	4–154

and 4035 respectively) got only 1 star by 1 FM. All four journals depicted in Figs. 5, 6, 7 and 8 have larger negative zones than positive zones and fewer overlapping points; within the zones, the patterns vary.

The most-cited articles and the most-scored F1000Prime articles. For each journal in Figs. 5, 6, 7 and 8, the one visible spike was further examined to understand how it was recommended in F1000Prime. In Table 7, the most cited article published in Cell and JAMA had only 1 FM assigned 1 star respectively; the article's ranking in F1000Prime was among the lowest in the long tail with ties in the rank at the 158th for the 31 rank-tied articles in the set of Cell's 188 articles and at the 34th for the 22 rank-tied articles in the set of JAMA's 55 articles respectively. The most-cited NEJM article was ranked 67th out of the 157 recommended articles in F1000Prime. Ranked the 5th, the most-cited Lancet article was the only article that had the closest ranks by the two measures.

On the other hand, the top 1 F1000Prime article for each journal (Table 8) received varied citation ranks: the 6th out of the 235 JAMA articles, the 15th out of the 272 Lancet articles, the 31st out of the 293 NEJM articles, and the 214th out of the 288 Cell articles. The peak citation year was 2–8 years after an article's publication year. These F1000Prime top recommended articles were classified by an FM as New Finding, which may indicate the type of articles FMs tend to recommend.

Because NEJM had the most interpolating zones and the more spikes (Fig. 8), the further analysis compared the two lists of top 5 articles ranked by their respective measures (Table 9 and Table 11). The 2nd and the 4th most-cited articles were also the 3rd- and the 4th-ranked articles by F1000Prime total stars. The classifications assigned to the 8 articles were Clinical Trial (7), New Finding (7), Changes Clinical Practice (4), Confirmation (4), Controversial (3), Technical Advance (2), Interesting Hypothesis (1), and Refutation. The

only two Technical Advance articles were the top 1 and top 2 cited articles sharing the same 4 classifications, but their F1000Prime ranks differ, 67th versus 3rd. The three Controversial articles were ranked the 1st, the 2nd, and the 5th respectively by F1000Prime, but very low in the ranking by WoS. The F1000Prime 5th ranked article was ranked the 177th out of the 293 articles. This article's citation counts from 2010 to 2018 are as follows: 9, 29, 31, 26, 34, 27, 21, 18, 18. How is this article compared with all the NEJM articles in the dataset?

To answer this derived question, the dataset of 293 NEJM articles was further analyzed to identify each article's peak citation year. The results indicate that 86.35% of the articles were being cited in their publication year, but none of the articles had peak citations in the publication year. The peak citation year occurred as early as the year following the publication year (Fig. 9). Figure 9 shows the trend of the peak citation year for the 293 articles: more articles reached citation peak in 2 years. This phenomenon corroborates the previously reported result that citations in medicine tended to rise early (Wang 2013). Because medical articles tended to reach peak citations early, many later rising articles could be easily overlooked in citation analysis using the popular 5-year citation window. The articles in Fig. 9 reached their respective peak citation counts in 2017 or 2018, which is 7 or 8 years after publication, may be cases worth further observations. As one of the reviewers suggested, were these late citation peak articles "sleeping beauty"? We intend to follow these articles to shed light on their citation patterns.

The 4th ranked article (Tables 9, 10) had its peak citation count of 475 in 2018 (16.92% of 2808 cites). This article shows a steady rising in citations from 2010 to 2018: 24, 188, 246, 312, 317, 374, 422, 450, 475; the trend is likely to continue. This article's citation pattern is similar to those of the classics in economics (Wang and Bownas 2005).

Content analysis of top-cited and top-recommended articles (RQ5)

The content analysis focused on the set of 8 NEJM articles (Tables 9, 11). Each of the 8 articles was reviewed by at least 2 FMs (or together with their AFMs). A total of 52 recommendations were analyzed using QDA Miner. The comments in these recommendations

Table 9 NEJM Top articles ranked by top 5 citation counts and top-5 F1000Prime scores

WoS	Cites	Peak Year ^a	Range	F1000Prime	Stars	FMs	Classified as ^b			
1st	6.640	2016	60–1101	(67th)	3	2	CP	CT	NF	TA
2nd	3443	2016	9–494	3rd	19	8	CP	CT	NF	TA
3rd	2887	2017	21–419	(51st)	4	2	CP	CT	CF	
4th	2808	2018	24–475	4th	17	8	CT	NF	CF	
5th	2790	2014	10–450	(16th)	9	4	CT	NF		
(177th)	213	2014	9–34	5th	14	8	CP	NF	CF	CN IH
(16th)	1435	2012	47–217	2nd	20	9	CT	NF	CF	CN RF
(31st)	898	2016	4–154	1st	23	13	CT	NF	CN	

^aPeak year is the year the paper received the most citations

^bCP Changes Clinical Practice, CT Clinical Trial, NF New Finding, CF Confirmation, CN Controversial, TA Technical Advance, IH Interesting Hypothesis, RF Refutation

The bold ranks (2nd, 3rd, and 4th) represent the articles that were ranked top 5 by both F1000Prime total stars and WoS total citations

Fig. 9 NEJM articles peak citation year after publication 2010

ranged between 55 and 805 words or between 2 and 51 sentences. Although the majority of the comments rephrased or highlighted the paper’s findings or methods, there were both positive and critical evaluations. For example, 59.9% of the comments mentioned the studies being “important,” “long-awaited results,” “noteworthy,” “excellent,” “exceptional,” “remarkable,” “interesting,” etc. About 27.0% of the comments claimed the paper was the *first* to report a study, adopt a method, or clarify a definition. Seven recommendations mentioned the impact of the papers as a potential “paradigm shift.” A substantial number of recommendations included suggestions for future research (41.8%). Although there were criticisms such as “lack major findings,” they constituted only about 3.8%.

Some cases showed the discrepancy between sentiment and star-rating. For example, the 4th-ranked paper was referred to as a “landmark study” by 3 FMs, who, however, assigned 3 different ratings: 1 star, 2 stars, and 3 stars, respectively. For the 7 recommendations that included a “paradigm shift” in comments, only 1 FM gave 3 stars while the other 6 FMs all gave 2 stars.

Discussion and conclusions

This study examined research and review papers published in four journals in biology and medicine in 2010 to answer five research questions. The range of data is the main limitation of this study. Thus, the findings are meant to provide evidence rather than conclusions. Although some of the findings corroborate with previous studies, differences are also found in the results that suggest further studies. In the analysis, we did not try alternative methods, such as to include all articles from WoS and assign zero scores to articles without F1000Prime recommendations, as in the study by Li et al. (2011). This could be a future study.

Table 10 Selected NEJM articles and their ranking by F1000Prime and WoS

F1000Prime Recommended Articles					WoS Cited Articles		
Rank	Scores	FMs	WoS Citations	DOI	Rank	Citations	DOI
1	23	13	898	10.1056/NEJMoa1005372	<i>1</i>	<i>6640</i>	<i>10.1056/NEJMoa1003466</i>
2	20	9	1435	10.1056/NEJMoa0912321	2	3443	10.1056/NEJMoa1008232
3	19	8	3443	10.1056/NEJMoa1008232	<i>3</i>	<i>2887</i>	<i>10.1056/NEJMoa0909530</i>
4	17	8	2808	10.1056/NEJMoa1000678	4	2808	10.1056/NEJMoa1000678
<i>5</i>	<i>14</i>	<i>8</i>	<i>213</i>	<i>10.1056/NEJMoa0903326</i>	5	2790	10.1056/NEJMoa1006448
6	14	6	1750	10.1056/NEJMoa1001286	6	2710	10.1056/NEJMoa1001294
7	13	6	2164	10.1056/NEJMoa1011205	7	2542	10.1056/NEJMoa0912217
8	13	6	475	10.1056/NEJMoa0908094	9	2272	10.1056/NEJMoa1002011
9	12	6	2710	10.1056/NEJMoa1001294	10	2164	10.1056/NEJMoa1011205
10	12	6	620	10.1056/NEJMoa0907118	12	1750	10.1056/NEJMoa1001286
11	11	5	580	10.1056/NEJMoa0810988	15	1449	10.1056/NEJMoa0904492
12	10	4	2272	10.1056/NEJMoa1002011	16	1435	10.1056/NEJMoa0912321
13	9	6	623	10.1056/NEJMoa0908127	<u><i>17</i></u>	<u><i>1431</i></u>	<u><i>10.1056/NEJMoa1007903</i></u>
14	9	5	157	10.1056/NEJMoa1001283	<u><i>18</i></u>	<u><i>1431</i></u>	<u><i>10.1056/NEJMoa1001282</i></u>
15	9	5	64	10.1056/NEJMoa0900158	19	1351	10.1056/NEJMoa0909494
16	9	4	2790	10.1056/NEJMoa1006448	20	1270	10.1056/NEJMoa0909883
17	9	4	1449	10.1056/NEJMoa0904492	22	1195	10.1056/NEJMoa0907847
18	9	4	409	10.1056/NEJMoa1000552	23	1189	10.1056/NEJMoa0907839
19	9	3	581	10.1056/NEJMoa0905370	24	1172	10.1056/NEJMoa0908721
20	8	4	520	10.1056/NEJMoa1001593	25	1159	10.1056/NEJMoa0907929
21	7	4	1021	10.1056/NEJMoa0909905	31	898	10.1056/NEJMoa1005372
26	7	3	960	10.1056/NEJMoa0912614	34	872	10.1056/NEJMra0912063
29	6	4	637	10.1056/NEJMoa0907355	39	773	10.1056/NEJMoa1008433
31	6	3	832	10.1056/NEJMoa0907667	48	692	10.1056/NEJMoa1002965
37	5	3	1172	10.1056/NEJMoa0908721	60	612	10.1056/NEJMoa0904130
47	5	2	407	10.1056/NEJMoa0912461	64	580	10.1056/NEJMoa0810988
50	4	3	255	10.1056/NEJMoa1000500	65	573	10.1056/NEJMoa1001288
<i>51</i>	<i>4</i>	<i>2</i>	<i>2887</i>	<i>10.1056/NEJMoa0909530</i>	86	491	10.1056/NEJMSa0911535
66	3	3	294	10.1056/NEJMoa1000727	88	487	10.1056/NEJMoa0909809
<i>67</i>	<i>3</i>	<i>2</i>	<i>6640</i>	<i>10.1056/NEJMoa1003466</i>	90	475	10.1056/NEJMoa0908094
<u><i>68</i></u>	<u><i>3</i></u>	<u><i>2</i></u>	<u><i>1431</i></u>	<u><i>10.1056/NEJMoa1007903</i></u>	91	471	10.1056/NEJMoa0911783
69	3	2	1195	10.1056/NEJMoa0907847	92	467	10.1056/NEJMoa0907635
70	3	2	1159	10.1056/NEJMoa0907929	100	429	10.1056/NEJMoa1006885
<u><i>128</i></u>	<u><i>1</i></u>	<u><i>1</i></u>	<u><i>1431</i></u>	<u><i>10.1056/NEJMoa1001282</i></u>	<i>177</i>	<i>213</i>	<i>10.1056/NEJMoa0903326</i>

The bold or/and italicized articles are being ranked in one system as top 5 but relatively low by another system. For example, the 5th ranked F1000Prime article is the 177th ranked WoS article; the unique identifier is the DOI. To rank articles in each system, methods were developed to break the ties. For F1000Prime articles, WoS citations served as tie-breaker and vice versa (see the two italicized, underlined records). Some ties remained when the articles were not recommended in F1000Prime

RQ1: Four of the top 5 classifications in this study overlap with the top classifications reported by Bornmann and Haunschild (2015) except for Clinical Trial, which was the 2nd in this study, and Good for Teaching, which was the 4th in their study. Excluding these two classifications in comparison, the remaining four classifications would be

in the same order for both studies: New Finding, Confirmation, Interesting Hypothesis, and Technical Advance (Fig. 5). For individual journals, the differences are likely due to the nature of the journals. Cell, the biological journal, had New Finding, Interesting Hypothesis, and Technical Advance as the top 3 classifications (Table 4) but no article in Clinical Trial or Good for Teaching. All medical journals shared the same top 3 classifications: Clinical Trial, New Finding, and Confirmation. Further studies should examine data beyond 1 year and include additional journals.

RQ2: For JAMA, The Lancet, and NEJM, the articles in the set of F1000Prime received significantly more citations than other articles of the same journal for most of the 9 citing years including the publication year (Table 5). For Cell, however, no statistical significance was found between the 2 datasets of articles in any specific year or all years considered. In relation to the *RQ1*, Cell again differs from the other three journals and requires further studies to include more biological journals.

RQ3: F1000Prime rating scores and WoS citation counts were significantly correlated for Cell, The Lancet, and NEJM for the entire dataset and the upper quantile (> median). NEJM is the only journal that showed a significant correlation for the upper quartile dataset (≥ 75 th percentile or top 25%). JAMA, however, is the only journal that showed no significant correlation between the two measures (Table 6). One possible explanation is that JAMA had the smallest dataset 1 (Only 23.4% published articles were recommended in F1000Prime). Our results append the findings of a weak correlation reported by Waltman and Costas (2014). Within a larger dataset of multiple journals, some journals may show a strong correlation, while others show no correlation; if analyzed together, they may show weak correlations.

RQ4: The scaled F1000Prime scores and WoS citation counts allowed graphic comparisons of the two measures (Figs. 5, 6, 7, 8). These graphs suggest that the two measures did not share similar patterns when one measure was higher than the median; the other measure tended to be lower than the median. There are several zones except for JAMA to show how the two measures compared across all the articles in dataset 1. Even within the zone, it is not obvious how F1000Prime scores can predict citation counts despite significant correlations for three journals as reported in RQ3. This finding is in agreement with Wardle (2010) that F1000Prime scores could not predict citation trends.

RQ5: Although only as a snapshot analysis of the recommendations to the top articles, a few interesting observations can shed some light upon comments and ratings in F1000Prime. Besides summarizing or highlighting major findings or methods, the most commonly used words to introduce the articles implied positive sentiment, including “interesting,” “important,” “first,” “exceptional,” etc. Recommending FMs assigned different rating scores to the articles about which their written comments were in a similar sentiment. This finding would corroborate the low level of rating agreement between FMs reported by Bornmann (2015). FMs tended to assign the average rating to articles even though they used strong words such as “landmark study” or “paradigm shift” in comments. This finding may explain why the three-star ratings accounted for only 10% of the 28,138 recommendations in the study by Bornmann and Haunschild (2015), and 6.9% of the 132,622 recommendations in the study by Waltman and Costas (2014). Further studies should observe what criteria FMs use to assign ratings such as significance or quality and whether there exist disciplinary differences in criteria used by FMs.

Taking a closer look at the design of the rating system in F1000Prime, it is obvious that the three labels do not result in a true 3-point Likert scale, because they do not provide equal semantic distance and are open for interpretation by the raters. Raters may rate

Table 11 Top 5 WoS Cited and Top 5 F1000Prime Scored 2010 NEJM Articles

WoS (N = 293)	F1000Prime (N = 157)	Articles in the New England Journal of Medicine as published in 2010
1st	67th	Improved survival with ipilimumab in patients with metastatic melanoma, Aug 19; 363(8:711–723). https://doi.org/10.1056/nejmoa1003466
2nd	3rd	Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery, Oct 21; 363(17:1597–1607). https://doi.org/10.1056/nejmoa1008232
3rd	51th	Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR, Jun 24; 362(25:2380–2388). https://doi.org/10.1056/nejmoa0909530
4th	4th	Early palliative care for patients with metastatic non-small-cell lung cancer, Aug 19; 363(8:733–742). https://doi.org/10.1056/nejmoa1000678
5th	16th	Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer, Oct 28; 363(18:1693–1703). https://doi.org/10.1056/nejmoa1006448
177th	5th	Morphine use after combat injury in Iraq and post-traumatic stress disorder, Jan 14; 362(2:110–117). https://doi.org/10.1056/nejmoa0903326
13th	2nd	Stenting versus endarterectomy for treatment of carotid-artery stenosis, Jul 01; 363(1:11–23). https://doi.org/10.1056/nejmoa0912321
26th	1st	Neuromuscular blockers in early acute respiratory distress syndrome, Sep 16; 363(12:1107–1116). https://doi.org/10.1056/nejmoa1005372

differently in different classifications. The study by Mohammadi and Thelwall (2013) reported that New Finding was one of the classifications in which FFa rating scores and citation counts differed significantly. A possible explanation is that FMs are cautious about assigning the highest rating to the articles classified as New Finding, given that other FMs may post a dissenting comment (although no rating for Dissent) or the article may be retracted later (3 of the retracted articles were removed from our data analysis). As a post-publication expert recommendation system, F1000Prime's three-star rating may need to be further developed into a well-defined scale for metric use. It is also true that the F1000Prime created the rating scale with a quite different purpose, to recommend readings. As noted earlier, the current rating labels, *Good*, *Very good*, and *Exceptional*, were revised from the earlier rating labels: *Recommended*, *Must read*, and *Exceptional*, which indicates the difficulty in finding the right labels for the recommendation scale. Future studies should observe how F1000 FMs assign ratings to the articles they recommend and how F1000Prime users perceive the meanings of these ratings.

This study collected data from 4 journals in a single year (2010) to minimize data bias due to sporadic coverage of the years before F1000's merging of the two databases in 2009. The two databases, F1000 Biology and F1000 Medicine, were launched in 2002 and 2006 respectively. A previous study (Du et al. 2015) collected F1000Prime recommended articles published between 1999 and 2013. Although the researchers achieved a large dataset, there is a potential coverage bias for biology between 1999 and 2001 and for medicine between 1999 and 2005 because FMs likely read and recommend from current journals, and F1000Prime does not require FMs to systematically review journals. Adopting the longer citation window strategy by Bornmann and Leydesdorff (2015), this study collected citation data for as many years as possible from the publication year (2010). The peak citation year after publication varies for each article ranging from the following year to 8 years (the data collection ending year). For NEJM, none of the 293 articles had the most citations in the publication year (2010), although the majority of the articles (86.35%) were cited. The top cited articles show steady citations over the 9 years. The mode of the peak citation year is 2 years, but a small number of articles had much later peak citation years (Fig. 9). The articles receiving steady citations over a long period may be citation classics, and the articles showing late rising citations may be "sleeping beauties" with a longer hibernation period.

To conclude with caution, the percentage of research and review articles recommended by F1000 Faculty varies by journals (JAMA 23.4%, The Lancet 29.4%, NEJM 53%, and Cell 65.3%). The recommended articles were mostly classified as New Finding; the top classifications assigned by F1000 Faculty members for the three medical journals differed from the biological journal. Although there were correlations between F1000Prime scores and citations counts, scaled graphic comparisons of the two measures did not suggest whether F1000Prime scores could predict citation trends. Some F1000Prime top recommended articles received very low ranks by citation counts. Sleeping beauties may be the reason for the few articles with the longer citation peak windows (7 and 8 years in this study). F1000 Faculty's sentiments on articles did not always correspond to the rating scores they assigned. Further studies should include additional journals, especially biological journals and different publication years, to compare and extend the findings reported here. Studies of F1000Prime users should be another line of valuable research to understand F1000Prime users' perspectives on the rating scale, classifications, and expert comments.

Acknowledgements Open access funding provided by the School of Information Sciences, The University of Tennessee, Knoxville. Wang and Williams wish to acknowledge the 2019 Dean's Summer Research grant, the College of Communication and Information, The University of Tennessee, Knoxville. Wu wish to acknowledge the support of the National Natural Science Foundation of China (Grants 71874173 and 71273250) and the Academic Division of Mathematics and Physics of the Chinese Academy of Sciences (Grant 2018-M04-B-004). Special thanks to Adam Hembree for editorial help. We sincerely thank the two reviewers for their constructive comments that guided us to improve the manuscript by reanalyzing the data, adding additional quantitative and qualitative analysis, and making substantial revisions.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS ONE*, 4(6), e5910. <https://doi.org/10.1371/journal.pone.0005910>.
- Benos, D. J., et al. (2007). The ups and downs of peer review. *American Journal of Physiology—Advances in Physiology Education*, 31(2), 145–152.
- Bornmann, L. (2015). Interrater reliability and convergent validity of F1000Prime peer review. *Journal of the Association for Information Science and Technology*, 66(12), 2415–2426. <https://doi.org/10.1002/asi.23334>.
- Bornmann, L., & Haunschild, R. (2015). Which people use which scientific papers? An evaluation of data from F1000 and Mendeley. *Journal of Informetrics*, 9(3), 477–487. <https://doi.org/10.1016/j.joi.2015.04.001>.
- Bornmann, L., & Haunschild, R. (2016). 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. <https://doi.org/10.1002/asi.23729>.
- Bornmann, L., & Haunschild, R. (2018). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PLoS ONE*, 13(5), e0197133. <https://doi.org/10.1371/journal.pone.0197133>.
- Bornmann, L., & Leydesdorff, L. (2013). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. *Journal of Informetrics*, 7(2), 286–291. <https://doi.org/10.1016/j.joi.2012.12.0034>.
- Bornmann, L., & Leydesdorff, L. (2015). Does quality and content matter for citedness? A comparison with para-textual factors and over time. *Journal of Informetrics*, 9(3), 419–429. <https://doi.org/10.1016/j.joi.2015.03.001>.
- Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgments of experts? *Journal of Informetrics*, 9(29), 408–418.
- Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 6(11), 2215–2222.
- Bornmann, M., Leydesdorff, L., & Wang, J. (2013). Which percentile-based approach should be preferred for calculating normalized citation impact values? An empirical comparison of five approaches including a newly developed citation-rank approach (P100). *Journal of Informetrics*, 7(4), 933–944. <https://doi.org/10.1016/j.joi.2013.09.003>.
- de Solla Price, D. J. (1961). *Science Since Babylon*. New Haven, CT: Yale University Press.
- Doane, D. P., & Seward, L. E. (2011). Measuring skewness: A forgotten statistics? *Journal of Statistics Education*, 19(2). <http://www.amstat.org/publications/jse/v19n2/doane.pdf>. Accessed 11 Nov 2019.
- Dorch, S. B. F. (2012). On the citation advantage of linking to data: Astrophysics. *Journal of eScience Librarianship*. Retrieved from <https://www.researchgate.net/publication/229070295>. Accessed 11 Nov 2019.
- Du, J., Tang, X., & Wu, Y. (2015). The effects of research level and article type on the differences between citation metrics and F1000 recommendations. *Journal of the Association for Information Science and Technology*, 67(12), 3008–3021. <https://doi.org/10.1002/asi.23548>.

- Haunschild, R., & Bornmann, L. (2018). Field- and time-normalization of data with many zeros: An empirical analysis using citation and Twitter data. *Scientometrics*, 116(2), 997–1012.
- Hirschauer, S. (2015). How editors decide; Oral communication in journal peer review. *Human Studies*. <https://doi.org/10.1007/s10746-014-9329-x>.
- Li, X., Thelwall, M., & Giustini, D. (2011). Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2), 461–471.
- Mohammadi, E., & Thelwall, M. (2013). Assessing non-standard article impact using F1000 labels. *Scientometrics*, 97(2), 383. <https://doi.org/10.1007/s11192-013-0993-9>.
- Pontis, S., Blandford, A., Greifeneder, E., Attalla, H., & Neal, D. (2015). Keeping up to date: An academic researcher's information journey. *Journal of the Association for Information Science and Technology*, 68(1), 22–35.
- Small, H. (2004). Why authors think their papers are highly cited. *Scientometrics*, 60(3), 305–316.
- Tressoldi, P. E., Giofré, D., Sella, F., & Cumming, G. (2013). High impact=high statistical standards? Not necessarily so. *PLoS ONE*, 8(2), e56180.
- Vardell, E., & Swogger, S. E. (2014). F1000Prime: A faculty of 1000 tool. *Medical Reference Services Quarterly*, 33(1), 75–84.
- Waltman, L., & Costas, R. (2014). F1000 recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science & Technology*, 65(3), 433–445.
- Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, 94(3), 851–872.
- Wang, P., & Bownas, J. (2005). Research classics and citation analysis. In *Proceedings of the 2005 ASIST annual meeting* (Charlotte, NC, October 28–November 2) <https://asistdl.onlinelibrary.wiley.com>; <https://doi.org/10.1002/meet.14504201202>
- Wang, P., & White, M. D. (1999). A cognitive model of document use during a research project. Study II: Decisions at reading and citing stages. *Journal of American Society for Information Science*, 50(2), 98–114.
- Wardle, D. A. (2010). Do 'Faculty of 1000' (F1000) ratings of ecological publications serve as reasonable predictors of their future impact? *Ideas in Ecology and Evolution*, 3, 11–15.
- Wets, K., Weedon, D., & Velterop, J. (2003). Post-publication filtering and evaluation: Faculty of 1000. *Learned Publishing*, 16(4), 249–258. <https://doi.org/10.1087/095315103322421982>.

Affiliations

Peiling Wang¹ · Joshua Williams¹ · Nan Zhang² · Qiang Wu²

✉ Peiling Wang
peilingw@utk.edu

✉ Qiang Wu
qiangwu@ustc.edu.cn

¹ School of Information Sciences, The University of Tennessee, 1345 Circle Park Dr., Knoxville, TN 37932-0341, USA

² School of Management, University of Science and Technology of China, 96 Jin Zhai Rd., Hefei 230026, China