Is Peer-Reviewing Worth the Effort?

Kenneth Church, Raman Chandrasekar, John E. Ortega and Ibrahim Said Ahmad
Institute for Experiential AI, Northeastern University
{k.church, r.chandrasekar, j.ortega, i.ahmad}@northeastern.edu

Abstract

How effective is peer-reviewing in identifying important papers? We treat this question as a forecasting task. Can we predict which papers will be highly cited in the future based on venue and "early returns" (citations soon after publication)? We show early returns are more predictive than venue. Finally, we end with constructive suggestions to address scaling challenges: (a) too many submissions and (b) too few qualified reviewers.

1 Introduction

1.1 Prioritization as a Forecasting Task

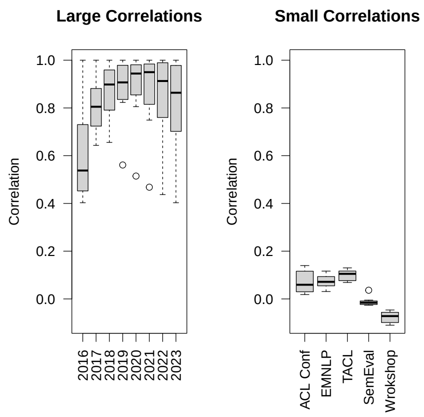


Figure 1: Early Returns (left) $\gg$ Venue (right), based on correlations (ρ) from Tables 2-3. Data is based on Semantic Scholar (S2) (Wade, 2022), where the venue field refers not only to conferences, but also to journals and more.

How effective is peer-reviewing in identifying important papers? Since readers cannot afford to read everything, should they prioritize papers in top venues, or something else? Following Davletov et al. (2014); Ma et al. (2021), we treat this question as a forecasting task. Can we predict which papers will be highly cited in the future? Both venue and "early returns" (citations soon after publication) are statistically significant, but early citations have larger correlations with future citations as shown in Figure 1. This figure will be explained in more detail in section 3. Data for figures and tables is posted on GitHub.[1]

Abramo et al. (2019) also found early citations to be more predictive than venue (impact): "the role of the impact factor in the combination becomes negligible after only two years from publication."

1.2 H-Index and Impact

In some organizations, authors are encouraged to publish in top tier venues, using statistics such as h-index (Hirsch, 2005) and impact (Garfield, 2006) to rank authors, venues, countries (Hyland, 2023) and more. We use similar summary statistics to show that conditioning on early citations is more effective than conditioning on venue. That is, we group papers by venue and by early citations (one year after publication), and summarize citations for the fourth year after publication with h (h-index) and μ (impact). Results do not depend too much on the details of these definitions of early and future citations because citations are highly correlated over time

When we discuss Tables 4-5, h and μ are better for papers conditioned on early citations than for papers conditioned on venue. In particular, papers in less selective venues (Workshops/ArXiv) with a few early citations tend to have more citations in the future than papers in more selective venues.

In addition to h and μ, Tables 4-5 report N (number of papers in each group) and σ (standard deviation). N will be used in discussions of inclusiveness and σ will be used in discussions of robustness. We will suggest prioritizing papers with early citations is more effective than prioritizing by venue:

[1] https://github.com/kwchurch/is-peer-reviewing-worth-the-effort

Proceedings of the 31st International Conference on Computational Linguistics, pages 3589–3599
January 19–24, 2025. ©2025 Association for Computational Linguistics

1. More selective: ρ (correlation), h, μ (impact)
2. More inclusive: N (number of papers)

These observations are robust, as will be shown by replications over a number of conditions including papers from different sources (ACL, PubMed, ArXiv) and papers with different publication dates.

2 Related Work

2.1 Metrics: H-index and Impact Factor

There is considerable work on metrics of success such as impact factor (Garfield, 2006) and h-index (Hirsch, 2005). Both of these summary statistics are computed over a group of papers, where papers are typically grouped by author or by venue, depending on whether one is interested in measuring success by author or by venue. We will group papers in additional ways such as papers with T or more citations in the first year after publication in order to compare scores of success by early returns with scores by other factors such as venue.

Impact factor, μ, is simply the average of citation counts for papers in the group, and h-index, h, is the number of papers in the group with h or more citations. Many journals report impact factors. Google Scholar ranks venues by h5, a variant of h-index, computed over the last five years. In addition to top venues,[2] Google also provides details for many fields such as Computational Linguistics.[3]

2.2 Numerous Challenges to Reviewing

The peer-review process, despite being an integral part of academic scholarship, has been a subject of criticism on multiple fronts (Jefferson et al., 2002):

> *the practice of peer review is based on faith in its effects, rather than on facts.*

In this work, we assume reviews and other assessments of value should be leading indicators of future citations, following suggestions we have made elsewhere (Church, 2005, 2020). While this assumption may be controversial, it provides an objective path forward. There are, of course, numerous challenges in reviewing processes; the first three challenges below are discussed in Sections 2.2.1-2.2.3; scale/growth is discussed in 2.3.2.

1. Poorly defined tasks/incentives
2. Validity and Reliability
3. Vulnerabilities, Cheating and Ethics
4. Scale: Exponential growth
5. Subjectivity/Biases (Lee et al., 2013; Huber et al., 2022; Smith et al., 2023)
6. Time and Cost (De Vries et al., 2009)

2.2.1 Purpose of peer-reviewing?

What is the purpose of peer-reviewing? The task is not very well-defined. According to Rogers and Augenstein (2020), “reviewers and area chairs face a poorly defined task forcing apples-to-oranges comparisons.” An evaluation of biomedical research publications (Chauvin et al., 2015) concluded: “The most important tasks for peer reviewers were not congruent with the tasks most often requested by journal editors in their guidelines to reviewers.”

2.2.2 Validity and Reliability

There is considerable discussion of validity and reliability in Experimental Psychology (Krippendorff, 2018). Evaluations of the reliability of peer-reviewing are worrisome. Cortes and Lawrence (2021) revisited an experiment based on NIPS-2014 (now known as NeurIPS): “From the conference 10% of the papers were randomly chosen to be reviewed by two independent program committees... results showed that the decisions between the two committees was better than random, but still surprised the community by how low it was.”

The follow up study looked at review scores and future citations. They failed to find a significant correlation for accepted papers (their figure 6). For rejected papers that appeared elsewhere, the correlation was not large (their figure 8).

A recent evaluation of reviews (Goldberg et al., 2023) found “many problems that exist in peer reviews of papers—inconsistencies, biases, miscalibration, subjectivity—also exist in peer reviews of peer reviews.”

2.2.3 Vulnerabilities, Cheating and Ethics

There are opportunities for authors, reviewers and other parties to use/abuse chatbots. A number of funding agencies (NIH[4] and ARC[5]) and journals

[2] https://scholar.google.com/citations?view_op=top_venues

[3] https://scholar.google.com/citations?view_op=top_venues&vq=eng_computationallinguistics

[4] https://nexus.od.nih.gov/all/2023/06/23/using-ai-in-peer-review-is-a-breach-of-confidentiality/

[5] https://www.arc.gov.au/sites/default/files/2023-07/Policy%20on%20Use%20of%20Generative%20Artificial%20Intelligence%20in%20the%20ARCs%20grants%20programs%202023.pdf

(Science, Lancet, JAMA) discourage/prohibit reviewers from uploading manuscripts to AI platforms that cannot guarantee confidentiality (Cheng et al., 2024).

Even before chatbots, much has been written about ethics and peer-reviewing: (Rockwell, 2006; Souder, 2011; Remuzzi, 2023). There have always been many ways to cheat. Advances in technology create new and better ways to cheat, as well as new and better ways to catch cheating.

In this work, we will use citations, which admittedly can be purchased/gamed[6] (Beel and Gipp, 2010). Spam is obviously a cat-and-mouse game but purchasing citations is unlikely to be successful for long. Given the correlations over time, cheaters would need to purchase citations for many years or else it is too easy to catch them by looking for anomalies in citation counts over time. Moreover, with h-index, it is too easy to find the small number of papers that contribute to the score. There are easier and more effective ways to cheat such as plagiarism and chat bots.

2.3 Related Work on Predicting Citations

This paper questions whether peer-reviewing is worth the effort. Prior work is more about improving predictions (subsubsection 2.3.1), or helping authors increase their citations (subsubsection 2.3.2).

2.3.1 Improving Predictions

There is a considerable body of work on predicting citations. Predicting citations can be viewed as a special case of time series prediction. There are many use cases, especially in finance: (Salinas et al., 2020). Prior work often focuses on methods: linear regression (Pobiedina and Ichise, 2016), negative binomials[7] (Onodera and Yoshikane, 2015), clustering (Davletov et al., 2014), nearest neighbors (Yan et al., 2011) and deep networks (Abrishami and Aliakbary, 2019; Ruan et al., 2020). There is considerable work on link prediction in the literature on GNNs (graph neural networks) using the ogbl-citation2 task in OGB (Open Graph Benchmark) (Hu et al., 2020). In more recent work, de Winter (2024) aims to "pave the way for AI-assisted peer review," using ChatGPT4 to analyze 2222 abstracts with 60 criteria. Using principal component analysis, three components are identified, of which two – about Accessibility & Understandability, and Novelty & Engagement, are linked to citation counts.

In addition to methods for predicting citations, there are also discussions of features:

1. Early Citations: Wang et al. (2013); Davletov et al. (2014); Abramo et al. (2019); Bai et al. (2019); Stegehuis et al. (2015); Ma et al. (2021); Yan et al. (2024)
2. Venue: Yan et al. (2011); Abramo et al. (2019)
3. Properties of authors: author rank, h-index, productivity, etc. Yan et al. (2011)
4. Contents of paper: Huang et al. (2022) predict citations based on sections (introduction, background, method, etc.) of a paper.

2.3.2 Advice to Authors

There is considerable advice to authors on how to increase citations. We have argued elsewhere (Church, 2017) that secondary sources are cited more than primary sources; the most cited papers often help others make progress, e.g., datasets, GitHubs, models on HuggingFace, benchmarks, tools, surveys, textbooks. By construction, the last word on a topic is not cited. The most cited paper is rarely the first, last or best; simplicity and accessibility are preferred over timing and quality.

Tahamtan et al. (2016) survey the literature on advice to authors, assigning prior work to 28 factors, which we have aggregated/condensed down to 8. Their 28 factors seem plausible, though it is not possible to discuss all 28 factors in this paper.

1. Intrinsic properties of paper: quality, length, number of references. Figures, charts and appendices can increase citations, but challenging equations can decrease citations.
2. Venue: metrics (μ, h), prestige, language.
3. Discipline/subject/topic/methodology
4. Accessibility and visibility of papers: Avoid pay walls (Lawrence, 2001; Eysenbach, 2006), and promote papers on social media/ArXiv.
5. Primary Source vs. Secondary Source: Textbooks and survey papers are highly cited, as are tools, benchmarks and datasets.
6. Demographics of author(s): Number of authors, self-citations, country, gender, age, reputation, productivity, affiliation, funding.

[6] https://www.science.org/content/article/vendor-offering-citations-purchase-latest-bad-actor-scholarly-publishing

[7] Negative binomials are a natural choice for highly skewed data. Citations tend to be highly skewed as can be seen from standard deviations (σ) in many of the tables in this paper. If citations were generated by a Poisson process, then $\sigma^2 \approx \mu$, but citations have long tails where $\sigma \gg \mu$ (in most cases).

7. Publication date: Since the literature is growing exponentially, doubling every 17 years (Bornmann et al., 2021; Redner, 2005), papers published recently tend to have more citations.
8. Early citations: Citations soon after publication are predictive of future citations, though there are exceptions such as "Sleeping Beauties" (van Raan, 2004).

Venue	Id in S2	2016	2017	2018	2019	2020	2021
NAACL	9724599	5	7	5	1	3	1
LREC	12260053	0	0	0	1	0	0
LREC	28309452	2	8	4	10	7	7
EMNLP	1380793	0	2	16	19	17	19
COLING	18649702	0	1	2	1	3	1
SemEval	17378758	0	0	0	2	0	0

Table 1: Citation counts from Semantic Scholar (S2) for a few ACL papers published in 2016.

3 Predictions Based on Citations

As suggested above, we will use a prediction task to show that early returns are more effective than venue. Figure 1 is based on citation counts from Semantic Scholar (S2) (Wade, 2022). For papers in ACL Anthology, PubMed and ArXiv, published between 2016 and 2019, we extracted citations by year, as illustrated in Table 1. There are slightly more than a million papers per year in PubMed, 100k/year in ArXiv and 3k/year in ACL. The next 3 subsections use these citations to:

1. Compute correlations (ρ) over time and venue
2. Compute h-index (h) and impact (μ) for papers grouped by early citations and venue
3. Forecast citations with regression

All 3 subsections demonstrate that early citations are more predictive of future citations than venue.

3.1 Correlations

The top of Table 2 focuses on 3710 ACL papers published in 2016. The correlation (ρ) of 0.80 between 2016 and 2017 compares the citation counts for these 3710 papers in 2016 and 2017. The bottom of Table 2 is similar except for the source of papers is now 1,026,798 PubMed papers. Both the top and bottom of Table 2 start with papers published in 2016. The correlation of 0.80 above between 2016 and 2017 drops slightly to 0.77 for PubMed papers.

Table 3 is like Table 2, but for venues. Venue is a binary indicator variable containing 1 if the paper appears in that venue and 0 otherwise. Figure 1 is based on correlations for ACL papers published in 2016. Figure 1 (left) is based on Table 2 (top), and Figure 1 (right) is based on Table 3 (top).

	2016	2017	2018	2019	2020	2021
3710 ACL Papers Pub. in 2016						
2016	1.00	0.80	0.66	0.56	0.51	0.47
2017	0.80	1.00	0.92	0.85	0.81	0.75
2018	0.66	0.92	1.00	0.98	0.94	0.88
2019	0.56	0.85	0.98	1.00	0.98	0.93
2020	0.51	0.81	0.94	0.98	1.00	0.98
2021	0.47	0.75	0.88	0.93	0.98	1.00
2022	0.44	0.70	0.82	0.88	0.95	0.99
2023	0.40	0.64	0.76	0.82	0.90	0.97
1,026,798 PubMed Papers Pub. in 2016						
2016	1.00	0.77	0.64	0.55	0.50	0.45
2017	0.77	1.00	0.90	0.82	0.75	0.68
2018	0.64	0.90	1.00	0.94	0.89	0.83
2019	0.55	0.82	0.94	1.00	0.94	0.90
2020	0.50	0.75	0.89	0.94	1.00	0.95
2021	0.45	0.68	0.83	0.90	0.95	1.00
2022	0.40	0.61	0.76	0.84	0.91	0.96
2023	0.35	0.54	0.69	0.78	0.86	0.93

Table 2: Citation counts (from Semantic Scholar) are highly correlated from one year to the next.

In addition to the main point, there are a number of interesting (though smaller) effects:

1. **Main point**: Correlations for early returns are much larger than correlations for venue.
2. **Prestige**: Top venues (the ACL main conference, EMNLP and TACL) have larger correlations with future citations than workshops.
3. **Forecasting horizon**: Because short-term forecasting is easier than long-term forecasting, correlations closer to the main diagonal of Table 2 are relatively large.
4. **Quantization**: Correlations for 2016 are relatively small because dates are quantized to years. There are two dates: year of publication and year of citation. Citation counts for the year of publication are relatively small because that is a partial year.
5. **Latency**: It takes time for papers to accumulate citations, and therefore, correlations improve for several years after publication.

These observations are robust. Tables 2-3 replicate the correlations for two types of sources of papers. The tables below replicate similar observations over two publication dates, using h and μ instead of ρ.

3.2 H-index and Impact

How can we identify papers that will be highly cited in the future? The previous section used cor-

3710 Papers in ACL Anthology (2016)				
	2016	**2017**	**2018**	**2019**
ACL Conf	0.140	0.136	0.096	0.068
EMNLP	0.031	0.116	0.103	0.084
TACL	0.069	0.111	0.130	0.120
SemEval	0.036	-0.005	-0.026	-0.024
Workshops	-0.110	-0.104	-0.094	-0.077
1,121,081 Papers in PubMed, ArXiv or ACL (2016)				
ACL	0.0086	0.0109	0.016	0.015
ArXiv	0.0255	0.0088	0.024	0.021
PubMed	-0.0212	-0.0012	-0.014	-0.013

Table 3: Correlations with venue are smaller.

relations (ρ). This section will use h-index (h) and impact factors (μ). Tables 4-5 group papers based on citations a year after publication, and report summary statistics of citations in the fourth year after publication. Table 4 does this The main observation is: conditioning on papers with early citations compares favorably to conditioning by venue.

1. **Exclusivity**: Papers with 20+ citations in the first year after publication are better than all 50 venues in Table 4 in terms of h and μ.
2. **Inclusivity**: There are more papers (N) with 20+ early citations than in most venues.
3. **Robustness**: We obtain similar results under a number of conditions including different publication years and different sources of papers.

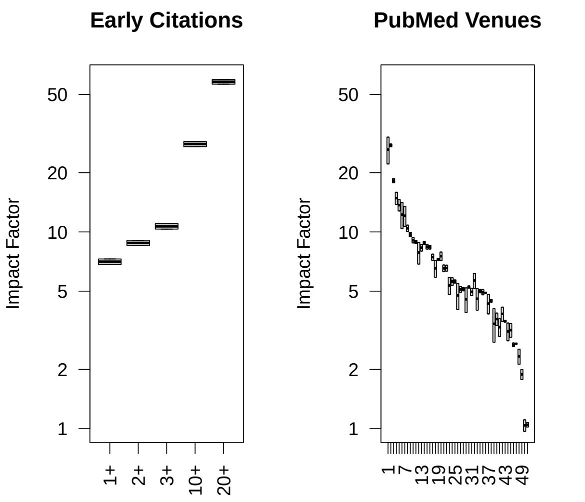


Figure 2: Impact factor (μ) from Table 4. Simple rule of thumb: for most venues, reviewers are no better than 1+ early citations in terms of μ; for all venues, reviewers are no better than 20+ early citations.

Figure 2 plots impact factors (μ) from Table 4, comparing early citations (left) with 50 PubMed venues (right). The figure shows that papers with 20+ early citations have a larger μ than all 50 venues. If we select on 1+ early citations, then $\mu \approx 6.8$ is better than 60% of venues in Table 4.

Note that h does not change much with thresholds on early citations. That is, h for 1+ citations is similar to h for 2+ citations because h is dominated by a few highly cited papers. As mentioned above, there are a few "sleeping beauty" papers that suddenly become highly cited after a few years, but that is unusual. The row for 0 early citations shows that papers with no early citations will have a few citations later on ($\mu \approx 1.4 \pm 2.3$). However, it is more common for papers that will be important to start off with more citations early on.

Table 5 is similar to Table 4 but for papers from different sources. In Table 5, the row for 3+ early citations is (usually) better than venues in terms of h and μ with an exception for TACL in 2016, because of a single highly cited outlier: (Bojanowski et al., 2017). It is risky to average over small samples of highly skewed numbers, as evidence by the large σ (standard deviations). Note that h is more stable than μ over 2016 and 2017.

Many ACL venues are highly selective in terms of μ, but ACL could improve inclusiveness (N) as well as exclusiveness (μ) by publishing more papers from preprint archives such as ArXiv with impressive early citations. We will discuss this suggestion in more detail in subsection 4.2.

3.3 Forecasting with Regression

We will use the regression model in Equation 1 to compare early returns and venue.

$$percentile_{year+4} \sim venue + factor(pmin(T, citations_{year+1})) \quad (1)$$

This model predicts the percentile of the paper in the fourth year based on the venue and early citations. Early citations are treated as a factor variable; thus, the model produces a coefficient for each count between 1 and T, as illustrated in Table 6.

This model performs a few simple transforms on both the input and output variables:

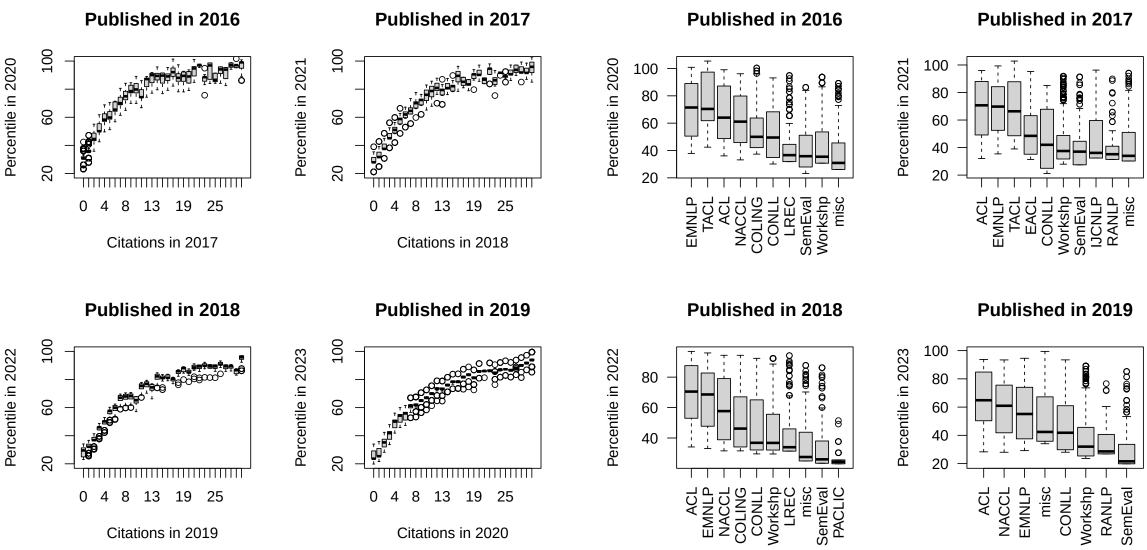
1. Percentile transform (Bornmann et al., 2012, 2014): Predict percentiles instead of raw counts. Percentiles are based on citations in fourth year after publication.
2. Thresholding transform: Since input citation counts have long tails, we use *pmin* to limit the number of factors in the regression to T.

Group	Published in 2016 h	median	μ	σ	N	Published in 2017 h	median	μ	σ	N
0 citations	47	1	1.4	2.3	265,090	33	1	1.3	2.1	268,696
1+ citations	292	3	6.8	19.8	761,729	298	4	7.3	19.1	808,772
2+ citations	291	5	8.5	22.8	557,824	298	5	9.1	22.0	591,657
3+ citations	291	6	10.3	26.1	414,641	298	7	11.0	25.2	438,445
10+ citations	289	17	27.1	54.2	81,125	297	19	28.8	52.1	84,876
20+ citations	288	37	56.4	97.3	21,329	295	41	59.5	92.8	22,119
Science	113	8	30.3	59.9	1732	97	3	22.1	48.9	1829
Nature	112	5	28.0	74.9	2094	109	4	27.1	70.7	2020
Nature Communications	88	12	18.7	25.9	3702	85	11	17.8	32.0	4505
J. Amer. Chemical Soc.	68	10	15.9	22.1	2414	68	8	13.8	20.0	2679
Proc of the Nat. Academy of Sciences of the USA	73	9	14.6	25.0	3629	68	8	12.8	19.4	3846
bioRxiv	55	6	14.1	41.9	1396	54	5	10.4	29.6	3121
Angewandte Chemie	70	7	13.5	21.3	2776	62	6	10.7	15.9	2795
Bioresource Technology	42	7	10.8	15.0	1612	40	7	10.0	11.9	1644
ACS Applied Materials...	49	7	10.0	10.7	4074	50	6	9.5	10.3	4933
Nutrients	34	5	9.4	16.6	825	38	6	8.8	14.9	1356
Frontiers in Plant Science	44	6	9.0	15.7	2045	41	6	8.7	14.0	2257
Physical Review Letters	48	5	8.8	13.4	2501	42	4	6.9	11.7	2674
Frontiers in Immunology	29	5	8.6	11.4	671	39	5	8.0	10.3	1901
Sci. of the Total Environ.	42	6	8.6	11.3	2508	46	5	8.9	15.2	2853
Frontiers in Microbiology	43	5	8.6	14.9	2175	47	5	8.2	14.2	2621
Food Chemistry	36	6	8.5	9.4	1930	31	6	8.2	7.8	1798
Chemosphere	35	5	7.7	12.1	1653	35	4	7.2	11.8	1916
Nanoscale	37	4	7.2	9.3	2166	31	4	5.9	7.8	2147
I.J. Bio. Macromolecules	28	5	7.2	8.6	1159	32	5	7.4	9.8	1616
I.J. of Molecular Sciences	37	4	7.1	11.4	2068	47	4	7.9	13.3	2746
Scientific Reports	67	4	6.8	14.5	20,860	64	4	6.3	22.2	25,006
Analytical Chemistry	29	5	6.8	7.5	1648	30	4	6.3	7.3	1817
BMJ Open	33	3	5.9	12.0	2016	31	3	4.8	8.1	2554
Molecules	34	3	5.8	12.2	1745	35	3	5.4	8.5	2247
Frontiers in Psychology	38	3	5.7	9.9	2074	32	3	5.5	12.2	2252
OncoTarget	42	4	5.5	6.8	7454	37	3	4.0	6.1	9282
Materials	24	3	5.3	7.8	1024	27	3	4.9	8.5	1473
J. of Biological Chemistry	26	4	5.2	6.5	2135	27	3	5.0	6.2	1852
Chemical Comm.	33	3	5.2	7.2	3046	25	3	3.9	4.7	2689
Italian Nat. C. on Sensors	34	3	5.2	12.9	2220	39	3	5.3	10.8	2962
British medical journal	41	0	5.2	37.2	1837	39	0	4.7	29.6	1670
I.J. Env. Res. and Pub...	24	3	5.2	7.1	1118	31	4	6.2	11.1	1575
Organic Letters	18	4	5.1	4.1	1646	16	3	4.0	3.4	1699
PLoS ONE	59	3	5.1	8.8	22,512	55	3	4.8	8.6	20,617
Environ. science and...	30	3	4.9	6.6	2430	32	3	4.9	7.5	2527
Chemistry	26	3	4.8	5.8	2271	23	3	3.8	4.5	2306
BioMed Research Inter.	25	3	4.5	8.8	1790	27	2	4.4	7.2	2005
Medicine	27	2	4.1	14.2	3275	22	2	2.8	8.5	3526
Optics Express	26	2	3.9	5.3	2871	23	2	3.4	5.0	2739
Physical Chem... - PCCP	25	2	3.6	5.4	3584	22	2	2.9	5.6	3258
RSC Advances	8	3	3.5	3.1	78	8	3	4.2	5.0	60
Biochemical... - BBRC	19	2	3.5	5.3	1744	22	2	3.6	6.6	2056
J. of Chemical Physics	23	2	3.4	7.8	2087	19	2	2.8	6.8	1944
Dalton Transactions	22	2	3.4	4.5	2085	18	2	2.9	3.8	1791
World Neurosurgery	15	2	2.7	3.7	1300	18	2	2.6	3.6	1999
Oncology Letters	15	2	2.7	3.5	1517	16	2	2.7	3.5	2207
Physical Review E	18	1	2.5	3.5	2284	18	1	2.1	3.7	2172
M. in molecular biology	18	1	1.8	4.9	2888	24	1	2.0	5.5	3612
BMJ Case Reports	9	1	1.1	1.7	1401	7	1	1.0	1.3	1689
Zootaxa	11	0	1.1	2.2	1967	9	0	1.0	3.0	1224
All other venues	274	2	5.1	17.1	878,782	273	2	4.9	18.8	913,401

Table 4: Deep dive into PubMed papers. A few early citations compare favorably to most venues. Early citations are based on first year after publication, and summary statistics are based on fourth year after publication.

Group	Published in 2016 h	median	μ	σ	N	Published in 2017 h	median	μ	σ	N
PubMed, ArXiv and ACL Anthology										
0 citations	48	1	1.3	2.3	292,566	35	1	1.3	2.1	295,467
1+ citations	345	3	6.9	24.7	828,515	339	4	7.4	24.0	883,685
2+ citations	345	5	8.7	28.6	604,536	338	5	9.2	27.8	645,904
3+ citations	345	6	10.6	32.9	448,541	338	6	11.3	32.0	479,097
10+ citations	343	17	28.6	70.1	88,490	337	19	30.0	67.5	95,105
20+ citations	341	37	61.4	127.8	23,593	336	41	62.9	122.1	25,613
ACL Anthology	73	2	9.9	59.0	3710	65	2	8.5	29.1	3030
ArXiv	236	2	6.4	47.0	101,176	234	1	6.1	58.4	110,184
PubMed	292	2	5.4	17.2	1,026,798	293	2	5.2	18.6	107,7437
Deep Dive into ACL Anthology										
0 citations	9	0	1.0	1.8	953	7	0	0.9	1.6	589
1+ citations	73	3	13.0	68.2	2757	65	3	10.3	32.2	2441
2+ citations	73	4	17.1	79.2	2025	65	4	12.9	36.0	1902
3+ citations	73	6	21.6	90.0	1550	65	5	15.7	39.8	1518
10+ citations	73	23	57.0	155.6	481	65	18	35.3	61.3	545
20+ citations	71	54	114.9	235.6	190	63	34	62.2	86.9	223
ACL Main Conf.	42	5	18.7	45.2	377	41	7	20.3	50.7	353
EMNLP	41	6	25.8	78.9	269	36	6	17.9	36.5	339
TACL	17	11	70.5	280.3	45	12	8	15.4	23.3	41
SemEval	15	1	4.7	16.3	230	14	1	5.5	23.5	208
Workshops	24	1	3.8	10.1	1111	30	1	4.2	12.5	1191

Table 5: Similar to Table 4, but for papers from different sources. Note ArXiv is better than PubMed in terms of μ.



(a) Predictions based on early citations.

(b) Predictions based on venue.

Figure 3: Boxplots of predictions from regression model for ACL papers. The bars are so narrow that they are hard to see on the left because early returns are more predictive than venue.

Coefficient	Large Set 2016	2017	Small Set 2016	2017
Intercept	15.7	15.3	36.7	32.8
ACL Anthology	4.3	1.5		
ArXiv	5.4	3.4		
PubMed	10.7	11.3		
TACL			6.8	6.5
EMNLP			1.2	2.6
COLING			0.6	NA
Workshops			-6.1	-4.9
misc			-10.6	-2.6
1 early	6.9	6.3	4.7	3.7
2 early	15.1	13.9	12.6	9.6
3 early	22.2	21.1	19.3	17.1
4 early	28.7	26.8	26.3	20.8
5 early	33.3	32.4	27.8	27.3
6 early	38.3	36.9	33.5	31.7
7 early	42.1	40.9	38.1	34.3
8 early	45.7	44.8	41.6	38.6
9 early	49.5	47.4	46.6	40.8
10+ early	59.4	59.8	54.6	54.1

Table 6: Coefficients for regression (with $T = 10$).

Because the literature is growing exponentially (Bornmann et al., 2021), care is required when comparing citations for papers published at different times (Newman, 2013). We address these concerns by fitting coefficients for each publication year.

Deep networks will likely produce better predictions, but our goal here is to estimate the value of peer-review. Is peer-review worth the cost, or should we publish more papers from ArXiv with impressive early citations?

Table 6 shows regression coefficients for $T = 10$ and two publication dates (2016 and 2017). The large set contains papers from PubMed, ArXiv and ACL. The small set is for ACL venues. The model produces coefficients for venues with 40 or more papers. Venues with less than 40 papers are assigned to *misc*. There is no coefficient for COLING in 2017 because there was no COLING meeting in 2017. To save space, some venues were omitted from Table 6.

As mentioned above, regression does not produce the best predictions in terms of loss, but it has advantages in terms of interpretability. The coefficients on early citations in Table 6 show that more early citations are better than fewer early citations. Papers with 10+ early citations are predicted to be in the 75^{th} percentile or better.

The boxplots in Figure 3 show predictions from the model with $T = 30$ for papers in the ACL Anthology published in 4 years between 2016 and 2019. The coefficients are fit four times, once for each publication year. For each year, predictions from the model are aggregated by early citations (left) and by venue (right).

The width of the bars indicates the influence of the other factor. The bars are so narrow on the left that they are hard to see, indicating that early citations are very predictive of future citations. Although venue may be statistically significant, it has relatively little consequence in practice.

These observations were confirmed by analysis of variance (ANOVA). The ANOVA shows that early citations account for much more of the variance than venue, as expected based on the discussion of correlations above.

Venues are sorted by median predictions (computed over the papers published in that year). While papers published in more prestigious venues rank higher than papers in less prestigious venues, the effect of venue is not only small, but also lacks robustness. Note that the ordering of venues varies from one year to the next: EMNLP is in the top three venues in all four plots, though it can be found in top place, second place and third place, depending on the publication year. Predictions based on early citations in Figure 3a are more consistent over the four publication years, indicating that early citations are more robust than venue. In particular, over all four panels, there is a consistent trend for predictions to increase with the number of early citations. The four panels in Figure 3a are more similar to one another than the four panels in Figure 3b.

4 Conclusions

4.1 Early Citations vs. Venue

We showed that "early returns" (citations soon after publication) are more predictive of future citations than venue. This conclusion is based on:

1. subsection 3.1: Correlations (ρ)
2. subsection 3.2: h-index (h) and Impact (μ)
3. subsection 3.3: Regression

These observations suggest a simple actionable rule-of-thumb (use early returns) that has advantages over current practice (reviewing) in terms of exclusivity, inclusivity and robustness:

1. **Exclusivity**: Simple rule of thumb: for most venues, 1+ early citations are as good as reviews in terms of μ; 20+ early citations are better than reviews for most (all) venues.
2. **Inclusivity**: There are more papers (N) with 1+ early citations than in most (all) venues.
3. **Robustness**: Results were replicated over several sources of papers and publication dates.

The rest of this paper will introduce two controversial suggestions: (1) early citations and (2) nominations to address two challenges (a) too many submissions and (b) too few qualified reviewers. Our goal is not so much to solve these challenges, but merely to jump start a discussion that might eventually lead to process improvements that will scale better than the status quo. We encourage the community, especially those that do not like (1) and (2), to offer alternative constructive suggestions.

4.2 DDI Alternative to Reviewing

A number of challenges for reviewing were mentioned: poorly defined tasks/incentives, validity, reliability, subjectivity, biases, time, cost, scale and cheating. Given these realities, is peer-reviewing worth the effort? Are there faster, cheaper and more effective alternatives?

1. Open Peer-Review (OPR) (List, 2017)
2. Don't Do It (DDI): Use early citations to reduce the load on peer-reviewing.

Since OPR "has neither a standardized definition nor an agreed schema of its features and implementations," Ross-Hellauer (2017), "proposes a pragmatic definition of [OPR] as an umbrella term for... peer review models... including making reviewer and author identities open, publishing review reports and enabling greater participation...."

The DDI alternative is even more pragmatic and constructive. Instead of reviewing papers, we suggest the community post papers on ArXiv, and use early returns to help readers, authors and committees address questions such as:

1. Readers: Who should read what?
2. Authors: Who should cite what for what?
3. Promotion and Award Committees:
   Who should be recognized for what?

4.3 New Role for Venues

What should be the role for venues under this suggestion? We suggest venues continue to publish high impact papers in their area that conform to their methods and practices, but to do so in a way that copes more effectively with scale. As mentioned above, the current system suffers from two concerns: (a) too many submissions and (b) too few qualified reviewers. We suggest introducing a process upstream of program committees to address both concerns. To reduce the load, program committees should focus on papers with impressive early citations, as well as papers nominated by a process described below in section 4.4.

In addition to the first concern, reducing the load, these suggestions also help with the second concern, identifying qualified/motivated reviewers. It should be easier for those who have cited the article to write a review since they have already read the article and most of the background material. They are not only better informed than a random reviewer, but they are also probably more sympathetic to the basic approach.

This proposal also simplifies the definition of the reviewing task. By the time reviewers see the paper, there is already considerable evidence of impact. The question for reviewers becomes more about judging fit than predicting impact.

4.4 Nomination Process

In addition to early citations, program committees should accept nominations of papers to review from thesis advisors and established researchers in industrial research laboratories, following precedents established by nomination processes for awards such as ACM Doctoral Dissertation.[8] To offset the reviewing load on society imposed by the nomination process, nominators should agree to review four papers for each paper they nominate. In this way, the proposed process addresses both concerns raised above: (a) too many submissions and (b) too few qualified/motivated reviewers.

5 Ethics

The proposed DDI method will not work with double-blind review, but people who have already cited the submission are unlikely to be biased against the submissions they have cited.

Mutual admiration societies have always existed in academia. There is a danger that the proposed DDI method will encourage those practices. However, citations leave an audit trail that makes it very easy for everyone to see what is happening. As the cliche goes, sunlight is the best disinfectant.

Reviewing is a controversial topic. From the perspective of a conference organizer, we should encourage controversial papers that engage the audience, and contribute significantly to the field.

[8] https://awards.acm.org/doctoral-dissertation/nominations

6 Limitations

Citation counts can be gamed. See discussion of cheating in subsubsection 2.2.3.

This work is largely limited to English since the venues we consider emphasize English.

There is a risk that the proposed DDI/nomination method will help the rich get richer; to compensate for this, there could be a process to encourage nominations from more diverse places.

References

Giovanni Abramo, Ciriaco Andrea D'Angelo, and Giovanni Felici. 2019. Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics*, 13(1):32–49.

Ali Abrishami and Sadegh Aliakbary. 2019. Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, 13(2):485–499.

Xiaomei Bai, Fuli Zhang, and Ivan Lee. 2019. Predicting the citations of scholarly paper. *J. Informetrics*, 13:407–418.

Joeran Beel and Bela Gipp. 2010. Academic search engine spam and Google Scholar's resilience against it. *Journal of Electronic Publishing*, 13.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Lutz Bornmann, Loet Leydesdorff, and Rüdiger Mutz. 2012. The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *ArXiv*, abs/1211.0381.

Lutz Bornmann, Loet Leydesdorff, and Jian Wang. 2014. How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics*, 8(1):175–180.

Anthony Chauvin, Philippe Ravaud, Gabriel Baron, Caroline Barnes, and Isabelle Boutron. 2015. The most important tasks for peer reviewers evaluating a randomized controlled trial are not congruent with the tasks most often requested by journal editors. *BMC medicine*, 13:1–10.

Kunming Cheng, Zaijie Sun, Xiaojun Liu, Haiyang Wu, and Cheng Li. 2024. Generative artificial intelligence is infiltrating peer review process. *Critical Care*, 28(1):149.

Kenneth Church. 2005. Last words: Reviewing the reviewers. *Computational Linguistics*, 31(4):575–578.

Kenneth Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Kenneth Church. 2020. Emerging trends: Reviewing the reviewers (again). *Natural Language Engineering*, 26(2):245–257.

Corinna Cortes and Neil D. Lawrence. 2021. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *ArXiv*, abs/2109.09774.

Feruz Davletov, Ali Selman Aydin, and Ali Cakmak. 2014. High impact academic paper prediction using temporal and topological features. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 491–498, New York, NY, USA. Association for Computing Machinery.

Dennis R De Vries, Elizabeth A Marschall, and Roy A Stein. 2009. Exploring the peer review process: what is it, does it work, and can it be improved? *Fisheries*, 34(6):270–279.

Joost de Winter. 2024. Can ChatGPT be used to predict citation counts, readership, and social media interaction? an exploration among 2222 scientific abstracts. *Scientometrics*, 129(4):2469–2487.

Gunther Eysenbach. 2006. Citation advantage of open access articles. *PLoS biology*, 4(5):e157.

Eugene Garfield. 2006. The history and meaning of the journal impact factor. *JAMA*, 295 1:90–3.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B. Shah. 2023. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *ArXiv*, abs/2311.09497.

Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for machine learning on graphs. *ArXiv*, abs/2005.00687.

Shengzhi Huang, Yong Huang, Yi Bu, Wei Lu, Jiajia Qian, and Dan Wang. 2022. Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. *Inf. Process. Manage.*, 59(2).

Jürgen Huber, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting, Stefan Palan, and Vernon L. Smith. 2022. Nobel and novice: Author prominence affects peer review. *Proceedings of the National Academy of Sciences*, 119(41):e2205779119.

Ken Hyland. 2023. Enter the dragon: China and global academic publishing. *Learned Publishing*, 36(3):394–403.

Tom Jefferson, Philip Alderson, Elizabeth Wager, and Frank Davidoff. 2002. Effects of Editorial Peer Review: A Systematic Review. *JAMA*, 287(21):2784–2786.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Steve Lawrence. 2001. Free online availability substantially increases a paper's impact. *Nature*, 411(6837):521–521.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17.

Benjamin List. 2017. Crowd-based peer review can be good and fast. *Nature*, 546:9–9.

Anqi Ma, Yu Liu, Xiujuan Xu, and Tao Dong. 2021. A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics*, 126(8):6803–6823.

Mark E. J. Newman. 2013. Prediction of highly cited papers. *Europhysics Letters*, 105.

Natsuo Onodera and Fuyuki Yoshikane. 2015. Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4):739–764.

Nataliia Pobiedina and Ryutaro Ichise. 2016. Citation count prediction as a link prediction problem. *Applied Intelligence*, 44:252–268.

Sidney Redner. 2005. Citation statistics from 110 years of Physical Review. *Physics Today*, 58:49–54.

Giuseppe Remuzzi. 2023. The ethics of peer review process. *Updates in Surgery*, 75(6):1391–1392.

Sara Rockwell. 2006. Ethics of peer review: a guide for manuscript reviewers. *URL: http://ori. dhhs. gov/education/products/yale/prethics. pdf.[Accessed on: 01-09-2013]*.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Tony Ross-Hellauer. 2017. What is open peer review? a systematic review. *F1000Research*, 6.

Xuanmin Ruan, Yuanyang Zhu, Jiang Li, and Ying Cheng. 2020. Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics*, 14(3):101039.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191.

Olivia M Smith, Kayla L Davis, Riley B Pizza, Robin Waterman, Kara C Dobson, Brianna Foster, Julie C Jarvey, Leonard N Jones, Wendy Leuenberger, Nan Nourn, et al. 2023. Peer review perpetuates barriers for historically excluded groups. *Nature Ecology & Evolution*, 7(4):512–523.

Lawrence Souder. 2011. The ethics of scholarly peer review: a review of the literature. *Learned Publishing*, 24(1):55–72.

Clara Stegehuis, Nelly Litvak, and Ludo Waltman. 2015. Predicting the long-term citation impact of recent publications. *Journal of informetrics*, 9(3):642–657.

Iman Tahamtan, Askar Safipour Afshar, and Khadijeh Ahamdzadeh. 2016. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, 107:1195–1225.

Anthony F. J. van Raan. 2004. Sleeping beauties in science. *Scientometrics*, 59:467–472.

Alex D Wade. 2022. The Semantic Scholar Academic Graph (S2AG). *Companion Proceedings of the Web Conference 2022*.

Dashun Wang, Chaoming Song, and Albert Ĺaszló Barabási. 2013. Quantifying long-term scientific impact. *Science*, 342:127 – 132.

Pengwei Yan, Yangyang Kang, Zhuoren Jiang, Kaisong Song, Tianqianjin Lin, Changlong Sun, and Xiaozhong Liu. 2024. Modeling scholarly collaboration and temporal dynamics in citation networks for impact prediction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2522–2526, New York, NY, USA. Association for Computing Machinery.

Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 1247–1252, New York, NY, USA. Association for Computing Machinery.