https://doi.org/10.51408/issi2025_080

Exploring the Application of Open Peer Review in Academic Evaluation: An Analysis of H1 Connect Recommended Papers

Liu Xiaojuan[1], Xiang Nannan[2], Yu Yao[3]

[1]*lxj_2007@bnu.edu.cn*, [2]*bnu_xnn@163.com*, [3]*yuyao990824@163.com*

School of Government, Beijing Normal University, No. 19, Xinjiekouwai Street, Haidian Beijing (China)

Abstract

The development of open peer review has provided a new perspective on academic evaluation. By exploring the relationship between peer review indicators and impact indicators including Citation and AAS, as well as delving into the value of papers from the perspective of peer reviewers, this research offers insights to improve academic evaluation systems. The study focuses on papers about three topics: *Cardiovascular Diseases*, *Respiratory Tract Diseases*, and *Neoplasms*. It utilizes open peer reviews from H1 Connect, analyzing them from two dimensions: review indicators and scientific research contributions. Regarding review indicators, attention is paid to the RNumber and the RStar. The analysis of contributions is based on the Becker Medical Library's research evaluation model, which is used to design a classification system for contribution types. This study employs the "GPT-4O-mini" model to extract sentences describing scientific research contributions from peer review texts, and then categorizes them according to the designed classification system. The findings reveal that, in terms of review indicators, there are significant differences across topics, with a notable positive correlation existing between the RNumber, RStar, Citation, and AAS. In terms of scientific research contributions, these contributions are primarily concentrated in the dimensions of Knowledge Advancement and Clinical Implementation, with slight differences in contribution types among the topics. Contributions regarding clinical trial outcomes and healthcare services are more prominent in *Cardiovascular Diseases*, while theoretical contributions are more apparent in *Respiratory Tract Diseases*. Regarding contribution co-occurrence, Knowledge Advancement and Economic and Community Benefits contributions often do not occur simultaneously. Papers that contribute to the discovery of new ideas, data methods, or clinical management and treatment are more likely to exhibit multiple types of contributions. Contributions to public health policies often appear separately. Generally, papers tend to focus on making significant contributions in one specific area, with the occurrence of multiple types of contributions being relatively rare. Academic evaluation should effectively integrate peer review with impact indicators, while deeply exploring the scientific research contributions of papers. It is crucial to consider both the diversity of contributions and the thematic differences to build a more comprehensive, scientific, and effective academic evaluation system.

Introduction

Peer review is a key mechanism to ensure the quality of publications, maintain academic integrity, and promote scholarly communication. The peer review process is intended to help improve research reporting and weed out work that does not meet the research community's standards for research production (Wolfram et al., 2020). It relies on the expertise and judgment of experts in the field to evaluate academic papers, project proposals, or research achievements to determine whether they fulfill the criteria for publication or funding. However, the traditional peer review process is often regarded as a closed and opaque "black box operation". Its information such as the decision basis, review texts, and reviewer identity is often not disclosed, which not only limits the transparency of the research process but also may lead to bias (Demarest et al., 2014; Fox & Paine, 2019), unfairness (Bravo et al., 2018) and inefficiency.

Open Science came into being to improve the transparency, fairness, and efficiency of scientific research and evaluation. It has become an important concept to promote the sustainable development of scientific research. Open Science advocates the openness and transparency of all facets of scientific research, and open peer review (OPR) is the last frontier of Open Science that has yet to achieve widespread adoption (Wang et al., 2016), has gradually become one of the means to overcome some limitations of traditional peer review.

Through the efforts of relevant institutions to enhance the transparency of the academic publishing process and oversee the peer review work, the credibility of peer review can be improved. This helps reduce unjust, unprofessional, and unnecessary evaluations of papers, thereby advancing the goals of follow-up reviews, peer review accountability, and review quality supervision (Wang, 2023). An increasing number of journals and conferences have started to implement the open peer review mechanism in recent years, and open peer review platforms such as H1 Connect, Publons, and Pubpeer have also emerged. These platforms significantly lessen the difficulty of obtaining peer review data and further enrich the types of peer review data, including review texts, review scores, review numbers, etc. Open peer review data are the tangible representation of expert opinions, with greater professionalism, transparency, and credibility than traditional citation data and altmetrics data. It also has rich value, offering a foundation for investigating the behavior of peer review, identifying the traits of expert reviews, and exposing the peer review process's working mechanism.

Focused on the open peer reviews from H1 Connect, this study analyzes the reviews from two dimensions: the numerical characteristics and the scientific research contributions. Additionally, impact indicators, including Citation and Altmetric

Attention Score（hereinafter abbreviated as AAS）, are incorporated to explore the relationship with peer review indicators. By integrating these dimensions and impact indicators, our goal is to delve deeper into the realm of peer reviews and investigate their potential value in academic evaluation. To be more specific, this study seeks to answer the following questions: What distribution characteristics can be observed in open peer review indicators? Are there differences across research topics? What is the relationship between peer review indicators, Citation and AAS? Do papers that receive higher recognition from peer reviewers tend to achieve higher impact? What research contributions are embedded in open peer review texts, and how are these contributions distributed and co-occurring?

Literature review

With the growing momentum of the open science movement, an increasing number of journals and publishers are joining the ranks of those sharing peer review data. Meanwhile, numerous open peer review platforms have emerged, laying a practical groundwork for peer reviews exploration. The development of technologies such as natural language processing and sentiment analysis provides technical support for the implementation of peer review mining. In addition to the review comments in the form of text (hereinafter abbreviated as "peer review texts"), there are also various forms such as review scores, review numbers, review stars, and review labels. Many scholars have carried out analysis and utilization research on different types of peer reviews.

Open peer review, Citation and AAS

Some studies have explored the effect of open peer review on citation and AAS. Zong et al. (2020), using PeerJ as an example, found that articles with open peer review history could be expected to have significantly higher citations than those with a traditional review pattern, but there would be variations among disciplines. However, some investigations reach different conclusions. According to Ni et al. (2021), there is no evidence of a citation advantage for the papers disclosing their peer review documents by taking *Nature Communications* as an example. Articles subjected to OPR have no obvious advantage in citation but a notably higher score in altmetrics (Cheng et al., 2024). Xie et al. (2024) revealed that different types of papers have significant differences in review scores and citations, and there is a positive correlation between review scores and citations.

Sentiment analysis of peer reviews

Peer review texts often contain rich sentimental information, reflecting the reviewers' overall attitude toward the research presented in the paper. Therefore, sentiment analysis is widely employed in peer review text mining, and most studies aim to classify the sentiment polarity of peer review texts. Wang et al. (2018) introduced sentiment analysis into peer review texts analysis for the first time. By using automatic identification, they detected sentence fragments with positive or negative connotations. These fragments, representing sentiment polarity, were then used to predict the final score of a paper. Based on the sentiment information in the authors' comments and the content of the peer review texts, Ghosal et al. (2020) developed the DeepSentiPeer model to forecast the overall recommendation score and ultimate decision of the work. Bravo et al. (2019) examined whether the language style of the reviewers changed after the journal opened the peer review report, using continuous numerical values to represent the sentiment polarity and subjectivity of the review texts. Lin et al. (2021) employed the sentiment analysis model to mine the sentiment polarity of open review texts. They used the titles, abstract, Twitter comments, and peer review texts as input to the model, with the average review scores as the actual score. The evaluation of the paper was based on the sentiment polarity of the review texts. Some scholars have further combined the sentiment polarity of peer review texts with citations. Zong et al. (2020) investigated the relationship between the sentiment polarity of peer review comments and citations using data from PubPeer, F1000, and ResearchGate. They discovered that in comparison to the comparable control pairings (articles without PPPRs), papers that obtained favorable post-publication peer reviews (PPPRs) had noticeably higher citations. However, the control group, which included papers with neutral or negative ratings and papers with both positive and negative reviews, did not differ significantly in citations.

Identification of elements in peer review text

Peer review texts on academic papers are typically long and structurally complex. Identifying the elements contained in them can help gain a deep understanding of the peer review mechanism and its value. At present, many studies have defined and identified the types of elements from different perspectives. Hua et al. (2019) divided the elements into evaluation, request, fact, reference, and quote. They then examined the effects of several models on element identification and found that the Bi-LSTM-CRF model had the best effect. Fromm et al. (2021) separated the elements into non-arguments, supporting arguments, and opposing arguments, and tested the performance of the Bert model in the argument extraction task. They also pointed out that peer review texts differ from other types of subjective texts (such as legal

documents and e-commerce reviews) in terms of length, tone, and wording. Chen et al. (2023) separated the elements into four categories, including overview, method, result, and highlights, using the step type definition in conjunction with the research corpus's content characteristics. They then evaluated the recognition efficacy of SVM, FastText, TextCNN, and BiLSTM models, concluding that the BiLSTM model performed the best. Ghosal et al. (2022), using the ICLR peer review dataset as an example, categorized peer review texts into four dimensions: the section of the paper that the review comments on (e.g., Introduction, Methodology, Data, Experiments), the aspect of the paper that the review addresses (e.g., Appropriateness, Originality or Novelty, Clarity), the purpose or the role of the review (e.g., Suggestion, Discussion, Question), and the significance of the review (e.g., Major Comment, Minor Comment, General Comment). Zhang et al. (2022), using 3329 comments from 690 papers published in the British Medical Journal (BMJ) as the research objects, analyzed the differences in the length distribution of reviewers' comments, the general distribution of words in comments and the position of reviewer comments. Wang et al. (2020) analyzed the review texts of papers published in journals such as *Cell* and *The Lancet* recommended by F1000Prime and found that the most frequently used words by experts included interesting, important, first, exceptional, etc.

In summary, existing research primarily focuses on the analysis of open peer review indicators, sentiment analysis, and element recognition based on peer review texts. While some studies examine the characteristics of open peer review data from various perspectives, most of them address only a limited number of indicators and rarely consider the inherent characteristics of the papers themselves. In this study, we take a more comprehensive approach by analyzing both the textual and numerical aspects of peer review data. Methodologically, most existing studies rely on machine learning and deep learning models to analyze the content of review texts. However, the generalizability, adaptability, and enhanced capabilities of large language models in feature extraction, semantic understanding, and multimodal learning provide models with significant advantages in identifying elements within peer review texts. In this study, we introduce large language models to extract research contributions. The aim is to comprehensively reveal the value of papers from the perspective of peer reviewers, enhance the understanding of post-publication open peer reviews, and provide insights into the application of peer review in academic evaluation within the context of open science.

Data and method

Data

Among the many open peer review platforms, H1 Connect (formerly F1000, F1000 Prime and Faculty Opinions) has been the most authoritative representative in the global biomedical field in the past twenty years. It brings together nearly ten thousand top experts in the field, aiming to further recommend and evaluate papers that have been published after traditional peer review. Therefore, this study uses H1 Connect as the source of open peer review data.

Neoplasms, *Cardiovascular Diseases*, and *Respiratory Tract Diseases* are characterized by high morbidity and mortality, severely affecting human health. This study selected academic papers on these three topics for research. First, a search was conducted in the PubMed database using “Neoplasms,” “Cardiovascular Diseases,” and “Respiratory Tract Diseases” as MeSH Major Topics, with the time frame limited to January 2015 to December 2020, and the document types restricted to "Article" or "Review." A total of 1,496,535 papers were retrieved, of which 10,810 were recommended by H1 Connect, including 9,580 articles and 1,230 reviews. Among the recommended papers, there are 3,526 papers on *Cardiovascular Diseases* (hereinafter abbreviated as C), 2,488 papers on *Respiratory Tract Diseases* (hereinafter abbreviated as R), and 5,640 papers on *Neoplasms* (hereinafter abbreviated as N). It should be noted that some papers belong to multiple topics simultaneously. Next, we collected the Citation and altmetrics data for the recommended papers using their DOI from Web of Science and Altmetric.com. Finally, we used a self-written Python program to scrape open peer reviews on H1 Connect, obtaining a total of 12,203 reviews The final dataset collected includes the topic, paper title, publication year (hereinafter referred to as Year), Citation, AAS, type of document (hereinafter referred to as Type), review number (hereinafter abbreviated as RNumber), review star (hereinafter abbreviated as RStar), and review text. The distribution of papers by publication year is shown in Table 1.

Table 1. Publication year distribution of papers.

Year	*Paper count*
2015	2023
2016	1999
2017	1939
2018	1775
2019	1657
2020	1417
Whole	10810

Method

(1) Extraction of scientific research contribution sentences

The scientific research contribution refers to the ability of the current research to improve, perfect and apply existing knowledge, theories or practices (Luo et al., 2021), including new theories, new methods, new technologies, new outcomes. Analyzing and evaluating these contributions is a necessary step in evaluating the quality of the paper and promoting knowledge innovation and disciplinary progress. Previous studies analyzing the scientific research contributions of papers were mainly based on abstract or full-text datasets and relied mainly on the authors' descriptions, which introduces a certain degree of subjectivity. In contrast, the insights and evaluations in peer review texts come from authoritative experts, making them an important reference for uncovering the paper's scientific research contributions. Therefore, this study further explores the scientific research contributions of papers based on peer review texts.

Traditional deep learning models rely heavily on large-scale, high-quality annotated data. The powerful contextual understanding ability of large language models enables them to achieve excellent performance in downstream tasks with only a small number of examples or direct prompts, thereby shifting the paradigm of information extraction tasks from fine-tuning to zero-shot/few-shot (Shi et al., 2024). This study uses the "GPT-4O-mini" model to extract scientific research contribution sentences from peer review texts. Firstly, combining with the definition of scientific research contribution, this study argues that the scientific research contribution sentence in peer review texts should meet both of the following conditions: (1) The sentence must explicitly mention the study. (2) The sentence must express the experts' recognition of the study’s value. This study designs the model prompt based on this, as shown in Figure 1. Secondly, a test sample of 1,000 review texts was constructed and manually annotated according to the two conditions. The extraction

performance of zero-shot, one-shot, and few-shot prompt strategies is tested respectively. Through experiments, it is found that in a few cases, the model's output might slightly change the original sentence. Therefore, further processing of the model's extraction results is necessary. By writing code to determine whether the extracted sentences are the original sentences of the review text, we match the sentences that do not meet the requirements to the original text based on cosine similarity. Next, the micro-average index is used to evaluate the extraction performance of different prompt strategies, and then the strategy with the best performance is selected to extract all review texts. Finally, the extraction results are manually verified.

You are a top scholar in the field of medicine. Below is a peer review text from an expert regarding a medical research paper. Please extract all the sentences from the expert that evaluate the value of the paper. Each extracted sentence must meet both of the following criteria:

1. The sentence must explicitly mention the study (e.g., "this study," "this article," "this paper," or specific descriptions of the study's content).
2. The sentence must express recognition of the study's value (e.g., "help," "valuable," "important," "novel," or "new method"). Avoid sentences that merely describe the study's content without including a value judgment.

Output requirements:

1. Only output the original extracted sentence(s) exactly as written, with no changes.
2. If multiple sentences are extracted, separate each sentence with a "\n\n".
3. If no relevant content is found, output "None.".

Here is an example:

Input: peer review text.
Output: sentence1. \n\n sentence2……

Input: ……

Figure 1. Model prompt.

The extraction performance of different prompt strategies is shown in Table 2. It can be found that the optimal F_1 value of the zero-shot strategy can reach 74.42%. Therefore, the zero-shot prompt strategy is used to extract all peer review texts. The scientific research contribution sentences have been extracted, totaling 7,290(including 279 non-original sentences, accounting for only 3.83%). After manual verification and filtering, 5021 sentences remain, involving a total of 3207 papers.

Table 2. Performance of different prompt strategies.

Prompt strategy	*P*	*R*	F_1
0-shot	71.14%	78.01%	74.42%
1-shot	69.08%	75.31%	72.06%
few-shot	59.94%	80.08%	68.56%

(2) Classification of scientific research contributions

The Becker Medical Library Research Evaluation Model, designed by Washington University School of Medicine, aims to go beyond traditional citation analysis indicators to comprehensively assess the value and impact of medical research. The model tracks research output, dissemination, and transformation, providing a comprehensive evaluation of biomedical research across five dimensions: advancement of knowledge, clinical implementation, legislation and policy, economic benefit, and community benefit. This study refers to the model and combines the actual characteristics of the extracted sentences to divide scientific research contributions into nine types from three dimensions. Relevant explanations and examples can be found in Table 3. Manual annotation is conducted based on this categorization system.

Table 3. Classification and explanation of the types of scientific research contributions.

Contribution Type	*Contribution Subtype*	*Explanation*	*Example*
1 Knowledge Advancement: Research outcomes contribute to the expansion and promotion of the knowledge system	1.1 Concepts & Theories	Initiating new research directions; proposing new theoretical frameworks, concepts, or hypotheses.	This study **creates a new paradigm** in critical care medicine.
	1.2 Insights & Findings	Formulating new insights, findings, conclusions, or confirmations during the research process.	These observations **add significant new insights** to our understanding.....

	1.3 Data & Methods	Constructing meaningful datasets; proposing or improving new methods, strategies, or pathways to research questions.	This paper **reports on a new methodology** .....
2 Clinical Implementation: Research outcomes contribute to the improvement of clinical practice	2.1 Medical Products	Research outcomes aid in the selection and development of medical products, such as pharmaceuticals, biomaterials, and medical devices.	Such genome-wide systematic and unbiased strategies could **help in developing a wide range of drugs**.....
	2.2 Clinical Management and Treatment	Research outcomes contribute to clinical decision-making, optimizing clinical management, or enhancing clinical treatment plans.	The data therefore **open new therapeutic avenues.**
	2.3 Clinical Trial Outcomes	Clinical trials have achieved valuable outcomes.	**The WINTHER clinical trial provides a glimpse of the value of** .....
3 Economic and Community Benefits: Research outcomes can enhance economic benefits or improve community welfare	3.1 Healthcare Services	Improving health conditions; enhancing health literacy; reducing service costs.	This may **help to lower resource use, costs, and enhance quality and value of care.**
	3.2 Morbidity & Mortality	Alleviating the disease burden; decreasing morbidity and mortality rates; increasing survival rates.	This review has **important implications for prevention** of VTE as a major cause of maternal **mortality and morbidity**.

3.3 Public Health Policy	Providing a scientific basis for the formulation of public health policies, guidelines and related measures.	This study **justifies the policy**.....

Results

Impact indicators of recommended papers

Citation analysis evaluates academic impact within a specific discipline; altmetrics emphasizes social impact on the public, and peer review provides an in-depth evaluation of a paper's content from an expert perspective. To analyze the characteristics of recommended papers from multiple perspectives and provide a reference for subsequent comparative analysis of peer review comments, this study first analyzes two commonly used impact indicators, Citation and AAS, to explore the impact of the recommended papers.

(1) Citation. The citation of recommended papers (Table 4, Figure 2) is highly dispersed, with a large span, ranging from a minimum of 0 to a maximum of 21,917, and an average of 203.95. Among them, papers on *Respiratory Tract Diseases* have a higher average Citation (243.94) and are the most dispersed, while the Citation of papers on *Cardiovascular Diseases* is generally concentrated at a lower level. Among these recommended papers, papers on *Respiratory Tract Diseases* have a higher effect on the academic community.

Table 4. Distribution of Citation on different topics.

	Mean	*Min*	*Max*	*SD*
Whole Data	203.95	0	21917	557.9242
Cardiovascular Diseases	163.96	0	5885	396.77
Respiratory Tract Diseases	243.94	0	21917	836.66
Neoplasms	216.57	0	9728	488.99

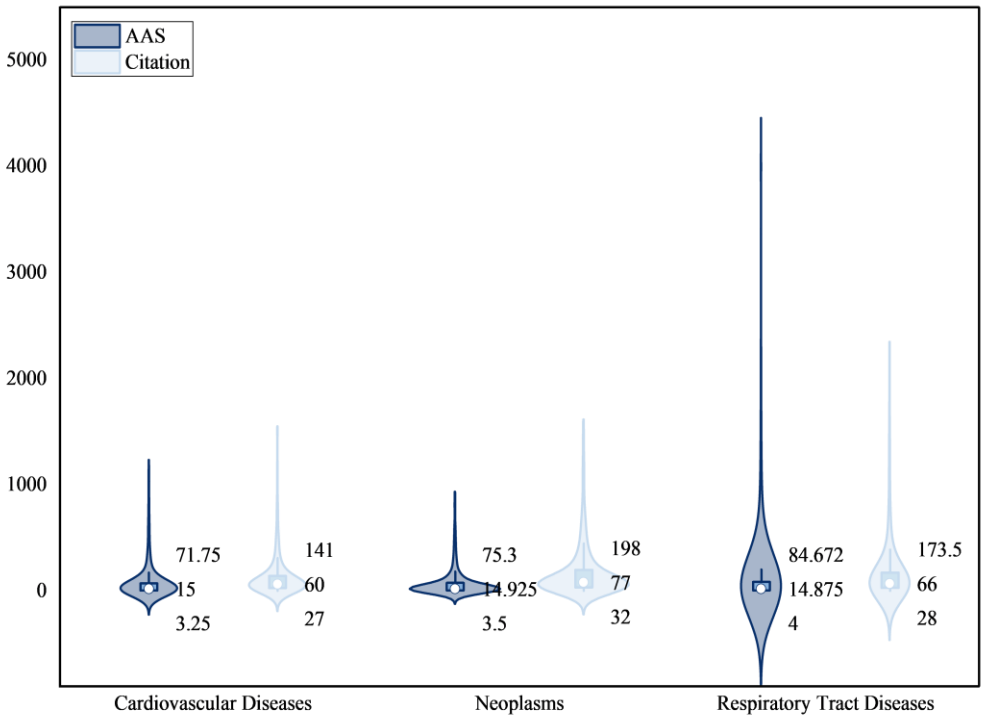

Figure 2. Distribution of impact indicators on different topics.

(2) AAS. The distribution of AAS is similar to that of Citation, with the data being highly dispersed and spanning a large range, from 1 to 32,243.46 (Table 5). The average values of AAS for each topic are significantly different. The papers on *Respiratory Tract Diseases* have a comparatively higher AAS, with an average value of two to four times that of other topics. While the AAS of papers on *Neoplasms* are concentrated at lower levels and have a lower degree of dispersion. It is evident that there are differences in the level of public attention towards papers on different topics. Papers belonging to *Respiratory Tract Diseases* generally have a higher and more scattered social impact, while papers on *Neoplasms* show more consistent levels of social attention.

Table 5. Distribution of AAS on each topic.

	Mean	*Min*	*Max*	*SD*
Whole Data	152.91	1	32243.46	782.56
Cardiovascular Diseases	120.60	1	12737.04	420.38
Respiratory Tract Diseases	341.31	1	32243.46	1531.80
Neoplasms	89.97	1	7380.74	250.70

Preliminary analysis reveals differences in the impact of papers across the three topics. To further examine these differences, this study performs differential tests on Citation and AAS. Due to the non-normal distribution of the data, the Kruskal-Wallis H test is conducted to analyze the data differences both among the three topics and between pairs of topics, as shown in Table 6. There is a statistically significant difference in Citation among the three topics ($H=70.682$, $p<0.001$), and a significant difference in AAS among the three topics ($H=10.820$, $p<0.01$). An analysis of pairwise topic differences is conducted, with each row in the table testing the null hypothesis that "the distributions of Topic 1 and Topic 2 are the same." The significance values have been adjusted by Bonferroni correction for multiple comparisons. Regarding Citation, significant differences in data distributions were observed between all pairs of the three topics. However, for AAS, the differences between topics varied. *Neoplasms* showed significant differences in data distributions compared to the other two topics, while the AAS data distributions for *Cardiovascular Diseases* and *Respiratory Tract Diseases* were nearly the same.

Table 6. Differential test of Citation and AAS across topics.

	H	*P_value*	*Group*	*P_value*
Citation	70.682	0.000***	C-R	0.003**
			C-N	0.000***
			R-N	0.000***
AAS	10.820	0.004**	C-R	1.000
			C-N	0.010**
			R-N	0.008**

* p<=0.05, ** p<=0.01, *** p<=0.001.

Peer review indicators of recommended papers

(1) RNumber. The number of reviews can reflect the degree of attention paid by the experts to the paper. The average of RNumber is 1.13, with the majority (90.63%) of papers recommended only once by experts, and a very small proportion (0.22%) receiving 5 or more recommendations. The highest RNumber obtained by a paper is 11. The pairwise distribution of the RNumber and the RStar is shown in the center scatter plot of Figure 3, where it is evident that there is a linearly positive correlation between the RStar and the RNumber of the paper. Papers with a higher RNumber tend to receive higher RStar. The distribution of RNumber and RStar for papers under various topics is displayed in the box plots on the top and right sides, respectively. It demonstrates that the publications on *Respiratory Tract Diseases*

have been recommended comparatively more frequently and are distributed more widely.

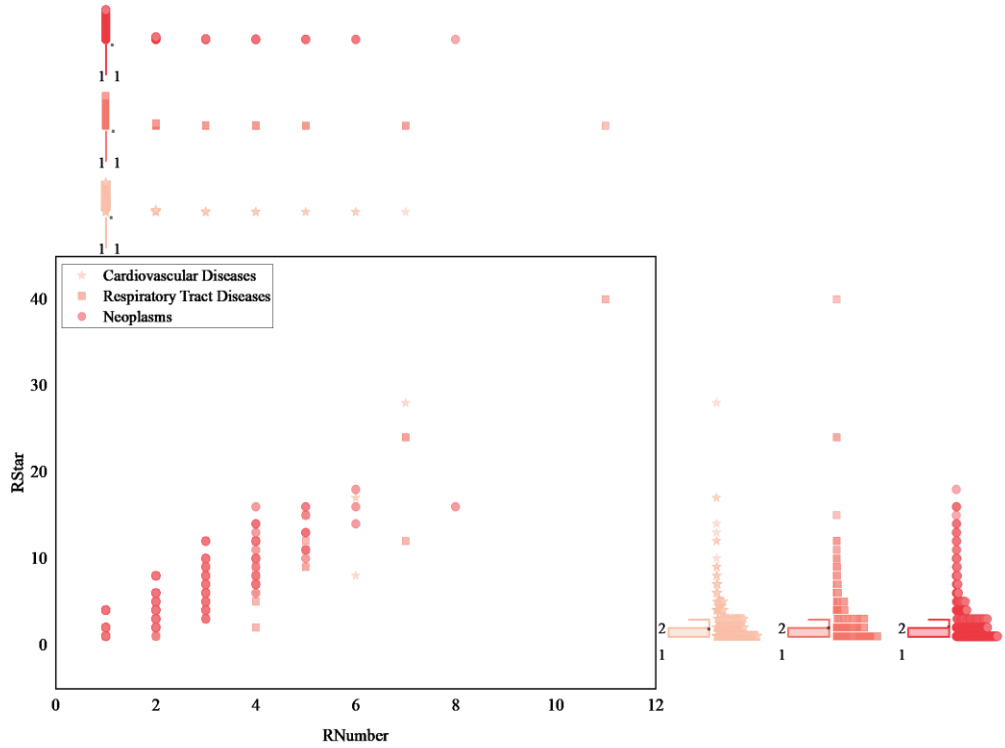

Figure 3. RNumber and RStar of papers on different topics.

(2) RStar. RStar can reflect the experts' recognition of the content and value of the paper. The RStar has a wide range, with a minimum of 1 and a maximum of 40. The mean and standard deviation of the RStar value are 1.96 and 1.64, respectively. RStar is typically low and concentrated. Only 0.46% of papers have an RStar of greater than ten, while the majority of papers (83.65%) are concentrated between one and two. Nearly half (49.70%) of the papers have an RStar of one. The papers on *Respiratory Tract Diseases* have a comparatively high RStar. Three of the four papers with RStar more than twenty are related to *Respiratory Tract Diseases*, while one belongs to *Cardiovascular Diseases*. On average, papers on *Neoplasms* received a higher average of 2.04, and the span of RStar obtained was also the smallest (1~18). In terms of dispersion, the RStar of papers on *Cardiovascular Diseases* is the most concentrated, while those on *Respiratory Tract Diseases* are the most dispersed.

To further investigate whether there are differences in distributions of peer review indicators among papers with different topics, we conducted difference tests on RNumber and RStar, respectively. Given that the tested data exhibited a non-normal distribution, non-parametric tests were employed. Specifically, this study utilized the

Kruskal-Wallis H test to analyze the differences in data among three topics and between each pair of topics. The results are presented in Table 7. The results indicated the presence of statistically significant differences in RNumber among the three topics ($H=10.860$, $p<0.001$), as well as marked differences in RStar across these topics ($H=38.837$, $p<0.001$). These results suggest that experts taking part in open peer review have varying levels of attention and recognition towards papers of distinct topics. Pairwise comparisons of topic differences were conducted. The null hypothesis that "the distributions of Topic 1 and Topic 2 are the same" was tested for each row in the table. The Bonferroni correction method was used to modify the significance values for multiple tests. In terms of the RNumber, *Cardiovascular Diseases* shows significant differences from the other two topics, while *Respiratory Tract Diseases* and *Neoplasms* are nearly the same. In terms of the RStar, *Neoplasms* is significantly different from the other two topics, while there is no significant difference between *Cardiovascular Diseases* and *Respiratory Tract Diseases*.

Table 7. Differential test of RNumber and RStar across topics.

	H	*P_value*	*Group*	*P_value*
RNumber	10.860	0.004***	C-R	0.030*
			C-N	0.006**
			R-N	0.884
RStar	38.837	0.000***	C-R	0.132
			C-N	0.000***
			R-N	0.003**

* p<=0.05, ** p<=0.01, *** p<=0.001.

Correlation test between peer review indicators and impact indicators of recommended papers

After conducting the Kolmogorov-Smirnov (K-S) test, it was determined that the RNumber, RStar, Citation, and AAS did not follow a normal distribution. Therefore, Spearman's correlation analysis was employed to assess the correlations among these indicators, with Spearman's rank correlation coefficient serving as a measure of the strength of these relationships. The results of the correlation test are presented in Figure 6. The upper right triangular area indicates the significance levels of the correlations, with the shape and color of the ellipses representing the positive or negative nature of the correlations. Positive correlations are depicted as upward-facing ellipses, where a darker color signifies a stronger correlation. The numerical values in the lower left triangular area represent the correlation coefficients, with

values closer to 1 indicating stronger positive correlations. It is observed that there is a significant positive correlation between the open peer review indicators of papers and impact indicators. Within each group of indicators, namely between RNumber and RStar, as well as between Citation and AAS, there are also significant positive correlations. Among them, the positive correlation between RNumber and RStar, and between Citation and AAS is higher (the two correlation coefficients are 0.68 and 0.48, respectively). This means that papers with more recommendations would be given higher review stars, and similarly, papers with higher citations would be given higher AAS.

RStar has a stronger positive effect on impact indicators than RNumber. The number of reviews positively affects both the Citation and AAS of a paper to a similar extent. The more reviews a paper receives, the wider its dissemination in academia and society, and the greater its impact. There is a strong positive correlation between RStar, Citation, and AAS, with RStar exerting a somewhat stronger positive effect on AAS. This suggests that papers that receive more positive reviews from experts will have higher citations and AAS.

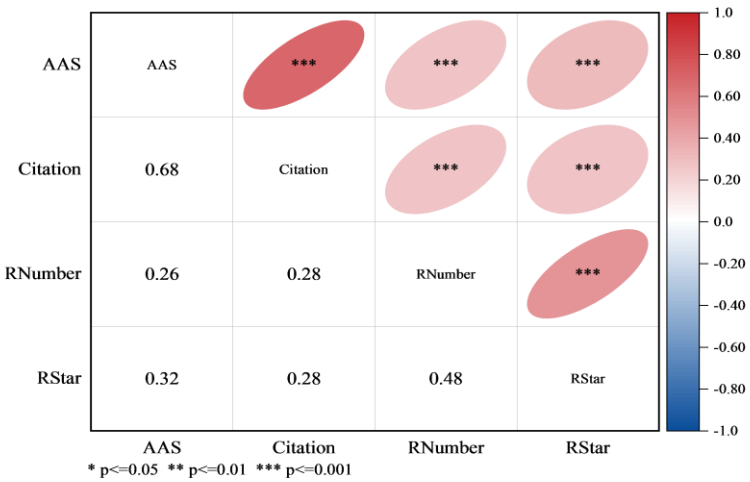

Figure 4. Correlation Test.

Spearman's correlation test was further performed on RNumber, RStar, Citation, and AAS under each topic, and the results are shown in Table 8. The correlation test findings of these variables for each topic show substantial positive correlations, which are basically consistent with the overall data. Notably, the strongest link is seen between Citation and AAS. The degree of correlation among different topics across various indicators varies. Except for a somewhat lower correlation between AAS and RStar compared with the situation in *Neoplasms*, *Respiratory Tract Diseases* shows

stronger relationships among all indicators than the other two topics. *Cardiovascular Diseases* has the poorest positive correlations among the indicators.

Table 8. Correlation test for different topics.

	Cardiovascular Diseases				*Respiratory Tract Diseases*			
	AAS	*Citation*	*RNumber*	*RStar*	*AAS*	*Citation*	*RNumber*	*RStar*
AAS	1.000				1.000			
Citation	.661**	1.000			.703**	1.000		
RNumber	.246**	.265**	1.000		.275**	.303**	1.000	
RStar	.223**	.192**	.453**	1.000	.323**	.328**	.501**	1.000

	Neoplasms			
	AAS	*Citation*	*RNumber*	*RStar*
AAS	1.000			
Citation	.699**	1.000		
RNumber	.272**	.270**	1.000	
RStar	.367**	.295**	.491**	1.000

Scientific research contributions of recommended papers

(1) Distribution of scientific research contributions

The review text can reflect various contributions of the paper in different aspects (Qin, 2020). Figure 4 shows the distribution of the nine contribution types in the three dimensions involved in the review. The paper's contributions are more prominent in the areas of knowledge advancement, followed by clinical implementation, with relatively less emphasis on economic and community benefits. Among these, the reviewers focus more on the insights and findings of the paper, its value for clinical management and treatment, and the data and methods used in the paper. This aligns with the findings of previous research. Some studies on the reviews of academic papers in different fields have found that research methods, as an

important part of the paper, are the focus of the reviewers (Han et al., 2022; Qin, 2020).

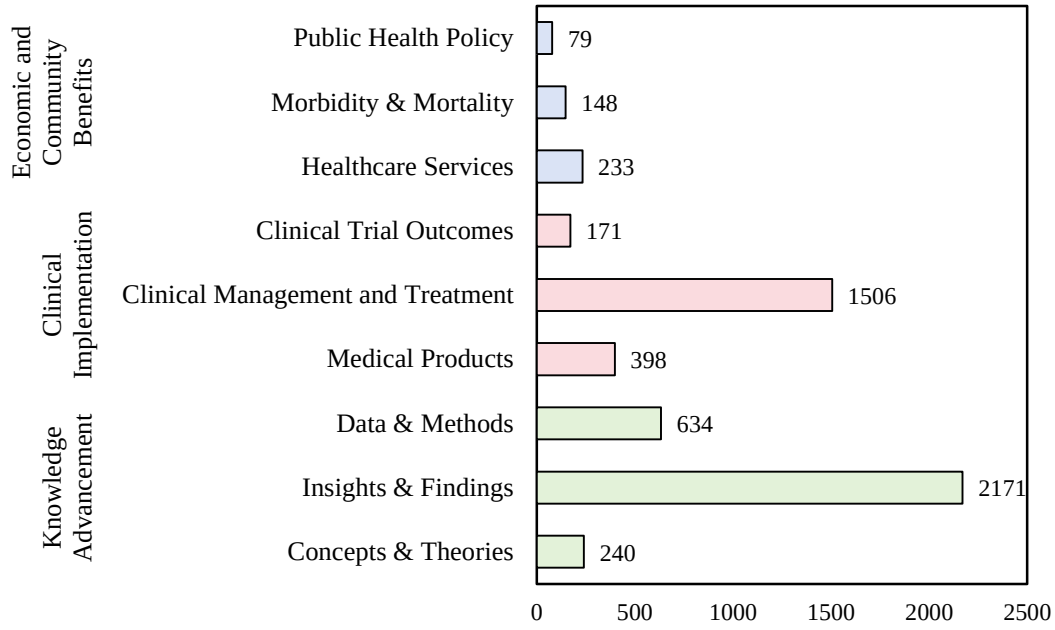

Figure 5. Distribution of types of scientific research contributions.

To better understand how scientific research contributions vary across papers on different topics, further exploration of their distribution is necessary (Figure 5). Analogous to the overall situation, it is observed that contributions in terms of insight discovery, clinical management and treatment, as well as data and methods dominate across all three topics. Slight variations exist among these topics. Specifically, contributions related to clinical trial outcomes and healthcare services are more pronounced in *Cardiovascular Diseases* compared with the other two topics, whereas conceptual and theoretical contributions are more evident in *Respiratory Tract Diseases*.

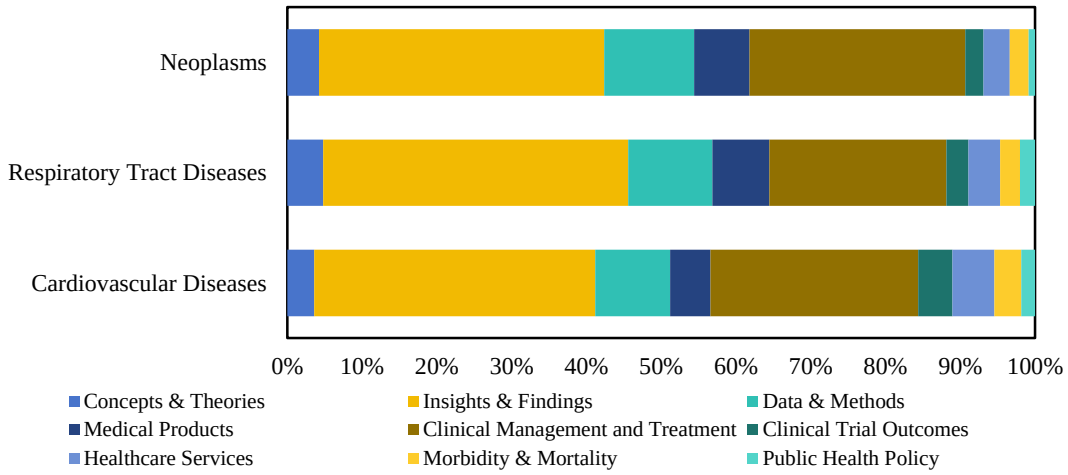

Figure 6. Distribution of the types of scientific research contributions of papers on different topics.

(2) Co-occurrence analysis between different types of scientific research contributions

Based on the overall distribution analysis of contribution types, to analyze the co-occurrence between different types of contributions helps gain a deeper understanding of the relationships or the influences between various contribution types. There are 2,145 papers demonstrating contributions in Knowledge Advancement, 1,669 papers exhibit contributions related to Clinical Implementation, and 413 papers present contributions in terms of Economic and Community Benefits. Notably, contributions of Knowledge Advancement and Clinical Implementation types tend to coexist more frequently, with 693 papers exhibiting both types of contributions. Following this, the coexistence of Clinical Implementation and Economic and Community Benefits contributions is observed in 241 papers. The coexistence of Knowledge Advancement and Economic and Community Benefits contributions is the least prevalent, occurring in 183 papers. Additionally, 97 papers exhibit contributions across all three types.

It is demonstrated that breakthroughs in basic research often propel advancements in clinical practice. This close connection may stem from the trend in modern medical research. Modern medical research emphasizes the rapid translation of basic research into clinical applications, which is driven by the need to meet the demands of medical practice. Knowledge Advancement and Economic and Community Benefits, the two types of contributions, often do not occur simultaneously. Knowledge Advancement typically involves basic research and theoretical innovation, with a primary focus on the academic sphere. In contrast, economic and community benefits are often derived from applied research. There is a gap between basic research and the generation of

significant economic and societal benefits. Additionally, there are inherent differences in research goals between basic and applied research. These disparities in goals cause scientists to concentrate more on a single area when conducting scientific research, which reduces the possibility of making both kinds of contributions in one paper.

In terms of more specific contributions, the majority (64.20%) of the nine scientific research contributions appear independently. There are 899 papers (28.03%) that demonstrate distinct scientific research contributions simultaneously. At most, six types of contributions appear simultaneously, but there is only one such paper. This shows that a study usually focuses on a single aspect to make outstanding contributions, and multiple contributions are less likely to occur at the same time. Table 9 shows the pairwise co-occurrence of scientific research contributions. The numbers in the table represent the count of papers in which contributions co-occur, indicating how many papers possess both contributions simultaneously. A darker shade in a cell signifies a greater intensity of co-occurrence of contributions. Among them, the most frequently co-occurring scientific research contributions are "Insights & Findings" and "Clinical Management and Treatment" (436), followed by "Insights & Findings" and "Data & Methods" (197), "Data & Methods" and "Clinical Management and Treatment" (143), as well as "Clinical Management and Treatment" and "Healthcare Services" (125). This implies that papers are more likely to generate other kinds of contributions when they contribute to the fields of idea creation, data methodologies, or clinical management and therapy. "Public Health Policy" contribution occurs infrequently with other kinds of contributions; in other words, public health policy is a relatively independent contribution.

Table 9. Co-occurrence of types of scientific research contributions.

Contribution type	1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3
1.1	0	85	37	17	57	7	8	6	3
1.2	85	0	197	106	436	28	62	56	20
1.3	37	197	0	43	143	16	27	15	15
2.1	17	106	43	0	55	9	27	15	5
2.2	57	436	143	55	0	38	125	65	20
2.3	7	28	16	9	38	0	8	10	6
3.1	8	62	27	27	125	8	0	13	5
3.2	6	56	15	15	65	10	13	0	5
3.3	3	20	15	5	20	6	5	5	0

High RStar - Low Citation papers and Low RStar - High Citation papers

The analysis results show a significant positive correlation between RStar and Citation of papers recommended by H1 Connect. This section explores some exceptions underlying this correlation. We define papers with the RStar in the top 10% but Citation in the bottom 10% as "High RStar – Low Citation papers (HR - LC)," totaling 39. Papers with RStar in the bottom 10% but Citation in the top 10% are termed "Low RStar – High Citation papers (LR - HC)," amounting to 293. An analysis focused on these two groups of papers, covering RStar, Citation, AAS, publication years, paper types, and other relevant attributes, is conducted to preliminarily identify characteristics of papers where the level of reviewer recognition significantly differs from Citation. The results are presented in Figure 7. As shown in Figure 7(a), the topic distribution of the two specific sub-datasets is similar to that of the overall dataset, with *Neoplasms* having the largest proportion of papers and *Respiratory Tract Diseases* having the least. From the perspective of document types, as shown in Figure 7(b), Article is the main type, and it accounts for a larger proportion of the HR - LC papers. In terms of publication year, as shown in Figure 7(c), there is a big difference between the two specific sub-datasets, with HR - LC papers being published later, mostly in 2019 and 2020, while LR - HC papers are published earlier, since paper citations take time to accumulate. With a notable separation between the two, Figure 7(d) shows the distribution of AAS and Citation for HR - LC papers and LR -HC papers. HR -LC papers have an average AAS of just 5.25, with AAS values ranging from 1 to 48.1. LR -HC papers, on the other hand, have an average AAS of 629.15, and their AAS values range between 5.25 and 10528.266. This means that compared to papers with greater RStar, those with higher citations typically garner more social attention. The social attention received by LR - HC papers is significantly higher than that received by HR - LC papers.

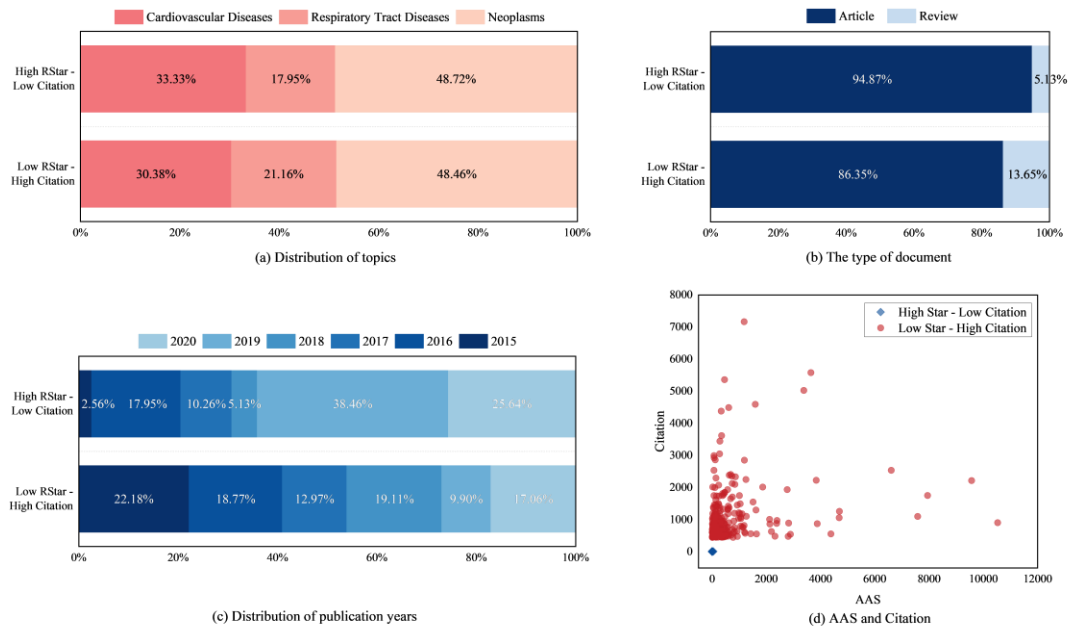

Figure 7. Characteristics of specific sub-datasets.

In the HR - LC and LR - HC papers, 21 papers (53.85%) and 57 papers (19.45%) respectively contained explicit scientific research contribution statements in their open peer review texts. The specific distribution is shown in Table 10. Similar to the overall distribution of scientific research contributions, "Insights & Findings" and "Clinical Management and Treatment" are the most common types of contributions. In the LR - HC papers, only these two types account for more than 10%, making them the dominant contributions. For the HR - LC papers, in addition to these two types, contributions related to "Data & Methods" are also notable. The scientific research contributions of HR - LC papers are concentrated in specific areas, with no contributions related to "Clinical Trial Outcomes," "Healthcare Services," or "Public Health Policy." In contrast, LR - HC papers address contributions across all nine types. This suggests that papers with greater contributions to economic and community benefits tend to receive higher Citation and lower RStar, while papers focused more on theoretical innovation and clinical applications, with fewer contributions to economic and community benefits, often receive higher RStar but lower Citation.

Table 10. Distribution of scientific research contributions of special sub-datasets.

	1.1	*1.2*	*1.3*	*2.1*	*2.2*	*2.3*	*3.1*	*3.2*	*3.3*
High RStar - Low Citation	5.56 %	25.00 %	22.22 %	8.33 %	36.11 %	0.00 %	0.00 %	2.78 %	0.00 %
Low RStar - High Citation	6.25 %	46.25 %	2.50%	5.00 %	26.25 %	2.50 %	5.00 %	1.25 %	5.00 %

Discussion & conclusion

This study analyzed peer review data and impact indicators of papers on *Cardiovascular Diseases*, *Respiratory Tract Diseases*, and *Neoplasms*, revealing significant differences in the distribution of relevant indicators among different topics. At the same time, a significant correlation between peer review data and impact indicators was verified. Additionally, the study found that the scientific research contribution types of the papers exhibited clustering. The validity and reliability of open peer review data have been somewhat confirmed by this study, which also offers helpful references for better application of peer review data in academic evaluation practice. The results of peer review judge the value of a paper from the perspective of experts, while traditional citation and altmetrics consider the quality and influence of a paper from the perspective of scholars and the public. These indicators all play an important role in scientific evaluation. These three evaluation methods complement each other and together provide a strong basis for the evaluation of scientific research outcomes.

In terms of topic differences, this study conducted a statistical analysis of papers on *Cardiovascular Diseases*, *Respiratory Tract Diseases*, and *Neoplasms*. The analysis revealed differences in RNumber, RStar, Citation, and AAS among the papers in these three topics, indicating that the performance of papers across different evaluation perspectives is influenced by the research topic. Further pairwise comparisons of the topics revealed that there were statistically significant differences between some topics ($P<0.05$), which highlights the need to consider the characteristics and priorities of different research fields when establishing the scientific research evaluation system. For example, *Cardiovascular Diseases* may focus more on clinical outcomes and the impact on healthcare services, while *Neoplasms* may be evaluated based on its contribution to drug development as well

as clinical management and treatment. By adopting differential evaluation criteria for specific topics, with each topic being assessed based on its unique aspects, the academic evaluation system can more accurately capture the true contribution and value of research in different fields.

In terms of the relationship between indicators, this study conducted Spearman correlation analysis to explore the relationship between open peer review indicators and impact indicators. The results showed significant positive correlations among RNumber, RStar, Citation, and AAS. Peer review indicators, along with Citation and AAS evaluate scientific research from different perspectives, with varying emphases. This suggests that peer review data and impact indicators should complement each other in research evaluation. The consistency observed also indicates that peer review is an effective scientific evaluation method. Furthermore, compared to the delayed nature of citation, open peer review can help predict a paper's impact and identify valuable research with greater potential for academic and social impact after publication.

In terms of scientific research contributions, the study found that most papers recommended by H1 Connect tend to focus more prominently on one specific area of contribution, and the probability of multiple contributions occurring is relatively low.. The findings also reveal that while there are slight differences in the distribution of contribution types among the three topics, most research papers primarily focus on advancing insights and findings or contributing to clinical management and treatment. This indicates that in the field of biomedicine, academic research plays a crucial role not only in advancing the boundaries of disciplines and expanding knowledge systems, but also in optimizing clinical decision-making, improving treatment strategies, and ultimately enhancing public health outcomes. Another important finding is that when papers have contributions in viewpoint discovery, data methods, or clinical management and treatment, they are more likely to trigger other types of contributions, while contributions related to public health policy less frequently co-occur with other types of contributions. This suggests that there remains a gap between biomedical research findings and the translation into policy. Papers in the biomedical field often focus on theoretical innovation, technological breakthroughs, or clinical applications, typically centered on specific diseases. The development of public health policies requires not only scientific evidence but also a comprehensive consideration of factors such as implementation challenges, economic costs, and other multifaceted aspects. Consequently, this highlights the need for a more comprehensive approach to evaluating scientific research, one that accounts for the diversity of contributions across different research areas. Rather than

relying solely on a single indicator, academic evaluation should incorporate multiple indicators to reflect a paper's contributions across various dimensions.

In summary, the results of this study highlight the value of peer review in academic evaluation. In practice, it is crucial to recognize the multiple contributions research can make and consider the unique characteristics of different research fields. A more comprehensive and diversified academic evaluation system, which should include impact indicators and peer review data, will better capture the multifaceted nature of scientific contributions. As research fields continue to evolve and become increasingly specialized, the evaluation system must adapt to ensure that it accurately reflects the diversity and influence of scientific work. Thus, it can promote a more open and comprehensive academic evaluation process.

There are also some limitations in this study. Due to the characteristics of the H1 Connect platform, the data samples selected in this study belong to the field of biomedicine, and there may be differences between different topics. In the future, the scope of the research can be further expanded to other fields to validate the generalizability of the conclusions. In addition, the peer review process is affected by multiple factors, such as the reviewer's research interests. In the future, other dimensions can be supplemented to explore the differences in peer review behavior under the influence of multiple factors.

References

Bravo, G., Farjam, M., Grimaldo Moreno, F., Birukou, A., & Squazzoni, F. (2018). Hidden connections: Network effects on editorial decisions in four computer science journals. *Journal of Informetrics*, *12*(1), 101–112.

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., & Squazzoni, F. (2019). The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications*, *10*(1), 322.

Chen, C., Cheng, Z., Wang, C., & Li, L. (2023). Identification and utilization of key points of scientific papers based on peer review texts. *Journal of the China Society for Scientific andTechnical Information*, *42*(5), 562–574.

Cheng, X., Wang, H., Tang, L., Jiang, W., Zhou, M., & Wang, G. (2024). Open peer review correlates with altmetrics but not with citations: Evidence from Nature Communications and PLoS One. *Journal of Informetrics*, *18*(3), 101540.

Demarest, B., Freeman, G., & Sugimoto, C. R. (2014). The reviewer in the mirror: Examining gendered and ethnicized notions of reciprocity in peer review. *Scientometrics*, *101*(1), 717–735.

Fox, C. W., & Paine, C. E. T. (2019). Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecology and Evolution*, *9*(6), 3599–3619.

Fromm, M., Faerman, E., Berrendorf, M., Bhargava, S., Qi, R., Zhang, Y., Dennert, L., Selle, S., Mao, Y., & Seidl, T. (2021). Argument Mining Driven Analysis of Peer-Reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(6), Article 6.

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, *17*(1), e0259238.

Ghosal, T., Verma, R., Ekbal, A., & Bhattacharyya, P. (2020). DeepSentipeer: Harnessing sentiment in review texts to recommend peer review decisions. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*,1120–1130.

Han, R., Zhou, H., Zhong, J., & Zhang, C. (2022). Characterizing Peer Review Comments of Academic Articles in Multiple Rounds. *Proceedings of the Association for Information Science and Technology*, *59*(1), 89–99.

Hua, X., Nikolov, M., Badugu, N., & Wang, L. (2019). *Argument mining for understanding peer reviews*. *1*, 2131–2137.

Lin, Y., Wang K., Ding K., & Xu, K. (2021). Quantitative research on qualitative evaluation of academic papers. *Information Studies:Theory & Application*, *44*(8), 28–34.

Luo, Z., Cai, L., Qian, J., & Lu, W. (2021). Research on the recognition of innovative contribution sentences of academic papers. *Library and Information Service*, *65*(12), 93–100.

Ni, J., Zhao, Z., Shao, Y., Liu, S., Li, W., Zhuang, Y., Qu, J., Cao, Y., Lian, N., & Li, J. (2021). The influence of opening up peer review on the citations of journal articles. *Scientometrics*, *126*(12), 9393–9404.

Qin, C. (2020). Exploring the distribution regularities of referees' comments in IMRAD structure of academic articles. *18th International Conference on Scientometrics and Informetrics, ISSI 2021*, 1527–1528.

Shi, Z., Zhu L., & Le, X. (2024). Material Information Extraction Based on Local Large Language Model and Prompt Engineering. *Data Analysis and Knowledge Discovery*, *8*(7), 23–31.

Wang, K., & Wan, X. (2018). Sentiment Analysis of Peer Review Texts for Scholarly Papers. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 175–184.

Wang, P., Hoyt, J., Pöschl, U., Wolfram, D., Ingwersen, P., Smith, R., & Bates, M. (2016). The last frontier in open science: Will open peer review transform scientific and scholarly publishing? *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–4.

Wang, P., Williams, J., Zhang, N., & Wu, Q. (2020). F1000Prime recommended articles and their citations: An exploratory study of four journals. *Scientometrics*, *122*(2), 933–955.

Wang, H. (2023). Value, mechanism and strategies of open peer review for periodicals. *Acta Editologica*, *35*(2), 147–151.

Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). *Scientometrics*, *125*(2), 1033–1051.

Xie, W., Jia, P., Zhang, G., & Wang, X. (2024). Are reviewer scores consistent with citations? *Scientometrics*, *129*(8), 4721–4740.

Zhang, G., Wang, L., Xie, W., Shang, F., Xia, X., Jiang, C., & Wang, X. (2021). "This article is interesting, however": Exploring the language use in the peer review comment of articles published in the BMJ. *Aslib Journal of Information Management*, *74*(3), 399–416.

Zong, Q., Fan, L., Xie, Y., & Huang, J. (2020). The relationship of polarity of post-publication peer review to citation count: Evidence from Publons. *Online Information Review*, *44*(3), 583–602.

Zong, Q., Xie, Y., & Liang, J. (2020). Does open peer review improve citation count? Evidence from a propensity score matching analysis of PeerJ. *Scientometrics*, *125*(1), 607–623.