Original Paper

Detecting Invisible Triangular Citations in Biomedical Research Using a Dual-Threshold Framework Based on Textual Similarity and Bibliographic Coupling Strength: Method Development and Validation Study

Yunmei Liu[1,2], PhD; Jingyi Xu[1], MSc; Lei Huang[3,4], PhD, MD

[1]School of Cultural Heritage and Information Management, Shanghai University, Shanghai, China

[2]National Experimental Base for Intelligent Evaluation and Governance, Fudan University, Shanghai, China

[3]National Key Laboratory of Immunity and Inflammation, Changhai Clinical Research Unit, The First Affiliated Hospital of Naval Medical University/Changhai Hospital, Naval Medical University, Shanghai, China

[4]Department of Gastroenterology, National Clinical Research Center for Digestive Diseases, The First Affiliated Hospital of Naval Medical University/Changhai Hospital, Naval Medical University, Shanghai, China

Corresponding Author:

Lei Huang, PhD, MD
National Key Laboratory of Immunity and Inflammation, Changhai Clinical Research Unit
The First Affiliated Hospital of Naval Medical University/Changhai Hospital, Naval Medical University
168 Changhai Road
Shanghai
China
Phone: 86 021-31166666
Email: lei.huang@alumni.dkfz.de

Abstract

Background: Invisible triangular citations (ITCs)—a form of citation malpractice where authors of a follower document (C) derive content from a seminal work (A) through an unacknowledged intermediary (B)—pose a significant threat to knowledge provenance and academic integrity. This concealed practice is particularly hazardous in biomedical research, where distorted attribution can propagate misinterpreted concepts into clinical guidelines and therapeutic protocols. Current detection methods lack robust, content-based models to identify such obscured citation pathways.

Objective: The study aims to establish and empirically validate a dual-threshold framework for detecting ITCs in biomedical research by integrating textual similarity and bibliographic coupling strength and to characterize the positional and semantic patterns of these concealed citations.

Methods: Using Web of Science's enriched cited references technology, we conducted a full-text citation analysis of 2948 source papers, 20,801 citing documents, and 37,542 citation sentences from the antibody-drug conjugate domain—a rapidly growing area of anticancer drug development, particularly in oncology. The framework combined textual similarity (cosine similarity) with bibliographic coupling strength (shared references). Across more than 2 million pairwise records, threshold selection and robustness were evaluated using Spearman rank correlation, sliding-window analysis, segmented linear regression, Bayesian information criterion improvement, enrichment factors, chi-square tests, Cramér *V*, threshold perturbation analysis, and expert review.

Results: Optimal detection thresholds were identified as cosine similarity ≥0.60 and coupling strength ≥10, beyond which citation behavior was more likely to show inconsistency between the formal citation and the substantive knowledge source. Among 2,085,004 effective pairwise records, 6165 ITC records were detected, accounting for 0.296% of all records. In a stratified random sample of 100 records independently and blindly reviewed by 3 experts, the expert confirmation rate was 85.00% (Wilson 95% CI 76.72%-90.69%). Positional analysis showed that ITCs (n=12,274) were concentrated in the Introduction (n=6023, 49.1%) and Discussion (n=5833, 47.5%), whereas Results (n=255, 2.1%) and Methods (n=163, 1.3%) accounted for much smaller proportions. Citation position was statistically associated with both cosine similarity and bibliographic coupling strength, although the effect sizes were weak. Representative cases showed that ITCs may involve not only verbatim copying but also strategic rewriting through lexical substitution and syntactic restructuring.

Conclusions: This empirically validated dual-threshold framework provides a transferable model for ITC detection across biomedical disciplines. Cosine similarity and bibliographic coupling strength were significantly but weakly positively correlated (Spearman ϱ=0.065, 95% CI 0.064-0.066; *P*<.001), supporting a dual-threshold rather than a single-indicator model and suggesting that authors replicate reference structures while paraphrasing text. The vulnerability of the Introduction and Discussion sections highlights systemic weaknesses in knowledge transmission, with implications for clinical practice. This study pioneers the application of enriched cited references technology to safeguard attributional integrity and offers actionable governance strategies for research integrity.

JMIR Med Inform 2026;14:e96265; doi: 10.2196/96265

Keywords: invisible triangular citation; citation position; textual similarity; cosine similarity; coupling strength; biomedical research; citation malpractice; Enriched Cited References; ECR; ECR technology

Introduction

Background

In the progression of academic research, citations play an indispensable role in providing theoretical foundations and empirical support for research outcomes. They not only reflect the inheritance and development of scholarly achievements but also serve as crucial indicators for measuring research value and impact. Triadic citation, as a distinctive citation phenomenon, has gradually captured researchers' attention. Liu et al [1] pioneered the study of triadic citation behavior, introducing concepts such as the original document (A), intermediary document (B), and follower document (C). Through data analysis, they demonstrated the existence of triadic citations in real-world citation networks with significant coverage. Among these, invisible triangular citation (ITC)—a specialized form of triadic citation—manifests when authors of the follower document (C) acquire information about the original document (A) via the intermediary document (B) yet cite only the original document (A) without acknowledging the intermediary (B) in their actual references [2].

ITCs, an improper citation practice, pose significant hazards by potentially misleading researchers' judgments about the genuine quality and value of scholarly publications. For instance, certain documents may establish and propagate unverified or erroneous scientific claims by distorting the original meaning of source literature, thereby creating unreliable information chains that confer "unwarranted authority" on the original document (A) [3]. This phenomenon occurs when document (C) cites the ostensibly authoritative document (A) through the intermediary document (B) without acknowledging (B), thereby obscuring the traceability of citation origins. Consequently, erroneous conclusions may propagate through citation networks, while remaining undetectable within visible citation chains. This concealment heightens the difficulty of identification by journal review systems.

Concurrently, improper citation practices persistently occur within the biomedical domain. Peoples et al [4] found that up to 25% of the scientific literature in biomedical disciplines contains citation inaccuracies, resulting in the dissemination of erroneous or misinterpreted data. Research further reveals that citation error rates in biomedical literature range between 20% and 25%, posing significant risks such as amplifying misinformation and compromising research validity [5]. These statistics underscore the pervasive nature of improper citations in the biomedical field. Such practices may lead to the development of clinical guidelines or policies based on flawed evidence, thereby misleading physicians' diagnostic and therapeutic decisions. This ultimately poses potential threats to patient health and safety.

Full-text citation content analysis, as an analytical approach to addressing improper citation practices, overcomes the limitations of traditional citation analysis that primarily relies on frequency counts by examining specific contextual occurrences of citations [6]. This methodology enables deeper comprehension of interrelationships among scientific publications and provides robust support for identifying ITCs. By using direct citation texts and semantic information between documents as research data [7], this facilitates citation context analysis through semantic similarity assessment [8] and enhances analytical accuracy by incorporating textual features such as citation positioning [9]. Web of Science launched its Enriched Cited References (ECR) technology in April 2021, with in-depth indexing information currently available only for papers conforming to the IMRAD (Introduction, Methods, Results, and Discussion) structure. This technology uses machine learning techniques to perform deep semantic analysis of citation content, with its core functionality focusing on automatically identifying citation sentences and classifying them according to the rhetorical function of cited references within citing documents [10-12]. Given the novelty of ECR functionality and the scarcity of related research [13], this study leverages this technology to conduct comprehensive full-text citation content analysis.

The theoretical innovation of this study lies in establishing a detection framework for ITCs using dual indicators—textual similarity and coupling strength—while conducting feature analysis from the perspective of citation position. Practically, it pioneers the construction of an ITC identification model specific to the antibody-drug conjugate (ADC) oncology domain, which provides a transferable framework for other medical fields. Additionally, by leveraging ECR technology to collect data on citation positions and citation sentences, this research broadens the practical implementation scenarios of ECR.

Literature Review

ITC is a composite form of citation malpractice that can be regarded as the intersection of omission (failing to cite original sources) and secondary citation (reproducing citations from intermediary sources without consulting the original publications) [14]. It combines the characteristics of both practices and poses new research challenges due to its concealed nature. Current research on citation malpractice primarily focuses on identification methods, manifestations, causal analysis, and governance strategies. As a specific subtype, studies on ITC similarly address these dimensions, though targeted research remains limited.

In the context of identification methods, early studies primarily used computer technology for the automatic detection of formatting irregularities [15-17] yet lacked robust identification models for nonnormative citation content behaviors. More recently, research has moved from format checking toward citation content analysis. Manual verification studies have compared citing statements with cited originals to determine whether a cited proposition is supported by the original publication [18] and have further distinguished inaccurate citations involving nonexistent findings, misinterpretation of conclusions, numerical errors, contextual errors, and content transferred from other sources [19]. Recent scholarly work has proposed innovative approaches integrating triadic citation structure analysis, textual similarity computation, and citation coupling strength measurement [20]. By combining textual similarity and citation coupling strength to identify potential instances of citation omission, researchers have provided new methods for analyzing content use relationships and citation associations between publications [21].

Regarding manifestation characteristics, previous studies have analyzed manifestation patterns across external dimensions such as bibliographic attributes (eg, language and document type), content features (eg, disciplinary domain), and author characteristics (eg, author's H-index) [21]. Additionally, scholars have developed multidimensional evaluation indicator systems specifically for ITCs, incorporating metrics such as citation content similarity, usage-to-citation conversion ratio, and coupling strength. These systems correlate document features with citation content to investigate contextual occurrence patterns and influencing factors, including journal impact, inherent citation influence, and publication time disparity [2].

Furthermore, analyzing the causes of citation malpractice forms the foundation for its governance. Regarding causal analysis and governance strategies, some scholars have previously examined forms of misconduct such as citation omission and secondary citation, proposing corresponding improvement recommendations [22-24]. Related research has also summarized the causes of citation malpractice as including authors' lack of scholarly rigor, failure to establish sound citation practices, and insufficient understanding of academic norms. Recent research has further found that inaccurate citations may propagate along citation chains. In a study of frequently cited biomedical papers, approximately one-fifth of inaccurate citations were related to existing inaccurate citation chains [19]. This finding suggests that when authors do not fully verify the original literature but reuse citation statements from other papers or rely on intermediary sources for the original information, early citation errors, interpretive deviations, or source omissions may accumulate in subsequent literature. ITCs are related to this chain-propagation pathway but are not identical to general citation error. General citation error focuses on whether the citing text is consistent with the cited publication, whereas ITCs also involve the relationship between the cited content and the knowledge source, unacknowledged intermediary contributions, and opaque attribution of knowledge provenance.

Corresponding governance measures have been proposed from the perspectives of various stakeholders, including authors, academic journals, affiliated institutions, and industry regulatory bodies [25]. However, these measures still mainly target explicit citation errors, formatting irregularities, and singular forms of citation malpractice, such as omission and secondary citation. Existing research has not yet sufficiently revealed the specific causes of the obscured intermediary knowledge pathways in ITCs nor has it provided targeted governance strategies corresponding to their composite citation structures.

In summary, current research on ITC exhibits the following limitations. First, regarding identification methods, there is a lack of ITC identification models established from the dimension of citation content. Second, the analysis of manifestation characteristics predominantly focuses on external features of documents, with insufficient exploration of the internal characteristics of citations. Finally, current causal analyses and governance strategies primarily address traditional, singular forms of citation malpractice, lacking research on composite citation malpractice. Consequently, this study will use a dual-metric approach using textual similarity and coupling strength to establish an identification model for ITC based on citation sentences. It will analyze the internal characteristics of citation positioning and conduct tailored research into the causes and governance of ITC behavior.

Aims

This study aimed to establish and empirically validate a dual-threshold framework combining cosine similarity and bibliographic coupling strength for identifying ITCs in biomedical research. Specifically, based on large-scale, full-text citation data from the ADC oncology research domain, this study aimed to (1) determine the identification thresholds for textual similarity and bibliographic coupling strength; (2) evaluate the robustness and interpretability of the framework through threshold sensitivity analysis, expert review, and representative case analysis; and (3) compare the distribution of ITCs across the IMRAD sections to characterize their positional features.

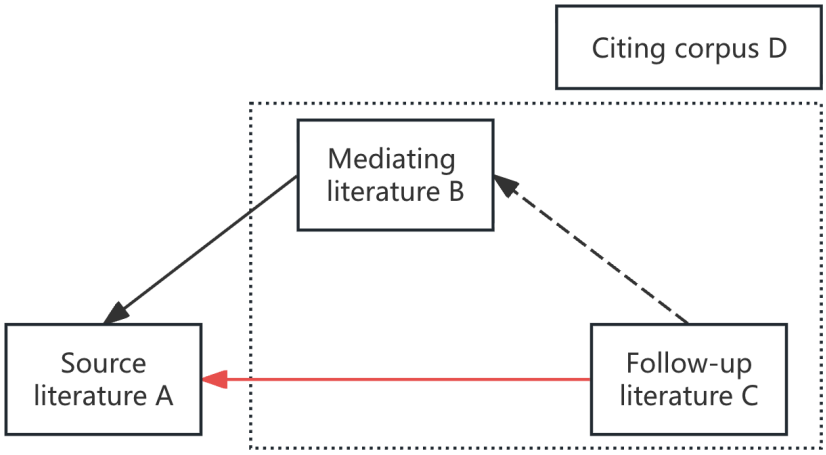
Methods

Data Source and Preprocessing

Concept Model

Triadic citation tracing can be categorized into 3 types: forward, reverse, and bidirectional [20]. Within this framework, the forward triadic citation model is based on the citation chain originating from the source publication (A). By tracing citations forward, this model identifies intermediary publications (B) and follower publications (C), enabling efficient tracking of diffusion pathways in emerging fields. As illustrated in Figure 1, ITC constitutes a specific form of triadic citation behavior. It occurs when authors of follower publications (C) access information about the source publication (A) indirectly through a mediating literature (B) yet cite only the source (A) without acknowledging the intermediary (B) in their actual citations [2]. To facilitate analysis, this study consolidates intermediary publications (B) and follower publications (C) into the citing publications set D (excluding data representing direct citation relationships between B and C). The focus is on the direct citation relationship between the source publication (A) and the citing publications set D, formally defined as $D = \{B_1, B_2,..., B_m\} \cup \{C_1, C_2,..., C_n\}$.

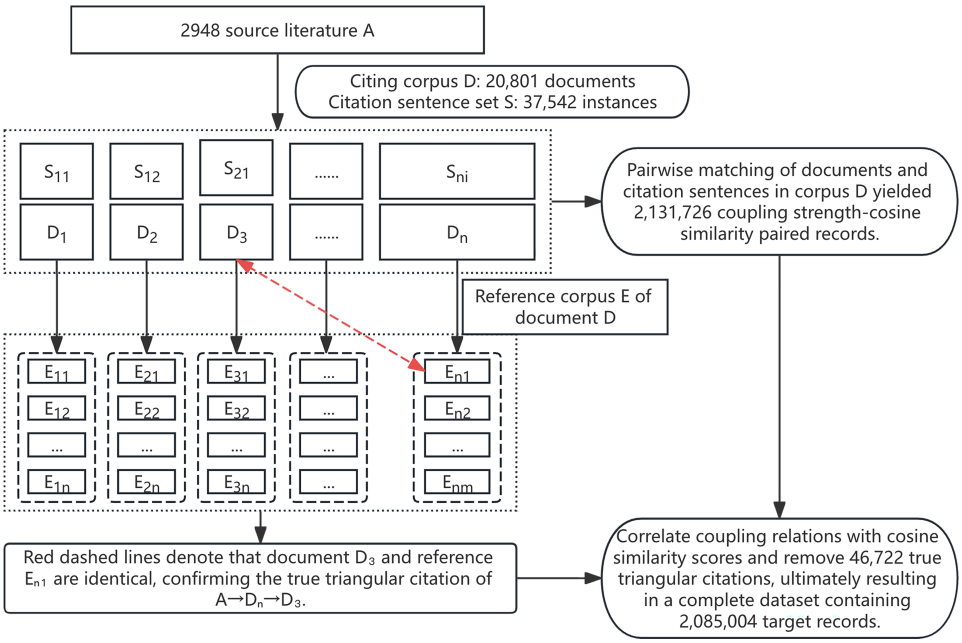
Figure 1. Conceptual model of invisible triangular citations. Dashed arrows denote unattributed citations. Red arrows indicate secondary citations, where sources were cited without consulting the original work.



Data Source and Processing Flow

The complete data processing flow is shown in Figure 2.

Figure 2. Data processing workflow.



Definition of the Original Literature: A Dataset

ADCs deliver cytotoxic drugs to tumor cells through monoclonal antibodies and represent an important research direction in precision oncology, characterized by rapid clinical translation, active technological iteration, and rapid growth of the literature [26-28]. Because knowledge in this field updates rapidly and publications are densely connected, researchers may obtain original information through intermediary literature, increasing the risk of ITCs. Therefore, this study uses the forward triadic citation model to investigate the ADC domain, enabling rapid identification of a substantial volume of subsequent extension research. The Web of Science Core Collection served as the data source, covering literature published from 2020 to May 2026. To precisely pinpoint core literature on ADCs within the cancer field, a compound retrieval strategy was implemented for literature collection. The search query was designed as follows:

topic field=(((“antibody-drug conjugate*” OR “antibody drug conjugate*” OR ADC OR ADCs OR “immune conjugate*” OR “immunoconjugate*”) AND (cancer* OR tumor* OR tumor* OR malignan* OR carcin* OR oncolog* OR neoplas* OR metastasa* OR adenocarcinoma* OR carcinoma* OR sarcoma* OR melanoma* OR lymphoma* OR leukemia* OR leukemia* OR myeloma* OR “breast cancer*” OR “lung cancer*” OR “gastric cancer*” OR “colorectal cancer*” OR “ovarian cancer*” OR “bladder cancer*” OR “urothelial carcinoma*” OR “pancreatic cancer*” OR “liver cancer*” OR “hepatocellular carcinoma*” OR “prostate cancer*” OR “cervical cancer*” OR “endometrial cancer*” OR “renal cancer*” OR “kidney cancer*” OR “head and neck cancer*” OR “glioblastoma*” OR “acute myeloid leukemia*” OR “acute lymphoblastic leukemia*” OR “chronic lymphocytic leukemia*”)))

In the topic field, “antibody-drug conjugate” and its common abbreviation were used to locate the core research area and were combined with other malignancy neoplasm–related terms. Data collection was conducted from May 6, 2026, to June 2, 2026, yielding 5436 publications.

First, as establishing triadic citation relationships requires the source publication (A) to have at least 2 citing publications, only publications with citation counts of at least 2 were retained (3836 publications). Subsequently, to further align with the research requirements for identifying ITCs under the ECR framework, data points in which a single publication had an ECR data volume ≤1 and retracted publications were excluded, focusing the analysis on valid citation chains. Through this screening process, a final collection of 2948 source publications (set A) was established, forming the foundational dataset for subsequent analysis.

Constructing a Collection of Cited Literature and Quoted Sentences

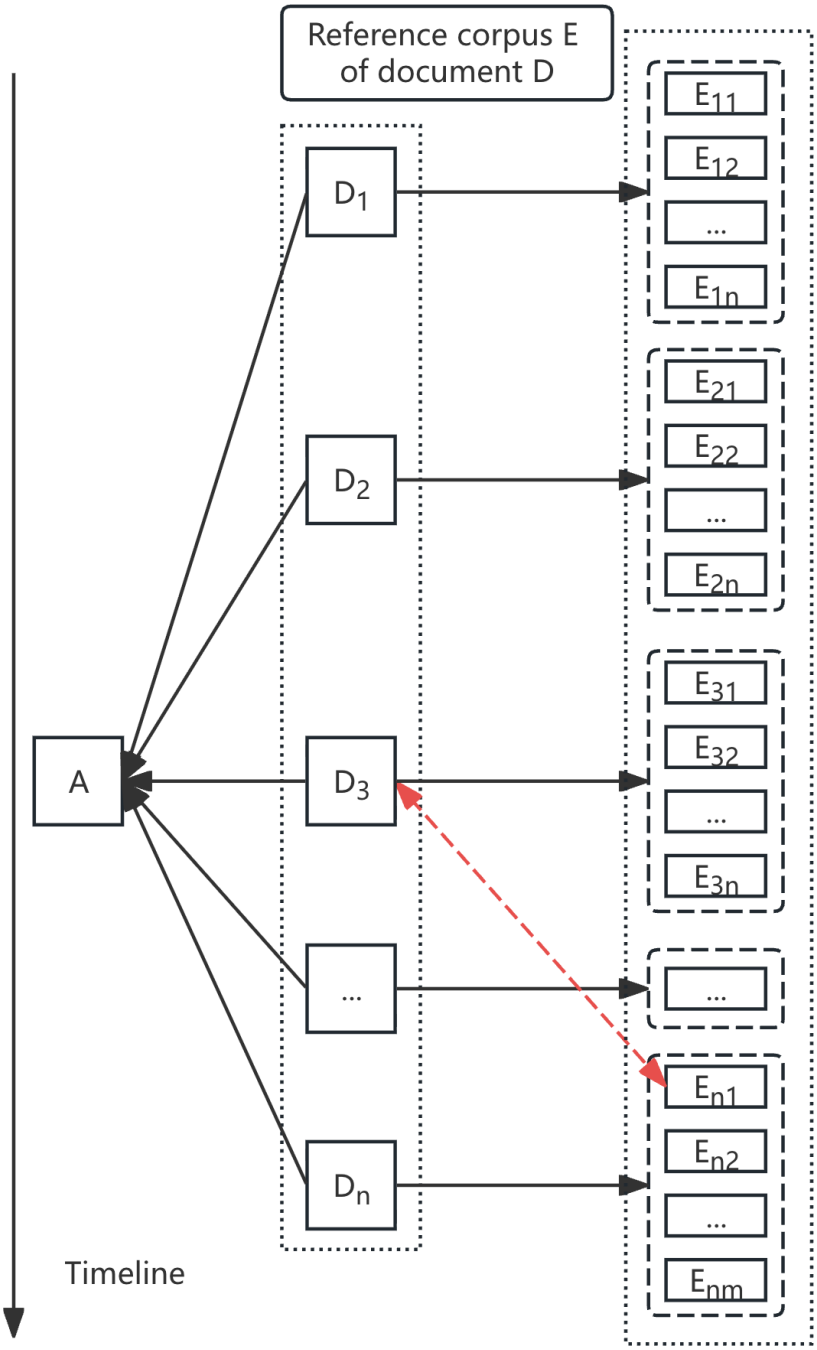
Starting from the original document collection (A), the ECR function was used to extract all citing documents along with the citation sentences from these citing documents referencing the original documents (A), thereby constructing the citing document collection $D=\{D_1, D_2,\ldots D_n\}$ and the citation sentence collection $S=\{S_{11}, S_{12},\ldots S_{ni}\}$, where D_n represents the nth citing document of the original document (A) and S_{ni} represents the *i*th citation sentence within D_n referencing the original document (A; each citing document contains at least one valid citation sentence). Ultimately, based on 2948 original papers (A), a total of 20,801 citing documents (D) and 37,542 citation sentence instances were collected. As shown in Table S1 in Multimedia Appendix 1, 99.9% of source publications (A) were papers; the distributions of citation functions and citation positions are shown in Tables S2 and S3 in Multimedia Appendix 1.

Collection of Invisible Triangular Reference Datasets

To address the research objective concerning ITCs, it is necessary to genuinely exclude existing triangular citation relationships between documents. The specific cleaning procedure is illustrated in Figure 3: first, extract the reference set $E=\{E_{n1}, E_{n2},\ldots E_{nm}\}$ for each citing document D_n, where E_{nm} denotes the *m*th reference in citing document D_n; subsequently, examine each reference E_{nm} to determine whether it coincides with any other document within the citing document collection D; if such coincidence is observed, the instance is identified as a genuine triangular citation and must be removed from the analytical dataset.

First, a pairwise analysis was conducted on the citing document collection $D=\{D_1, D_2,\ldots D_n\}$ to calculate the coupling relationships between documents. The coupling strength was defined as the number of shared references between 2 documents, determined by calculating the intersection of their reference sets for each pair. This process yielded a total of 846,650 coupling relationships along with their corresponding strength values. Second, a text similarity analysis was performed on the citation sentence collection $S=\{S_{11}, S_{12},\ldots S_{ni}\}$, calculating the cosine similarity between each pair of citation sentences. This study used an SBERT (sentence bidirectional encoder representations from transformers) sentence-vector approach by loading the pritamdeka/S-PubMedBert-MS-MARCO model through the sentence-transformers framework to encode the citation sentences and calculate cosine similarity between each pair of citation sentences. The bibliographic coupling strength and cosine similarity values were integrated, resulting in a combined dataset of 2,131,726 entries. Finally, by applying the cleaning rules established for genuine triangular citation relationships, 46,722 data entries corresponding to triangular citations were removed. This yielded a final effective pairwise dataset of 2,085,004 entries, establishing the data foundation for subsequent identification and analysis of ITCs.

Figure 3. Data cleaning process for true triangular citation relationships. Red-dashed lines indicate that document D_3 and reference En_1 are identical, confirming the true triangular citation of A→D_2→D_3.



To examine the distributional characteristics of the removed true triangular citation relationships, this study compared them with the retained data. Tables S4 and S5 in Multimedia Appendix 1 show that both types of data were mainly concentrated in background and discussion citation functions and in the Introduction and Discussion sections, but their distributional centers differed. Compared with the retained data, true triangular citation relationships were more concentrated in the Discussion, Methods, and Results sections and showed higher proportions of basis, support, and differ functions. These results indicate that the removed true triangular citation relationships were not random noise but explicit triadic citation pathways with clear positional and functional features. Therefore, removing this type of data before threshold identification was necessary to ensure a clear definition of the study object and to avoid interference from explicit triadic citations in threshold selection and subsequent analysis.

Ethical Considerations

This study analyzed only bibliographic and citation-context data from published literature in the Web of Science Core Collection. It did not involve human participants, human data, human tissue, patients, clinical interventions, or personal health information, and it did not involve animals, animal experimentation, or animal-derived biological materials. Therefore, it did not constitute human-participants research under 45 CFR §46.102(e) [29] or animal research under

the Animal Welfare Act [30] and the Public Health Service Policy on Humane Care and Use of Laboratory Animals [31]. Accordingly, no institutional review board/research ethics board or Institutional Animal Care and Use Committee application or approval was required.

Results

ITC Behavior Recognition

ECR Classification Framework

ECR technology uses machine learning methods to conduct in-depth analysis of citations, enabling the identification of the function and citation sentiment of the cited literature within the citing document [10]. The specific classification scheme is presented in Table S6 in Multimedia Appendix 1.

Correlation Between Bibliographic Coupling Strength and Cosine Similarity

This study calculated the Spearman rank correlation coefficient between bibliographic coupling strength and cosine similarity in the target dataset and found a significant but very weak positive correlation (Spearman ϱ=0.065, 95% CI 0.064-0.066; *P*<.001). This indicates that highly coupled publication pairs are more likely to be semantically similar, supporting the joint use of the 2 measures to identify ITCs. However, the correlation coefficient was low, indicating that the 2 indicators cannot substitute for one another. Bibliographic coupling strength mainly captures the shared knowledge base at the structural level, whereas cosine similarity captures semantic proximity at the citation sentence level; the 2 provide complementary information. Therefore, this study adopted dual-threshold joint screening rather than a single-indicator approach.

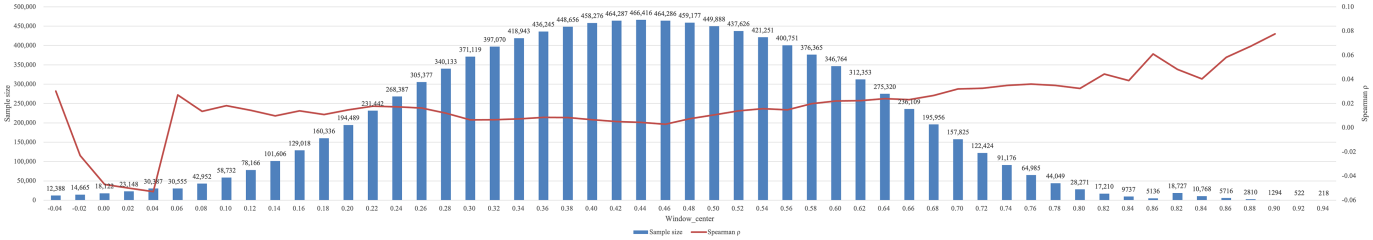
As shown in Table S7 in Multimedia Appendix 1, Spearman correlations between bibliographic coupling strength and cosine similarity were further calculated separately by citation functions and citation positions. Most strata were consistent with the overall results, showing significant but weak positive correlations. Only the Results position showed a nonsignificant and weak negative correlation. These findings indicate that the relationship between structural coupling and semantic similarity is affected by citation positions, so threshold setting should examine both the full sample and the stability across different citation functions and locations.

Cosine Similarity Threshold Setting

This study used a 2-step strategy to determine the cosine similarity threshold. First, sliding-window Spearman correlation analysis was used to examine the Spearman correlation between bibliographic coupling strength and cosine similarity across intervals of cosine similarity. A window width of 0.10 was used for the local observation interval, and the window moved rightward in steps of 0.02. For example, if one window was [0.40, 0.50], the next would be [0.42, 0.52]. Within each window, the Spearman correlation between log1*p*(coupling strength) and cosine similarity was calculated, and the window center was used as the x-axis value to plot the local correlation curve (Figure 4).

The sliding-window results showed that the local correlation between coupling strength and cosine similarity substantially fluctuated in low cosine similarity intervals, whereas the positive relationship gradually strengthened when cosine similarity entered the mid-to-high range. In particular, from 0.46 (ϱ=0.003, 95% CI 0.000-0.006; *P*=.048), the lower bound of the CI was nearly 0, indicating the end of the plateau and the beginning of a clearer strengthening of the positive relationship. Therefore, all intervals after SBERT ≥0.46 (a step size of 0.01) were included as candidate thresholds for segmented linear regression.

Figure 4. Sliding-window trends in Spearman correlations between bibliographic coupling strength and cosine similarity.



Second, segmented linear regression models were compared across candidate thresholds. The model was log1p(coupling strength)=β0+β1× cosine similarity+β2× max(0, cosine similarity-c), where c denotes the candidate threshold. Model comparison focused on the Bayesian information criterion (BIC), which evaluates both model fit and model complexity. A positive BIC improvement indicates that the segmented model has a lower BIC than the ordinary linear model, and a larger BIC improvement indicates stronger improvement after introducing the candidate threshold. As shown in Table 1, when cosine similarity=0.60, the BIC improvement reached the maximum value of 1344.33; the segmented model outperformed the linear model across all citation positions and functions; and β2 was significantly positive in all strata. Therefore, cosine similarity ≥0.60 was set as the textual similarity threshold.

Table 1. Segmented regression comparison of candidate thresholds for cosine similarity.

Candidate threshold C	BIC[a] improvement	Consistency: segmented BIC better than linear BIC	Consistency: β2 significantly positive
0.46	780.97	8/9	8/9
0.47	846.19	8/9	8/9
0.48	911.35	8/9	8/9
0.49	974.74	8/9	8/9
0.5	1034.82	8/9	8/9
0.51	1089.91	8/9	8/9
0.52	1138.55	8/9	8/9
0.53	1183.47	8/9	8/9
0.54	1223.38	8/9	8/9
0.55	1258.14	8/9	8/9
0.56	1287.44	8/9	8/9
0.57	1309.90	8/9	9/9[b]
0.58	1329.03	8/9	9/9
0.59	1341.81	8/9	9/9
0.60[b]	1344.33[b]	9/9[b]	9/9
0.61	1337.82	9/9	9/9
0.62	1323.04	9/9	9/9
0.63	1306.65	9/9	9/9
0.64	1288.99	9/9	9/9
0.65	1266.66	9/9	9/9
0.66	1238.27	9/9	9/9
0.67	1198.12	9/9	9/9
0.68	1148.83	9/9	9/9
0.69	1094.50	9/9	9/9
0.7	1046.04	8/9	9/9

[a]BIC: Bayesian information criterion.

[b]These values indicate the first threshold achieving full consistency (9/9) for the corresponding criterion. The threshold of 0.60 was selected because it produced the maximum BIC improvement (1344.33) and achieved full consistency across all strata.

Coupling Strength Threshold Setting

Bibliographic coupling strength showed a pronounced long-tailed distribution, with $k \leq 5$ covering 1,948,454 (93.45%) of the total sample 2,085,004, reflecting the common baseline pattern of citation behavior. As shown in Figure 5, the sample-size decay ratio (N_{k-1}/N_k) exhibited irregular oscillations and rebounds after $k \geq 15$, suggesting that the data were dominated by a very small number of cases and no longer supported stable statistical inference.

Therefore, k values from 6 to 14 were set as candidate thresholds, and the high-similarity enrichment factor (EF) was calculated as $p1/p0$, where $p1$ is the proportion of SBERT ≥0.60 in the qualified group (coupling≥ k), and $p0$ is the proportion of SBERT ≥0.60 in the unqualified group (coupling<k). As shown in Table 2, among the candidate thresholds, a coupling strength of 10 produced the highest EF for high similarity and the strongest discrimination. In addition, EF exceeded 1 across different citation functions and locations, indicating discriminative capacity. Therefore, 10 was selected as the bibliographic coupling strength threshold.

Figure 5. Changes in the sample-size decay ratio across levels of bibliographic coupling strength.

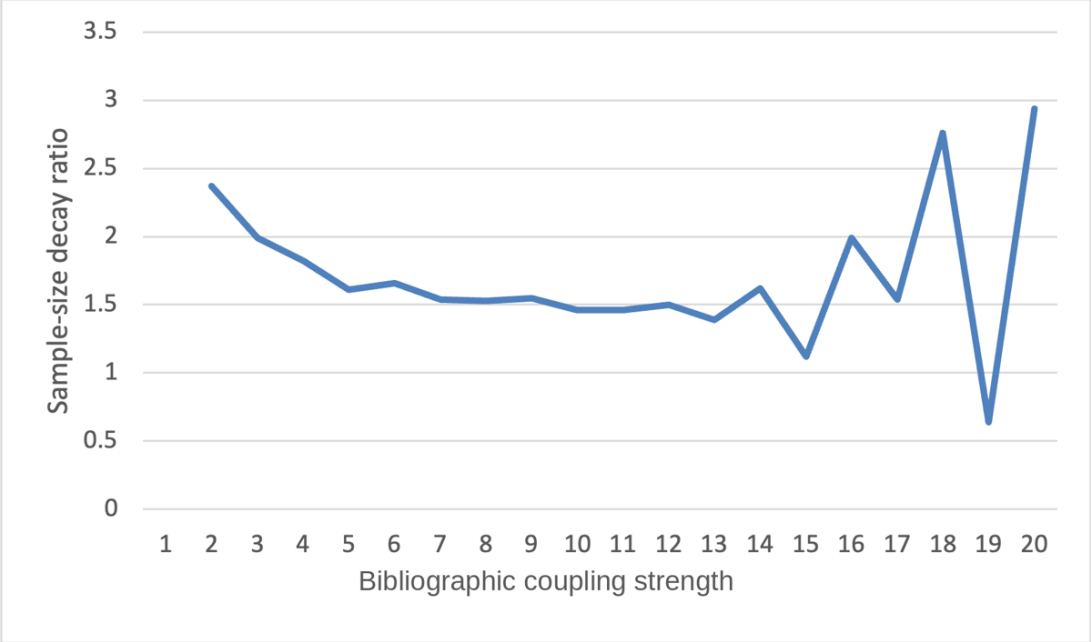


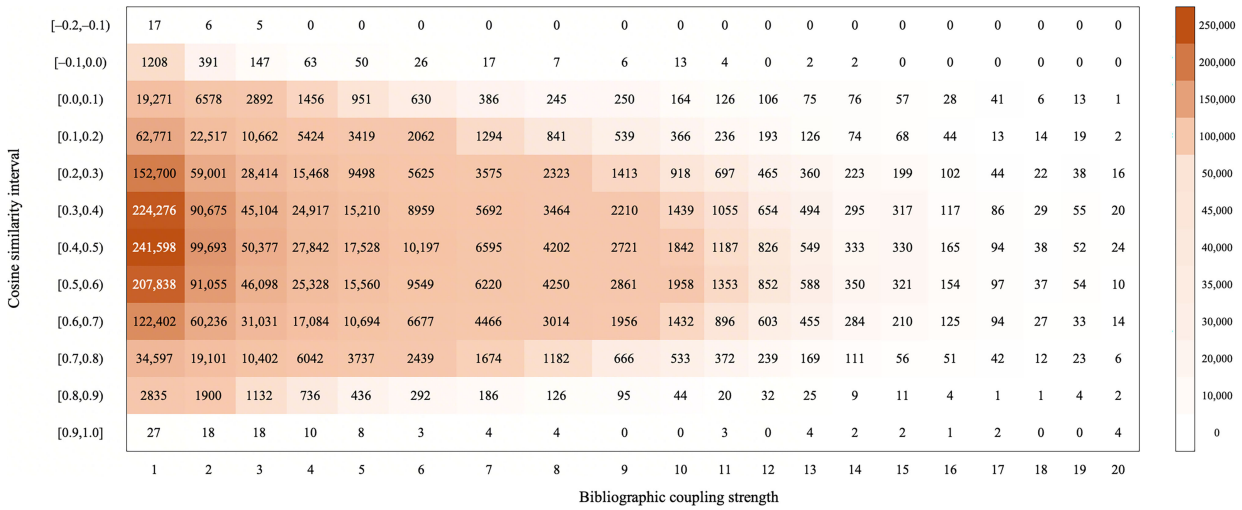
Table 2. Comparison of candidate thresholds for bibliographic coupling strength based on high-similarity enrichment factors (EFs).

Bibliographic coupling strength *k*	High-similarity EF	Consistency (EF >1)
6	1.28	9/9
7	1.30	9/9
8	1.32	9/9
9	1.31	9/9
10	1.33	9/9
11	1.31	8/9
12	1.32	8/9
13	1.32	8/9
14	1.30	8/9

Coupling Strength-Cosine Similarity Heatmap

Based on 2,085,004 pairwise comparisons, this study generated a heatmap (Figure 6), illustrating coupling strength (x-axis) against cosine similarity (y-axis) to provide a visual representation of the data distribution.

Figure 6. Heatmap of the joint distribution of cosine similarity intervals and bibliographic coupling strength.



The results reveal that samples were mainly concentrated in regions of moderate cosine similarity and low bibliographic coupling strength. This distribution indicates that citation sentences within the ADC oncology research field usually share similar research objects, disease types, drug mechanisms, and technical terminology. Thus, cosine similarity does not cluster heavily in very low ranges but naturally aggregates in midrange intervals. This moderate similarity primarily reflects topical and terminological similarity arising from the same research domain and should not be directly equated with ITCs. In contrast, bibliographic coupling strength exhibited a clear long-tailed distribution: out of 2,085,004 pairwise records, 1,948,454 (93.45%) had coupling strength no greater than 5, 2,057,397 (98.68%) were below 10, and only 27,607 (1.32%) reached 10 or above. Therefore, the region of both high similarity and high coupling strength exhibited a scattered distribution and anomalies, which may arise from ITC behavior causing abnormally elevated textual similarity and coupling strength. Based on these results, this study established a unified identification threshold—cosine similarity ≥0.60 and coupling strength ≥10—as the unified identification threshold for ITC behavior to separate anomalous instances.

Detection Results Under the Dual-Threshold Framework

Under the dual-threshold condition of cosine similarity ≥0.60 and bibliographic coupling strength ≥10, 6165 ITC records were ultimately screened, accounting for 0.296% of all effective pairwise records. As shown in Table 3, among the full sample of 2,085,004 pairs, 351,395 (16.85%) pairs met only the SBERT threshold of ≥0.60; 27,607 (1.32%) pairs met only the coupling threshold of ≥10; and 6165 (0.296% of the full sample, approximately 0.30%) pairs met both conditions simultaneously. These 6165 pairs crossed the thresholds on the 2 independent dimensions of high semantic similarity and high structural coupling.

Table 3. Detection results obtained using the dual-threshold framework (N=2,085,004).

Coupling group	SBERT[a] ≥0.60, n (%)	SBERT <0.60, n (%)	Total, n (%)
Coupling ≥10	6165 (0.30)	21,442 (1.03)	27,607 (1.32)
Coupling <10	345,230 (16.56)	1,712,167 (82.12)	2,057,397 (98.68)
Total	351,395 (16.85)	1,733,609 (83.15)	2,085,004 (100)

[a]SBERT: sentence bidirectional encoder representations from transformers.

Robustness, Expert Review, and Representative Cases

To examine the stability of the dual-threshold framework, this study constructed 9 parameter combinations around cosine similarity thresholds of 0.55, 0.60, and 0.65 and bibliographic coupling strength thresholds of 9, 10, and 11, using 0.60 and 10 as the baseline thresholds. The results showed that as the thresholds increased, the number of ITCs decreased continuously without abnormal jumps; the baseline combination maintained high overlap with adjacent combinations, with all Jaccard coefficients at least 0.40. Under different threshold combinations, the samples remained mainly distributed in the Introduction and Discussion sections, indicating that threshold perturbation primarily affected the scale of identification rather than the basic contextual structure of the samples. The threshold sensitivity analysis is shown in Table S8 in Multimedia Appendix 1.

For manual validation, 3 experts with backgrounds in information science, scientometrics, or biomedical literature analysis were invited to participate in semistructured interviews and independent, blinded review. From the ITC data identified by the dual-threshold model, 100 records were selected using stratified random sampling based on citation function and citation positions. Reviewers saw only the 2 citation-sentence texts, not similarity scores or coupling strength, and they judged whether a possible citation relationship existed. Table S9 in Multimedia Appendix 1 shows that 85 samples received unanimous support from all 3 experts, corresponding to a unanimous expert-confirmation rate of 85.00% (Wilson 95% CI 76.72%-90.69%). Expert judgments also showed high intuitive agreement: complete agreement among the 3 experts was 85.00%, and pairwise agreement ranged from 86.00% to 98.00%. Although Fleiss κ was only 0.0097 due to category imbalance, the high intuitive agreement and confirmation rate indicate that the model-hit cases had good interpretability.

To bridge algorithmic identification and scholarly practice, 2 representative cases were further selected (Tables 4 and 5). The first occurred in the Introduction. The highly cited source literature A, focused on ADCs for precision chemotherapy delivery across tumor types. Publications B and C both cited A in the Introduction, and C was published after B. The citation sentences in B and C both concerned the application of ADCs in human epidermal growth factor receptor 2 (HER2)–positive metastatic breast cancer and were highly similar in research object, argumentative focus, and expression structure, with a cosine similarity of 0.94 and a coupling strength of 13. However, C cited only source literature A and did not cite the earlier mediating literature B, whose expression pathway was highly similar. This case suggests that follower publications within the Introduction may adopt the intermediary's summary and problem-organization pathway of the original study, forming a potential ITC that retains original attribution while obscuring the intermediary pathway.

Table 4. A representative invisible triangular citation case in the Introduction section.

Introduction	Source literature A	Mediating literature B	Follow-up literature C
Published	March 2022	April 4, 2024	August 2025
Document type	Article	Article	Article
Citation position	—[a]	Introduction	Introduction
Title	ADC[b] smart chemotherapy delivery across tumor histologies	Optimal sequential strategies for ADC in MBC[c]: evaluating efficacy and cross-resistance	Prognostic index for predicting outcomes of trastuzumab deruxtecan in HER2[d] expressing MBC: a real-world multicenter study
Citation sentence	—	In recent years, several novel ADCs have been approved for the treatment of breast cancer, particularly HER2 positive MBC.	In recent years, ADCs have revolutionized the treatment of breast cancer, particularly HER2-positive MBC.

[a]Not applicable.
[b]ADC: antibody-drug conjugate.
[c]MBC: metastatic breast cancer.
[d]HER2: human epidermal growth factor receptor 2.

Table 5. A representative invisible triangular citation case in the Discussion section.

Discussion	Source literature A	Mediating literature B	Follow-up literature C
Published	October 2022	May 2023	April 18, 2025
Document type	Article	Article	Article
Citation position	—[a]	Discussion	Discussion
Title	Quantitative measurement of HER2[b] expression to subclassify ERBB2[c] unamplified breast cancer	Digital image analysis and assisted reading of the HER2 score display reduced concordance: pitfalls in the categorisation of HER2-low breast cancer	Machine learning prediction of HER2-low expression in breast cancers based on hematoxylin-eosin-stained slides
Citation sentence	—	Alternative assays for the identification of the HER2-low category have been proposed, including quantitative fluorescence assays or the use of DIA[d], including as a decision-aiding tool.	Alternative methods, such as quantitative fluorescence assays or DIA, have been proposed for the identification of the HER2-low category.

[a]Not applicable.
[b]HER2: human epidermal growth factor receptor 2.
[c]ERBB2: Erb-B2 receptor tyrosine kinase 2.
[d]DIA: digital image analysis.

The second case occurred in the Discussion section. Source literature A focused on the quantitative measurement of HER2 expression and its use in subclassifying Erb-B2 receptor tyrosine kinase 2 (ERBB2)–unamplified breast cancer. Publications B and C both discussed the identification and classification of HER2-low breast cancer, and C was published after B. The citation sentences of B and C were highly consistent around the core claim that alternative detection methods, such as quantitative fluorescence assays or digital image analysis (DIA), had been proposed for identifying the HER2-low category, with a cosine similarity of 0.85 and 10 shared references. Careful comparison showed that C did not copy B verbatim but retained the original semantic structure through synonym substitution and syntactic restructuring, for example, replacing "assays" with "methods," replacing "including" with "such as," and rewriting "the use of DIA" as "digital image analysis (DIA)." However, C still cited only the original publication A and did not cite the mediating literature B, which was closer in expression pathway and had explicitly presented quantitative fluorescence assays and DIA as decision-aiding tools for HER2-low classification.

In summary, the dual-threshold model using cosine similarity ≥0.60 and bibliographic coupling strength ≥10 can effectively identify and distinguish ITCs in the target dataset.

Analysis of ITCs' Positional Features

Overall Distribution of ITC Positions

Applying the identification threshold of "cosine similarity ≥0.60 and coupling strength ≥10," this study screened a total of 6165 qualifying records. The distribution of citation positions (Table S10 in Multimedia Appendix 1) showed that ITCs were mainly concentrated in the Introduction (n=6023, 49.1%) and Discussion (n=5833, 47.5%), whereas Results (n=255, 2.1%) and Methods (n=163, 1.3%) accounted for much smaller proportions. The proportions of ITCs in both the Introduction and Discussion exceeded the corresponding proportions among normal citations, indicating that these sections are high-occurrence locations for ITCs. Subsequently, chi-square tests were used to investigate the association between cosine similarity intervals or coupling strength and citation positions, further elucidating the positional characteristics of ITCs.

The Correlation Between Cosine Similarity and Citation Position

This study further analyzed the relationship between cosine similarity and citation positions among ITCs. The Pearson chi-square test showed a statistically significant association between cosine similarity and citation positions (χ^2=215.270;

P<.001; Cramér V=0.076). This result indicates that within ITCs, the distribution of citation positions differs statistically across cosine similarity intervals, although the effect size is small. In other words, cosine similarity is an important condition for identifying ITCs, but variation across similarity intervals does not strongly change the section-level distribution of ITCs.

As shown in Table S11 in Multimedia Appendix 1, the Introduction and Discussion sections dominated across all similarity intervals, particularly in the 0.60 to 0.80 interval. This indicates that ITCs do not depend on any specific similarity interval but consistently concentrate in the Introduction and Discussion once the dual-threshold criteria are met.

The Correlation Between Coupling Strength and Citation Position

This study further analyzed the relationship between bibliographic coupling strength and citation positions among ITCs. The Pearson chi-square test showed a statistically significant association between bibliographic coupling strength and citation positions (χ^2_9=140.233; P<.001; Cramér V=0.062). This result indicates that within ITCs, the distribution of citation positions differs statistically across coupling-strength intervals, although the effect size is small. That is, bibliographic coupling strength is associated with citation positions, but its practical influence is limited.

As shown in Table S12 in Multimedia Appendix 1, the Introduction and Discussion sections dominated across all coupling-strength intervals. The proportions for Introduction in the intervals of 10, 11 to 12, 13 to 15, and ≥16 were approximately 32.58% (n=1962), 34.97% (n=2106), 21.78% (n=1312), and 10.68% (n=643), respectively; for Discussion, they were approximately 32.32% (n=1885), 36.12% (n=2107), 21.84% (n=1274), and 9.72% (n=567). This indicates that the concentration of ITCs by section does not occur only at a specific coupling-strength level but remains relatively stable across levels. Methods accounted for 34.97% (n=57) of the ≥16 interval, higher than for the Introduction and Discussion, but because the Methods sample size was small, this should be treated as a supplementary observation rather than being overinterpreted. Results were mainly concentrated in the coupling-strength interval of 10, accounting for 45.49% (n=116).

Discussion

Main Study Findings

This study addressed the 3 objectives proposed in the Introduction. First, based on 2,085,004 effective pairwise records, it determined dual thresholds of cosine similarity ≥0.60 and bibliographic coupling strength ≥10 and identified 6165 ITC records, accounting for 0.296% of all data. Second, threshold sensitivity analysis showed that threshold perturbation mainly affected the number of ITCs without changing their basic section distribution; among a stratified sample of 100 records, 85 received unanimous support from the 3 experts, indicating that the model-hit records had some interpretability. Third, ITCs were mainly distributed in the Introduction (n=6023, 49.1%) and Discussion (n=5833, 47.5%), whereas Results (n=255, 2.1%) and Methods (n=163, 1.3%) accounted for relatively small proportions. Associations between citation positions and textual similarity or bibliographic coupling strength reached statistical significance, but Cramér V values were 0.076 and 0.062, respectively, indicating a weak practical association.

Citation position was statistically associated with both textual similarity and bibliographic coupling strength, but the effect sizes were small. This suggests that the semantic and structural dimensions provide complementary information, but neither alone is sufficient to explain the formation of ITCs. Information-seeking cost, citation propagation mechanisms, source memory, and awareness of scholarly attribution can serve as theoretical lenses for understanding these distributional features, but this study did not directly test their causal effects. Therefore, these mechanisms should be treated as hypotheses for future research.

Interpretations, Implications, and Comparisons to Existing Literature

Compared with previous research, this study extends both the research object and the identification methods. Existing citation-accuracy studies have mainly manually compared citing statements with the cited originals to determine whether citations support relevant claims and to identify nonexistent findings, misinterpretations, numerical miscitations, and transfers of content from other sources [18,19]. Triadic citation and omission studies have begun to combine citation networks, textual similarity, and bibliographic coupling to identify potential intermediary pathways [2,20,21]. Building on these studies, the present work jointly analyzes semantic similarity at the full-text citation sentence level and structural coupling reflected by shared references and evaluates ITCs through threshold sensitivity analysis, expert review, and section-location comparison. The findings are consistent with prior evidence that citation errors may propagate along citation chains but further emphasize that the interpretive or organizational contributions of intermediary publications may not be transparently presented. It should be noted that this study identifies suspected intermediary-dependence pathways rather than factually confirmed cases of citation misconduct.

The Introduction emerged as the predominant locus for such malpractice (n=6023, 49.1% of identified cases), reflecting the vulnerability of contextual framing to secondary sourcing. This vulnerability stems from the convergence of the principle of least effort and information foraging theory. The former holds that individuals tend to choose behavioral paths that reduce time and information-processing costs [32], whereas the latter suggests that information users weigh expected benefits against search costs and prioritize sources that provide high information gain at a lower cost [33]. Therefore, when authors rapidly establish a research background in the Introduction, they may be more likely to rely on mediating literature B, which has

already summarized, selected, and organized source literature A, because B provides a low-cost information entry point, an interpretive framework, and ready-made citation cues. However, B's contribution often appears as background integration, contextual organization, or provision of citation leads and may therefore be treated as an auxiliary reading pathway and be weakened or omitted in formal citation.

The Matthew effect further suggests that research outputs with greater reputation and visibility are more likely to receive continuing attention and credit [34]. When A is regarded as the original source or authoritative basis for a claim, citing A conforms more closely to the normative expectation of citing original literature and provides more direct formal legitimacy for the argument. Consequently, ITCs may occur when authors acquire, understand, and organize information about A through B but retain only the citation to A in the text, without adequately presenting B's intermediary contribution to knowledge selection, interpretation, and referral.

Similarly, the Discussion was the second most common location of such malpractice, accounting for 47.5% (5833/12,274) of identified cases, suggesting that the stages of result interpretation and theoretical integration may also constitute high-risk contexts. Unlike the Introduction, which mainly establishes background, the Discussion typically interprets findings, compares prior results, addresses inconsistent evidence, and proposes theoretical implications. In this context, authors often need to rapidly invoke existing mechanisms, similar findings, or conceptual frameworks that can explain their own results. Mediating literature B, which has already selected, interpreted, and theoretically organized source literature A, may therefore become a low-cost entry point, since it not only helps authors understand how A's results are interpreted and used but also provides an argumentative pathway that can be directly embedded in the Discussion.

Source monitoring theory can further explain why B may be omitted. Source monitoring refers to the process by which individuals judge and attribute the sources of memories, knowledge, and beliefs; this process may involve source confusion or misattribution [35]. Cognitive psychology experiments show that individuals may remember previously encountered ideas while forgetting or misjudging their specific sources, producing inadvertent plagiarism or source-attribution errors [36]. In idea-generation tasks, inadequate source monitoring can also increase the likelihood of unintentionally using others' prior ideas [37]. Therefore, authors may retain interpretive frameworks, comparative perspectives, or theorized expressions obtained from B but fail to preserve the source pathway through which they were obtained.

Conversely, out of 12,274 records, the Results (n=255, 2.1%) and Methods (n=163, 1.3%) sections demonstrated remarkable resistance to ITCs, highlighting the methodological sanctity accorded to conclusions, data, and technical procedures in biomedical research. When describing final results, core mechanisms, experimental procedures, or analytical frameworks, the nonnegotiable requirement for direct source verification naturally forms a barrier against intermediary dependence. Exceptions may arise when intermediary publications function as language or access gatekeepers—for example, when the original work exists only in a difficult-to-access language or behind a prohibitive paywall—forcing authors to obscure citation pathways to conceal their lack of access to the original source.

The biomedical implications of these findings extend beyond academic integrity to tangible research quality and patient safety concerns. ITCs distort knowledge inheritance chains, potentially propagating misinterpreted concepts across generations of literature. Within clinical domains such as ADC research—where mechanistic insights, drug efficacy, and safety directly inform therapeutic development—such distortions could cascade into flawed trial designs or clinical applications rooted in misattributed evidence. However, this study did not trace downstream citation chains of the identified ITCs and has no evidence that these cases have already caused specific clinical misunderstandings or consequences for clinical decision-making. The clinical impact should therefore be understood as a potential risk requiring further validation.

The vulnerability of the Introduction and Discussion sections is particularly consequential, as foundational inaccuracies embedded in literature reviews can systemically corrupt subsequent research narratives. When synthesizing, comparing, or critically evaluating prior ideas, researchers should distinguish the original evidence provided by A from the summaries and interpretive frameworks provided by B. If the understanding, summarization, or argumentative structure of A substantially derives from B, B's interpretive contribution should also be acknowledged. Mitigation requires multilayered governance strategies: standardized academic writing curricula should explicitly train researchers in ethical source tracing, highlighting the "efficiency trap" of overreliance on secondary sources during literature synthesis; journals must implement cross-disciplinary review panels empowered with algorithmic screening tools incorporating our dual-threshold model; funding agencies could mandate citation network transparency statements for grant applications; technologically, real-time citation tracking plugins should visualize knowledge provenance networks during manuscript preparation, enabling authors to self-correct attribution gaps before submission.

From the perspective of research ethics, even if some ITCs arise from information-search costs or source monitoring errors rather than deliberate deception, authors still bear responsibility for verifying sources and transparently acknowledging knowledge pathways. Research integrity means not only avoiding severe misconduct but also adhering to professional standards of accuracy, transparency, and traceability [38]. Therefore, theories from cognitive psychology and sociology of science can explain how ITCs may form but cannot eliminate authors' responsibility for scholarly attribution. Algorithmic identification should serve only as a risk signal and cannot directly determine authors' subjective intent or establish academic misconduct; judgments should

still be made in combination with the full-text context and human review.

Limitations

While establishing a robust methodological foundation, this study acknowledges limitations that require future resolution. The exclusive focus on ADC—though strategically advantageous for examining an emergent field with high citation velocity—constrains generalizability across biomedical subdisciplines with distinct citation cultures. Comparative studies across oncology, neuroscience, and epidemiology would reveal the disciplinary modulation of prevalence and manifestation patterns of ITCs. Additionally, incorporating semantic features such as citation sentiment, citation strength, and citation function could enhance detection specificity. The current data source was limited to Web of Science ECR data; future research should aggregate datasets across multiple domains to validate the universality of the dual-threshold model at a larger scale. Finally, expanding the investigation to nonbiomedical fields would test the model's universality and reveal how disciplinary epistemology shapes citation malpractice.

Moreover, the dual-threshold results support potential content and citation-propagation pathways but cannot directly prove mechanisms such as the principle of least effort, the Matthew effect, source-monitoring errors, or authors' subjective intent. These mechanisms should be regarded as possible theoretical explanations and hypotheses for future research. Subsequent studies could combine author interviews, manuscript version histories, reference-management records, literature-access traces, and bibliometric indicators of related publications to distinguish and directly test different mechanisms. Future work could also start with identified ITCs and trace downstream citing publications, classify them as basic research, clinical research, systematic reviews, and clinical guidelines, and invite experts in information science and clinical fields to jointly examine changes in claims across A, B, C, and downstream publications, thereby evaluating whether and how ITCs produce clinical consequences.

Conclusions

In synthesizing these insights, we conclude that the developed dual-threshold model provides a transferable framework for ITC detection, validated through large-scale empirical analysis within the ADC domain, and it pioneers novel applications of ECR technology. The results show that ITCs are mainly concentrated in the Introduction and Discussion sections; however, their associations with textual similarity and bibliographic coupling strength are weak, suggesting that the phenomenon cannot be explained by a single semantic or structural indicator. The proposed countermeasures—spanning pedagogical reform, peer review innovation, and technological augmentation—offer actionable pathways toward more transparent knowledge ecosystems. Future research must prioritize multidisciplinary validation studies and integrate richer linguistic and sentiment analyses to fully illuminate the complex ecology of citation malpractice. By addressing these concealed pathways of knowledge transmission, the scientific community can strengthen the ethical foundations of scholarly communication, which is particularly crucial in high-stakes domains where attributional integrity directly impacts human health.

Acknowledgments

The authors thank all referees for their important insightful comments and suggestions. They would also like to thank Aman Xu for the helpful discussion of the manuscript.

Generative AI Use

The authors used a generative AI tool (DeepL) to assist with language polishing, structural checking, and code optimization. All aspects of the study design, data collection, statistical analysis, interpretation of results, and the final textual decisions were completed by the authors and remain their responsibility. All AI-generated or AI-modified content was manually checked, revised, and confirmed by the authors before inclusion in the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (72304181), the Scientific Research Start-up Fund for Excellent Doctor (Youbo; Class A; LH) (The Third Batch) by The First Affiliated Hospital of Naval Medical University/Changhai Hospital, Naval Medical University, the Xueshu Xinxiu Talent Fund by Naval Medical University, and the Shenlan Qihang Talent Fund (The Third Batch) by Naval Medical University. The funders played no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

Data Availability

The publicly testable samples, threshold-calibration outputs, and custom analysis scripts used in the automated annotation process have been deposited in a public GitHub repository [39]. Because some raw records are restricted by database terms of use, the repository provides anonymized, deidentified, and reproducible derivative data and analysis code to support replication and independent verification.

Authors' Contributions

Conceptualization: YL (lead), LH (equal)
Data curation: YL (lead), JX (equal)

Formal analysis: YL (lead), JX (equal)
Funding acquisition: YL (lead), LH (equal)
Investigation: YL (lead), JX (equal)
Methodology: YL (lead), LH (supporting)
Project administration: LH
Resources: LH
Software: YL (lead), JX (equal)
Supervision: LH
Validation: YL (lead), JX (equal)
Visualization: YL (lead), JX (equal)
Writing – original draft: YL (lead), JX (equal), LH (supporting)
Writing – review & editing: YL (equal), JX (equal), LH (lead)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables (S1-S12) reporting document-type distributions, enriched cited reference citation functions and positions, true and invisible triangular citation relationships, bibliographic coupling/cosine similarity correlations, threshold sensitivity analyses, expert blind review results, and related distribution analyses.
[DOCX File (Microsoft Word File), 48 KB-Multimedia Appendix 1]

References

1. Liu Y, Yang L, Chen M. A new citation concept: triangular citation in the literature. J Informetr. May 2021;15(2):101141. [doi: 10.1016/j.joi.2021.101141]
2. Liu Y, Li R, Sheng X. Invisible triangular citations: an irregular citations in scientific papers [Article in Chinese]. J Modern Infor. 2024;44(4):127-141. URL: https://d.wanfangdata.com.cn/periodical/CiBQZXJpb2RpY2FsQ0hJU29scjkyMDI2MTAxMjE5NTI0NxINeGRxYjIwMjQwNDAxMhoId2tjd3dicXc%253D [Accessed 2026-10-03]
3. Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. BMJ. Jul 20, 2009;339:b2680. [doi: 10.1136/bmj.b2680] [Medline: 19622839]
4. Peoples N, Østbye T, Yan LL. Burden of proof: combating inaccurate citation in biomedical literature. BMJ. Nov 6, 2023;383:e076441. [doi: 10.1136/bmj-2023-076441] [Medline: 37931929]
5. Choi J, Gupta A, Kaghazchi A, Htwe TS, Baiocchi M, Spain DA. Citation inaccuracies in influential surgical journals. JAMA Surg. Aug 1, 2021;156(8):791-792. [doi: 10.1001/jamasurg.2021.1445] [Medline: 34037684]
6. Ding Y, Zhang G, Chambers T, Song M, Wang X, Zhai C. Content-based citation analysis: the next generation of citation analysis. J Assn Inf Sci Tec. Sep 2014;65(9):1820-1833. [doi: 10.1002/asi.23256]
7. Huang Y, Bu Y, Ding Y, Lu W. Exploring direct citations between citing publications. J Infor Sci. Oct 2021;47(5):615-626. [doi: 10.1177/0165551520917654]
8. Asubiaro TV, Ajiferuke I. Semantic similarity-based credit attribution on citation paths: a method for allocating residual citation to and investigating depth of influence of scientific communications. Scientometrics. Nov 2022;127(11):6257-6277. [doi: 10.1007/s11192-022-04522-3]
9. Budi I, Yaniasih Y. Understanding the meanings of citations using sentiment, role, and citation function classifications. Scientometrics. Jan 2023;128(1):735-759. [doi: 10.1007/s11192-022-04567-4]
10. Citation context in Web of Science. Clarivate; 2022. URL: https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/dlm_uploads/2022/05/202205-WoS-citation-context-final.pdf [Accessed 2026-08-20]
11. Clarivate launches new web of science to accelerate the pace of research and discovery. Clarivate. 2021. URL: https://clarivate.com/news/clarivate-launches-new-web-of-science-to-accelerate-the-pace-of-research-and-discovery/ [Accessed 2026-08-20]
12. Enriched cited references—citation function class. Clarivate. 2025. URL: https://webofscience.zendesk.com/hc/en-us/articles/27713957430033-Enriched-Cited-References-Citation-Function-Class [Accessed 2026-08-20]
13. Lin GG, Hou HY, Liang GQ, Hu ZG. Research on the diversity of citation motivation based on enriched cited references [Article in Chinese]. Infor Stud Theory Appl. 2024;47(2):61-69. URL: https://d.wanfangdata.com.cn/periodical/qbllysj202402009 [Accessed 2026-10-03]
14. Su GL, Xu XJ. Seven unfavorable behaviors in citing references in social science monographs [Article in Chinese]. Library Work and Study. 2002(2):15-17. URL: https://tinyurl.com/bdhhdndv [Accessed 2026-10-03]

15. Su XN. Quality control of data in citation indexes [Article in Chinese]. J Lib Sci China. 2001(2):76-78. URL: https://tinyurl.com/3ahbkapr [Accessed 2026-11-03]
16. Brennan D. Simple export of journal citation data to Excel using any reference manager. J Med Libr Assoc. Jan 2016;104(1):72-75. [doi: 10.3163/1536-5050.104.1.012] [Medline: 26807057]
17. Li J, Wu Y, Deng P, Leng F. Automatic data processing strategy of citation anomie based on feature fusion [Article in Chinese]. Data Anal Know Dis. 2020;4(5):38-45. URL: https://d.wanfangdata.com.cn/periodical/CiBQZXJpb2RpY2FsQ0hJU29scjkyMDI2MTAxMjE5NTI0NxIReGR0c3FianMyMDIwMDUwMDQaCGh0bzQ1eDd6 [Accessed 2026-11-03]
18. Smith N, Cumberledge A. Quotation errors in general science journals. Proc R Soc A. Oct 2020;476(2242). [doi: 10.1098/rspa.2020.0538]
19. Pavlovic V, Weissgerber T, Stanisavljevic D, et al. How accurate are citations of frequently cited papers in biomedical literature? Clin Sci (Lond). Mar 12, 2021;135(5):671-681. [doi: 10.1042/CS20201573] [Medline: 33599711]
20. Yang WX, Deng SH, Hu HT, Kong J, Yan XH. A comparative study on the difference of literature position in triangular citation [Article in Chinese]. J Inf Res Manage. 2024;14(1):131-145. URL: https://d.wanfangdata.com.cn/periodical/xxzyglxb202401014 [Accessed 2026-11-03]
21. Yang WX, Deng SH, Yang J, Zhang Q, Wang H. Analysis of the performance characteristics of literature missing citations behavior based on non-parametric test [Article in Chinese]. Infor Sci. 2024;42(7):76-83. URL: https://d.wanfangdata.com.cn/periodical/qbkx202407010 [Accessed 2026-11-03]
22. Zhu H, Zhang D, Sun Y, Ren H. Analysis of citation errors in references of high-impact medical journals [Article in Chinese]. Chinese J Sci Tech Period. 2012;23(2):243-247. URL: https://d.wanfangdata.com.cn/periodical/zgkjqkyj201810012 [Accessed 2026-11-03]
23. Lv YP. Analysis and discussion on poor citation practices in academic papers and preventive measures [Article in Chinese]. Library Work and Study. 2012(11):53-56. URL: https://tinyurl.com/a6sdyw5j [Accessed 2026-11-03]
24. Dai Y, Zhang SX. Analysis of the phenomenon of intentional omission of citations in academic papers [Article in Chinese]. J Modern Infor. 2005(4):48-50. URL: https://tinyurl.com/27skcxa5 [Accessed 2026-11-03]
25. Wang ZB. Manifestations, causes and governance of citation anomie in academic journal papers [Article in Chinese]. China Publishing J. 2020(21):46-50. URL: https://tinyurl.com/5n7maxan [Accessed 2026-11-03]
26. Ghose T, Blair AH. Antibody-linked cytotoxic agents in the treatment of cancer: current status and future prospects. J Natl Cancer Inst. Sep 1978;61(3):657-676. [Medline: 80453]
27. Fu Z, Li S, Han S, Shi C, Zhang Y. Antibody drug conjugate: the “biological missile” for targeted cancer therapy. Signal Transduct Target Ther. Mar 22, 2022;7(1):93. [doi: 10.1038/s41392-022-00947-7] [Medline: 35318309]
28. Tsuchikama K, Anami Y, Ha SYY, Yamazaki CM. Exploring the next generation of antibody-drug conjugates. Nat Rev Clin Oncol. Mar 2024;21(3):203-223. [doi: 10.1038/s41571-023-00850-2] [Medline: 38191923]
29. Protection of human subjects. Electronic Code of Federal Regulations (eCFR). URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46 [Accessed 2026-10-21]
30. Animal welfare act. US Department of Agriculture, National Agricultural Library. URL: https://www.nal.usda.gov/animal-health-and-welfare/animal-welfare-act [Accessed 2026-10-21]
31. PHS policy. Grants & Funding (National Institutes of Health). URL: https://olaw.nih.gov/policies-laws/phs-policy.htm [Accessed 2026-10-21]
32. Zipf GK. Human Behavior and the Principle of Least Effort. Addison-Wesley Press; 1949. URL: https://psycnet.apa.org/record/1950-00412-000 [Accessed 2026-10-21]
33. Pirolli P, Card S. Information foraging. Psychol Rev. 1999;106(4):643-675. [doi: 10.1037/0033-295X.106.4.643]
34. Merton RK. The Matthew effect in science: the reward and communication systems of science are considered. Science. Jan 5, 1968;159(3810):56-63. [Medline: 5634379]
35. Johnson MK, Hashtroudi S, Lindsay DS. Source monitoring. Psychol Bull. Jul 1993;114(1):3-28. [doi: 10.1037/0033-2909.114.1.3] [Medline: 8346328]
36. Brown AS, Murphy DR. Cryptomnesia: delineating inadvertent plagiarism. J Exp Psychol Learn Mem Cogn. 1989;15(3):432-442. [doi: 10.1037/0278-7393.15.3.432]
37. Marsh RL, Landau JD, Hicks JL. Contributions of inadequate source monitoring to unconscious plagiarism during idea generation. J Exp Psychol Learn Mem Cogn. 1997;23(4):886-897. [doi: 10.1037/0278-7393.23.4.886]
38. Steneck NH. Fostering integrity in research: definitions, current knowledge, and future directions. Sci Eng Ethics. Jan 2006;12(1):53-74. [doi: 10.1007/pl00022268] [Medline: 16501647]
39. Invisible-triangular-citations-ADC. GitHub. URL: https://github.com/Norah0205/Invisible-Triangular-Citations-ADC [Accessed 2026-10-21]

Abbreviations

ADC: antibody-drug conjugate
BIC: Bayesian information criterion
DIA: digital image analysis
ECR: Enriched Cited References
EF: enrichment factor
ERBB2: Erb-B2 receptor tyrosine kinase 2
HER2: human epidermal growth factor receptor 2
IMRAD: Introduction, Methods, Results, and Discussion
ITC: invisible triangular citation
SBERT: sentence bidirectional encoder representations from transformers

Edited by Ivan Steenstra; peer-reviewed by Farshid Danesh, Zhentao Liang; submitted 27.Mar.2026; final revised version received 08.Oct.2026; accepted 08.Oct.2026; published 17.Nov.2026

Please cite as:
Liu Y, Xu J, Huang L
Detecting Invisible Triangular Citations in Biomedical Research Using a Dual-Threshold Framework Based on Textual Similarity and Bibliographic Coupling Strength: Method Development and Validation Study
JMIR Med Inform 2026;14:e96265
URL: https://medinform.jmir.org/2026/1/e96265
doi: 10.2196/96265

© Yunmei Liu, Jingyi Xu, Lei Huang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 17.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.