C O R R E S P O N D E N C E

References to nowhere: an audit of fabricated citations across 2·5 million biomedical papers

Maxim Topaz[*], Nir Roguin, Pallavi Gupta, Zhihong Zhang, Laura-Maria Peltonen

mt3315@cumc.columbia.edu

School of Nursing (MT, NR, PG, ZZ), Data Science Institute (MT, ZZ), Columbia University, New York, NY, USA; VNS Health, New York, NY 10032, USA (MT); Tel Aviv Sourasky Medical Center, Tel Aviv, Israel (NR); Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel (NR); Department of Health and Social Management, University of Eastern Finland, Kuopio, Finland (L-MP); Wellbeing Services County of North Savo, Kuopio, Finland (L-MP); Wellbeing Services County of Southwest Finland, Turku, Finland (L-MP); Department of Nursing Science, University of Turku, Turku, Finland (L-MP)

Scientific literature depends on the integrity of its references. Each reference implicitly asserts that a verifiable source exists and supports the claims being made. When references point to non-existent studies, readers, reviewers, and policy makers are unable to evaluate the evidence.

Fabricated references (references whose claimed titles correspond to no existing publication) can arise from paper mill activity, intentional misconduct, or uncritical use of artificial intelligence (AI) writing tools.[1] Large language models (LLMs) generate plausible sounding but fictitious references, a well documented failure mode; previous studies estimate that 30–69% of LLM-generated references in biomedical contexts are fabricated.[2,3] These references are often correctly formatted, attributed to real researchers, and bear plausible publication dates, making them difficult to detect by conventional peer review.[4] No systematic audit of reference integrity across the biomedical literature had been conducted until now.

We present findings from a reference-integrity audit of 2·5 million biomedical papers spanning 3 years, showing that fabricated references are embedded in the peer-reviewed literature at scale, and that the rate of fabrication is accelerating.

We developed an automated reference verification system scanning PubMed Central's Open Access subset from Jan 1, 2023, to Feb 18, 2026: 2 471 758 papers and 125 615 773 structured references. We extracted references from full-text extensible markup language, retaining those with a PubMed identifier (PMID). Of 125·6 million references, 97·1 million (77%) carried a PMID and were verified; the remaining 23%, predominantly non-indexed references to websites, books, and grey literature, were excluded. For each verified reference, we retrieved the bibliographical record for the claimed identifier from PubMed and Crossref and compared it with the citing paper's claimed metadata with the use of text-similarity scoring, and mismatches were flagged.

Flagged references underwent sequential filters to minimise false positives: automated pattern detection removed parsing artefacts, and an LLM (Claude 3.5 Haiku; Anthropic, San Francisco, CA, USA) screened remaining candidates to distinguish genuine fabrications from formatting discrepancies such as informally abbreviated titles. For example, a reference listed as *Depression and anxiety in young adults with ID* corresponds to the real indexed title *Depression and anxiety symptoms during the transition to early*

adulthood for people with intellectual disabilities and is probably a reference error, not a fabrication. The model was applied zero-shot without fine-tuning or modification of model weights.

References passing all filters were verified against PubMed (approximately 37 million records), Crossref (more than 160 million digital object identifiers), OpenAlex (more than 250 million scholarly works),[5,6] and Google Scholar (which indexes journals, preprints, conference proceedings, theses, and grey literature).[7] A reference not found in any database was classified as a fabricated reference; one found but linked to an incorrect identifier was a reference error (appendix p 2). Precision of our automated reference verification system was 91% (Fleiss' κ=0·71, indicating moderate agreement in about seven of every 10 cases), measured in a 500-entry masked validation with three independent reviewers; this design estimates precision but not recall.

Among 97·1 million verified references, we identified 4046 fabricated references across 2810 papers (illustrative examples are shown in the appendix p 5). In 2023, approximately one in 2828 papers contained at least one fabricated reference. By 2025, this had risen to one in 458 and in the first 7 weeks of 2026, one in 277 papers had at least one fabricated reference. The fabrication rate increased more than 12 times, from approximately four per 10 000 papers in 2023, to 51·3 per 10 000 papers in the fourth quarter of 2025, reaching 56·9 per 10 000 papers in early 2026 (figure).

A 2025 paper on ureteroileal anastomotic techniques in an open access oncology journal contained 18 (60%) fabricated references of 30 verified; each fabricated reference was tailored to the paper's narrow surgical topic, attributed to real urologists, and bore claimed publication years of 2023 or 2024.[8] Further examples, such as a references about rheumatology biomarkers linked to an identifier for a study of nematode worms, are shown in the appendix (p 5).

Beyond individual papers, we identified patterns consistent with paper mill activity: the same two authors appeared across 11 papers in a single surgical journal in 2025, with 15 fabricated references covering CRISPR diagnostics, AI-guided nanovaccines, and gut microbiome biomarkers, all sharing a core co-authorship pair. Most affected papers (91%, n=2564) contained one or two fabricated references; 246 contained three or more. Review articles had a fabrication rate that was 57% higher than other paper types (16·7 per 10 000 vs 10·6 per 10 000; p<0·0001; appendix p 7).

The sharp inflection in mid-2024 coincides with the expected publication lag following widespread LLM adoption, although increased paper mill activity and changes in journal indexing practices might also have contributed. LLMs became broadly available in late 2022 and 2023; with submission-to-publication times of 100–200 days,[9] LLM-assisted papers would appear in PubMed Central from mid-2024 onward.

The fabricated references we identified were not obviously defective: topically specific, correctly formatted, attributed to real researchers, and beared plausible publication dates. Systematic reviews have found that approximately one in four references in medical journal articles contains errors,[4] confirming that reference verification is not standard in peer review. Automated reference verification can close this gap.

The implications extend beyond individual papers. Paper mill articles have been included in systematic reviews informing clinical guidelines;[10] when a guideline cites a paper with a partly fictional reference list, the evidence chain for treatment decisions is compromised. Of the 2810 affected papers, 98·4% had received no publisher action at the time of our audit (appendix p 7).

Several limitations deserve mention. The exclusion of 23% of references with no PMIDs could bias estimates in either direction: fabricated references might be more common among non-indexed sources, including grey literature, websites, and books (undercounting), or they can be preferentially associated with PMIDs (overcounting). PubMed Central open access does not cover the full biomedical literature. The early 2026 data span only 7 weeks. This design estimates precision but not recall; fabricated references that evaded all pipeline filters are not counted. Some references might exist in sources outside our four databases. Our system identifies the problem, not its cause.

We recommend four actions. First, publishers should integrate automated reference verification into submission workflows before peer review begins; verification tools exist, and the barrier to adoption is institutional rather than technological. Second, indexing services should add integrity metadata to article records so that downstream users can assess the reliability of references. Third, publishers should retroactively screen existing publications and issue corrections or retractions when fabricated references compromise a paper's conclusions. Fourth, fabricated references do not currently exist as a discrete category in major research integrity databases; establishing this category would enable systematic tracking and accountability.

When references point to non-existent studies, the evidence they claim to support is fictional. Routine automated verification can close this gap before fabricated references reach the published record.

Figure

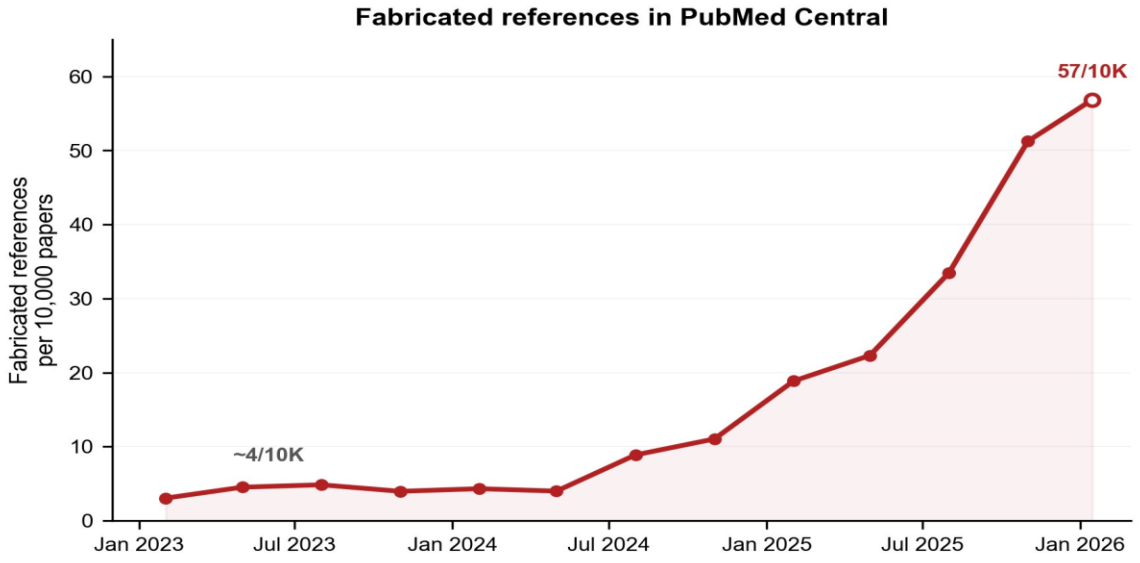


Figure: Quarterly rate of fabricated references per 10 000 papers in PubMed Central from January, 2023, to February, 2026. The fabrication rate remained stable at approximately four per 10 000 papers throughout 2023 (dashed line). Beginning in mid-2024, the rate rose sharply, reaching approximately 57 per 10 000 by early 2026. Each datapoint represents one calendar quarter. The open symbol indicates an incomplete quarter (Jan 1 to Feb 18, 2026); all filled symbols represent complete calendar quarters.

Contributors

MT conceived and designed the study, developed the automated reference verification system, conducted the analysis, and wrote the first draft. NR contributed to study design, data interpretation, and manuscript revision. PG and ZZ contributed to data processing and manuscript revision. L-MP contributed to study design, data interpretation, and critical revision of the manuscript. All authors approved the final version.

Declaration of interests

We declare no competing interests.

Data sharing

The aggregate dataset and a detailed description of the pipeline logic are available upon reasonable request to the corresponding author. Case-level data identifying individual papers and authors are available to editors and to researchers who can demonstrate appropriate safeguards and are not publicly released.

AI use

During the preparation of this work the authors used Claude (Anthropic) in order to assist with code development, manuscript drafting, and editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1 Else H, Van Noorden R. The fight against fake-paper factories that churn out sham science. Nature 2021; 591: 516–19.

2 Athaluri SA, Manthena SV, Kesapragada VSRKM, et al. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 2023; 15: e37432.

3 Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023; 13: 14045.

4 Jergas H, Baethge C. Quotation accuracy in medical journal articles—a systematic review and meta-analysis. PeerJ 2015; 3: e1364.

5 Priem J, Piwowar H, Orr R. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv 2022; published online April 29. https://doi.org/10.48550/arXiv.2205.01833 (preprint).

6 Culbert JH, Hobert A, Jahn N, et al. Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. Scientometrics 2025; 130: 2475–92.

7 Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods 2020; 11: 181–217.

8 Ren C, Xiao M, Zhu J, Tong W, Yi F. Comparative analysis of ureteroileal anastomotic stricture rates: Bricker versus Wallace techniques in ileal conduit urinary diversion—a single-surgeon study with BMI-matched design and long-term follow-up excluding cancer recurrence bias. Front Oncol 2025; 15: 1613772.

9 Andersen MZ, Fonnes S, Rosenberg J. Time from submission to publication varied widely for biomedical journals: a systematic review. Curr Med Res Opin 2021; 37: 985–93.

10 Tang G, Cai H. Citation contamination by paper mill articles in systematic reviews of the life sciences. JAMA Netw Open 2025; 8: e2515160.

S U P P L E M E N T A R Y A P P E N D I X

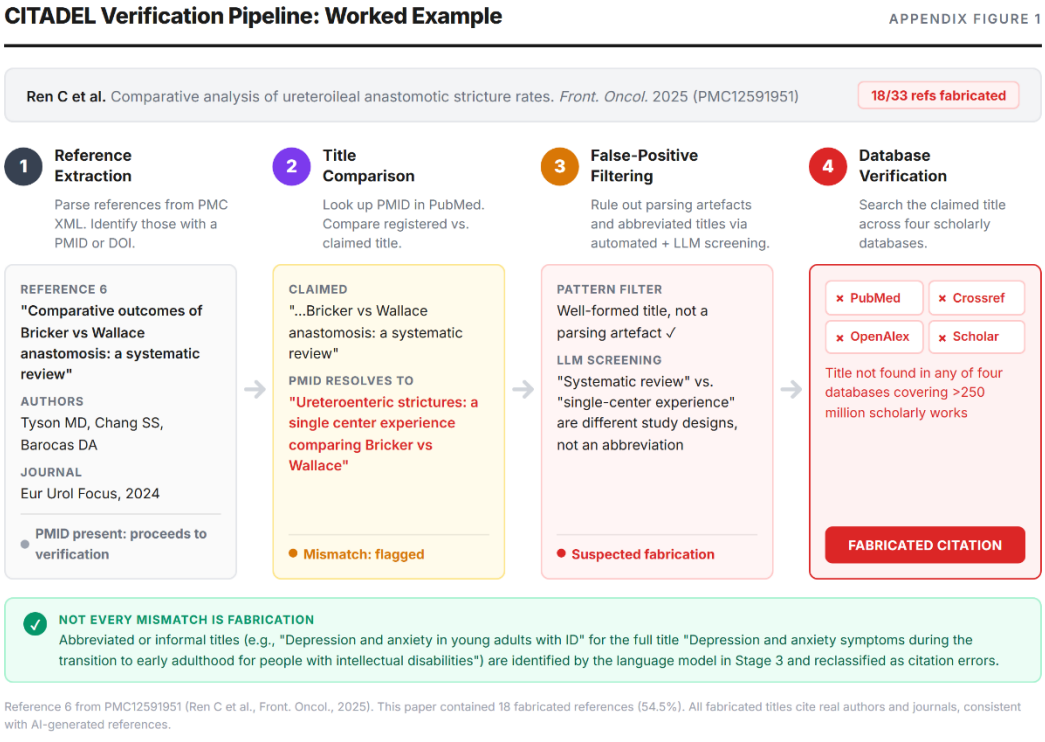
References to nowhere: an audit of fabricated citations across 2·5 million biomedical papers

Topaz M, Roguin N, Gupta P, Zhang Z, Peltonen L-M

Contents

Appendix 1: CITADEL Verification Pipeline

The CITADEL system (Citation Integrity Testing And Detection of Erroneous Literature) is an automated citation verification pipeline developed to detect suspected fabricated references at scale. It processes full-text XML from papers in the PubMed Central Open Access subset and applies sequential verification stages to each reference bearing a resolvable identifier. The pipeline is designed to minimize false positives by applying a series of filters before any reference is classified as a suspected fabricated reference. Figure A1 illustrates the four stages using a worked example.



Reference 6 from PMC12591951 (Ren C et al., Front. Oncol., 2025). This paper contained 18 fabricated references (54.5%). All fabricated titles cite real authors and journals, consistent with AI-generated references.

Stage 1: Reference Extraction

Full-text XML files were downloaded from the PubMed Central Open Access subset for all papers published between January 1, 2023 and February 18, 2026. Structured references were extracted from each paper's reference list. References bearing at least one resolvable identifier (a PubMed identifier or a digital object identifier) were retained for verification. References lacking any resolvable identifier, including citations to websites, books, government reports, and grey literature, were excluded from analysis. Of 125·6 million total references extracted, 97·1 million (77%) carried a resolvable identifier and proceeded to Stage 2.

Stage 2: Bibliographic Record Comparison

For each retained reference, the bibliographic record associated with the claimed identifier was retrieved from PubMed and Crossref and compared to the record claimed in the citing paper using automated text-similarity scoring. The comparison encompassed bibliographic details including title, journal, and publication year. References whose claimed and retrieved records did not reach a minimum similarity threshold were flagged as potential mismatches and passed to Stage 3. References whose claimed and retrieved records were sufficiently similar were classified as verified and excluded from further analysis.

Stage 3: False-Positive Filtering

Flagged references underwent two sequential filters to minimize false positives before any suspected fabrication classification was assigned. First, an automated pattern-detection step removed parsing artifacts: instances in which metadata elements such as author lists, consortium names, database names, or website addresses were misextracted from the XML as reference titles. Second, a large language model screening step helped distinguish genuine fabrications from formatting discrepancies such as informally abbreviated or truncated titles. The model received the claimed title, resolved metadata, and contextual information, and assessed whether any discrepancy was plausibly explained by citation error rather than fabrication. The language model (Claude 3.5 Haiku, Anthropic, San Francisco, CA, USA) was applied. Cases judged plausibly consistent were reclassified as citation errors and removed from the suspected fabrication pool; all others proceeded to Stage 4.

Stage 4: Database Verification

Each remaining reference title was searched across four scholarly databases: PubMed (over 37 million records), Crossref (over 160 million digital object identifiers), OpenAlex (over 250 million scholarly works), and Google Scholar (indexes journals, preprints, conference proceedings, theses, and grey literature). A reference whose claimed title could not be matched in any of the four databases was classified as a suspected fabricated reference. A reference whose title was found in at least one database was reclassified as a citation error and excluded from the fabricated reference count.

Validation

Classification accuracy was assessed through a blinded validation study. A stratified random sample of 500 flagged entries drawn across all study years was independently reviewed by three trained reviewers who had no access to the pipeline's automated classifications. Reviewers used external bibliographic tools to independently verify or refute each flagged entry. Inter-rater reliability was assessed using Fleiss' kappa. Pipeline precision was 91% (Fleiss' kappa = 0·71). This design estimates precision; recall cannot be estimated from this design because it does not capture references that the pipeline failed to flag.

Scope and Limitations

CITADEL identifies references whose claimed bibliographic details do not correspond to any verifiable publication in major bibliographic databases. Several limitations apply. First, the pipeline verified only

references with resolvable PubMed identifiers or digital object identifiers; the 23% of references lacking these identifiers were excluded. Second, the pipeline operates on the PubMed Central Open Access subset, which does not represent the full biomedical literature. Third, all thresholds were set conservatively to favor specificity over sensitivity; the reported counts represent a lower bound on the true prevalence. Fourth, the pipeline does not determine the cause of a suspected fabricated reference. Fifth, the language model screening step may introduce differential misclassification rates across subject areas or time periods. Sixth, a small proportion of references classified as suspected fabrications may exist in sources outside the four databases used for verification. The system architecture is designed to scale to non-biomedical disciplines and preprint servers.

Code and Data Availability

The pipeline code and detailed implementation documentation are available upon reasonable request to the corresponding author. The aggregate dataset of suspected fabricated references is available upon reasonable request. Case-level data identifying individual papers and authors are available to editors and to researchers who can demonstrate appropriate safeguards and are not publicly released. A summary of findings and interactive visualizations are available at https://www.maxtopaz.com/citadel.

Appendix 2: Illustrative Examples of Suspected Fabricated References

Among the suspected fabricated references in our dataset, a recurring pattern involved reference titles precisely tailored to the citing paper's topic, but attached to identifiers (PMIDs and DOIs) belonging to entirely unrelated published studies. Each suspected fabricated reference assembled real components (a real journal name, a valid identifier from that journal, and plausible author names) into a citation that appeared credible on cursory inspection but could not be verified. The following three examples are drawn from different papers in our dataset.

Example A

Source paper: Impact of Risk Mitigation Strategies on Non-Fatal Injuries in the Construction Sector in Qatar: A Retrospective Analysis (*International Archives of Occupational and Environmental Health*, 2025) PMC11972227

Reference [14] is cited in the following context: "The increase in ICU admissions in the post-implementation group suggests that more individuals survived initial injuries and required intensive care, aligning with findings by Doe and Smith."

Cited title	"Impact of enhanced safety protocols on ICU admissions in the construction industry: A longitudinal analysis"
Cited authors	J Doe; R Smith
Cited journal/year	Journal of Occupational and Environmental Medicine (2023)
PMID 36730737 resolves to	"Predictors of Suicide and Differences in Attachment Styles and Resilience Among Treatment-Seeking First-Responder Subtypes"
DOI resolves to	"Occupational Balance and Depressive Symptoms During the COVID-19 Pandemic"

Note: The PMID and DOI each point to a different paper in the same journal. Neither has any connection to construction safety or ICU admissions. The claimed title does not correspond to any indexed publication.

Example B

Source paper: Biomarker-Guided Imaging and AI-Augmented Diagnosis of Degenerative Joint Disease (*Diagnostics*, 2025) PMC12154452

Reference [90] is cited in the following context: "MRI excels in soft tissue visualization and compositional assessment; CT offers superior bone detail and faster acquisition, with emerging dual-energy techniques adding material differentiation capabilities."

Cited title	"A Protocol for the Use of DMM/PTX-Induced Mouse Models of Osteoarthritis and Rheumatoid Arthritis"
Cited authors	E. Krustev; D. Rioux; J.J. McDougall
Cited journal/year	Current Protocols (2021)
PMID 34767311 resolves to	"Three-Dimensional Fruit Tissue Habitats for Culturing Caenorhabditis elegans"
DOI resolves to	Same paper (PMID and DOI are consistent)

Note: The claimed title combines two real methodologies into a protocol that has not been published. Both the PMID and DOI point to an unrelated paper about culturing nematode worms in fruit tissue. The claimed title does not correspond to any indexed publication.

Example C

Source paper: Nociplastic Pain: The Prominent Role of Non-Neuronal Cells in Central and Peripheral Sensitization (*Frontiers in Immunology*, 2026) PMC12886048

Reference [146] is cited in the following context: "Activation of P2X7 promotes astrocyte differentiation, and astrocytes can secrete inflammatory mediators, which further promote the activation of microglia and ultimately contribute to the development of chronic pain."

Cited title	"Microglial Modulation via Cannabinoid Receptor 2 Alleviates Fibromyalgia-Related Central Sensitization and Pain Hypersensitivity"
Cited authors	F. Chen; Y. Liu; H. Wang; X. Zhang; J. Li; K. Yang
Cited journal/year	Neuroscience (2023)
PMID 36813155 resolves to	"ChatGPT in Research: Balancing Ethics, Transparency and Advancement"
DOI resolves to	Same paper (PMID and DOI are consistent)

Note: The claimed title combines three neuroscience concepts into a plausible-sounding study. Both the PMID and DOI point to an editorial about ChatGPT ethics in research. The claimed title does not correspond to any indexed publication.

Appendix 3: Additional Analyses

1. Retraction and Correction Audit

All papers containing at least one suspected fabricated reference were cross-referenced against the Retraction Watch database (68,910 entries as of February 2026) and against publisher records to identify any post-publication action.

Table A3-1. Publisher actions recorded for the 2,810 papers containing at least one suspected fabricated reference, as of February 2026.

Publisher action	n (%)	Notes
No action of any kind	2,765 (98·4%)	No retraction, correction, or expression of concern
Retraction	2 (0·1%)	Both for data integrity concerns unrelated to references
Erratum	43 (1·5%)	None addressed fabricated references
Total	**2,810 (100%)**	

Of the two retracted papers, neither retraction notice mentioned reference integrity; both cited concerns about the underlying data. Of the 43 errata, none addressed the suspected fabricated references identified by CITADEL.

2. Publication Type Breakdown

Table A3-2. Fabrication rate per 10,000 papers by publication type. * Significantly elevated vs overall rate (p<0·0001, Bonferroni-corrected).

Publication type	Affected papers	Corpus total	Rate/10,000
Systematic review*	79	45,798	17·2
Review*	519	311,565	16·7
Observational study	30	24,886	12·1
Journal article	2,752	2,417,050	11·4
Letter	22	22,099	10·0
Case report	139	139,968	9·9
Meta-analysis	24	25,741	9·3
RCT	28	31,274	9·0
Editorial	30	41,218	7·3
Comment	5	12,701	3·9
Overall	**2,810**	**2,471,758**	**11·4**

3. Distribution of Fabricated References per Affected Paper

Among all papers containing at least one suspected fabricated reference, the median number was 1 (IQR 1–1; mean 1·44; maximum 24).

Table A3-3. Distribution of suspected fabricated references per affected paper (n = 2,810).

Fabricated refs per paper	Papers, n	%
1	2,211	78·7%
2	353	12·6%

3	115	4·1%
4	55	2·0%
5	33	1·2%
6–10	33	1·2%
11 or more	10	0·4%
Total	**2,810**	**100%**