Fabricated citations: an audit across 2·5 million biomedical papers

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] To our knowledge, no systematic audit of reference integrity across the biomedical literature has 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 2–4). Precision of our automated reference verification system was 91% (Fleiss' κ=0·71, indicating moderate agreement in about seven of every ten 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 5–6). 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 7–8).

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.

See Online for appendix

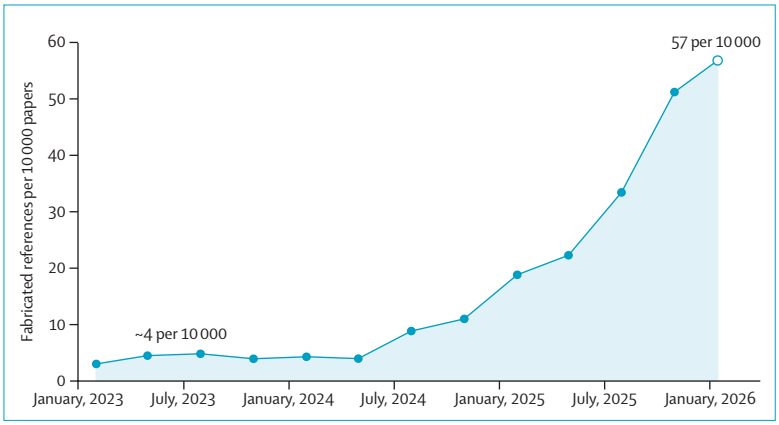

Figure: **Quarterly rate of fabricated references per 10 000 papers in PubMed Central from January, 2023, to February, 2026**
The fabrication rate remained stable at approximately four per 10 000 papers throughout 2023 (blue line). Beginning in mid-2024, the rate rose sharply, reaching approximately 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.

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.

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 had full access to the data in the study and had final responsibility for the decision to submit for publication. We declare no competing interests. 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.

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

Editorial note: The Lancet Group takes a neutral position with respect to territorial claims in published maps and institutional affiliations.

**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)

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/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.

The pharyngeal sanctuary: a challenge for zoliflodacin

The phase 3 trial by Alison Luckey and colleagues[1] validating zoliflodacin marks a pivotal moment in gonorrhoea management. As a novel agent targeting DNA GyrB subunit, zoliflodacin bypasses the GyrA-mediated resistance mechanisms that compromise fluoroquinolones, offering a viable option for use against multidrug-resistant strains.[1,2]

However, the trial data reveal important nuances regarding clinical efficacy. Although statistical non-inferiority was achieved, zoliflodacin's urogenital cure rate was 5·3% lower than the comparator (90·9% *vs* 96·2%).[1] This absolute deficit, although statistically acceptable, implies a clinical trade-off: five additional failures per 100 treated patients. This concern is amplified when considering extragenital sites. Microbiological cure rates for pharyngeal infections were notably suboptimal for both zoliflodacin (42 [79·2%] of 53 patients) and the ceftriaxone–azithromycin comparator (22 [78·6%] of 28 patients) in the intention-to-treat population.[1] The fact that neither regimen met even the 90% cure rate assumed in the study protocol underscores the pharynx as a place where antibiotic penetration is frequently insufficient to eradicate bacterial reservoirs.[1]

The persistence of *Neisseria gonorrhoeae* in the pharynx due to subtherapeutic drug exposure is a known driver for the selection of antimicrobial resistance.[3] Introducing a novel antibiotic class into an anatomical site where it is ineffective in approximately 20% of cases (a rate well below the efficacy required to suppress mutation selection) creates a high-risk environment for the emergence of zoliflodacin resistance.

Therefore, zoliflodacin should not be considered a monotherapy suitable for all anatomical sites without further guidance. Given the failure rates observed in this trial, it could be argued that if zoliflodacin is used for suspected or confirmed pharyngeal gonorrhoea, it must be accompanied by a mandatory test of cure at 7–14 days. Without rigorous follow-up to detect and manage these failures, this novel agent's use could be at risk of being squandered shortly after its introduction.

I declare no competing interests.

During the preparation of this work, I used Google Gemini to assist with grammar and punctuation. After using this tool, I reviewed and edited the content as needed and take full responsibility for the content of the publication.

Jian-Ying Wang
ao4256@ntpc.gov.tw

New Taipei City Hospital, New Taipei City 241, Taiwan

1 Luckey A, Balasegaram M, Barbee LA, et al. Zoliflodacin versus ceftriaxone plus azithromycin for treatment of uncomplicated urogenital gonorrhoea: an international, randomised, controlled, open-label, phase 3, non-inferiority clinical trial. *Lancet* 2026; **407:** 147–60.
2 Oyardi O, Yilmaz FN, Dosler S. Efficacy of zoliflodacin, a spiropyrimidinetrione antibiotic, against gram-negative pathogens. *Curr Microbiol* 2024; **81:** 241.
3 Unemo M, Lahra MM, Escher M, et al. WHO global antimicrobial resistance surveillance for *Neisseria gonorrhoeae* 2017–18: a retrospective observational study. *Lancet Microbe* 2021; **2:** e627–36.

Zoliflodacin: non-inferior, but not equivalent?

We congratulate Alison Luckey and colleagues [1] of the phase 3 zoliflodacin trial on the successful evaluation of a novel oral antimicrobial agent against uncomplicated gonorrhoea. Novel agents with distinct pharmacodynamic targets, such as GyrB inhibition by zoliflodacin, are urgently needed.

The purpose of a non-inferiority margin is to preserve a clinically meaningful proportion of the established treatment effect.[2] In the trial,[1] the 12% non-inferiority margin was met for the primary microbiological endpoint (estimated difference 5·3%, 95% CI 1·4–8·6) and is also within the 10% margin recommended in regulatory guidance by the US Food and Drug Administration and European Medicines Agency.[3,4]

Despite demonstrating regulatory non-inferiority, zoliflodacin (90·9%) was statistically inferior to ceftriaxone plus azithromycin (96·2%) for microbiological success at the urogenital site. Furthermore, clinical cure was lower for zoliflodacin (82%) than for ceftriaxone plus azithromycin (88%; estimated difference 6·7%, 95% CI 0·7–11·9).

These results contrast with the EAGLE-1 trial of gepotidacin.[5] In a similar setting using the same comparator, outcomes were closely aligned between groups: microbiological success was 92·6% with gepotidacin and 91·2% with ceftriaxone plus azithromycin (estimated difference –0·1%, 95% CI –5·6 to 5·5).

Regarding the microbiologically evaluable population, no case of urogenital bacterial persistence was seen for ceftriaxone plus azithromycin in both trials[1,5] or for gepotidacin in EAGLE-1, whereas zoliflodacin was associated with 15 (3·2%) of 475 cases of microbiological failure.[1] Importantly, both the zoliflodacin and EAGLE-1 trials demonstrated the high efficacy of ceftriaxone plus azithromycin as the primary regimen,