Analysing discussion text and references of paper mill articles in high-impact factor cancer journals

Whamond A [1], Barnett AG [1, 2] and Byrne JA [1, 3]

1. School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, NSW, Australia
2. School of Public Health and Social Work, Queensland University of Technology, Queensland, Australia
3. NSW Health Statewide Biobank, NSW Health Pathology, Camperdown, NSW, Australia

Identifying potential paper mill articles by visually scanning article text could help readers avoid them. We conducted an exploratory, cross-sectional, descriptive analysis of discussion text and references in high-impact factor (IF) cancer journals, defined as IF≥7 in Clarivate categories 'oncology', 'biochemistry & molecular biology', or 'cell biology'. We filtered the Retraction Watch database for retraction reason “paper mill”, subject “cancer”, and journals on our high-IF list. For comparison article cohorts, we randomly sampled original cancer articles in (i) the same n=8 journals as retracted articles, and (ii) n=10 independent journals with high-IF maintained ≥20 years. We found 22 retracted paper mill articles published between 2016-06-21 and 2022-06-16. Both comparison groups (n=100 articles each) were restricted to articles published between 2016-01-01 and 2024-10-31. For each article, we calculated percentage of references first cited in each section. We then classified every discussion sentence as ‘background’, ‘summary’, ‘comparison’, ‘inference’, ‘limitation’, or ‘future direction’, and recorded whether sentences cited new references. Paper mill discussions had the highest percentages of background sentences that cited new references with a median (IQR) of 32% (20% to 43%); followed by articles from the same journals with 18% (11% to 25%); and independent journals 11% (8% to 17%). Recognising superficial and redundant “second introductions” in discussion sections could help readers quickly identify potentially problematic articles.