

Improving machine learning-based paper mill detection in cancer research using full text, image and citation data

Arjun Mukherjee¹

ar.junmukherjee2007@gmail.com

Carnegie Mellon University

Research Article

Keywords: paper mills, machine learning, research integrity, BERT, cancer research, retraction, natural language processing

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11097507/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background: Paper mills, organizations that create and sell authorship rights to fraudulent research papers, are detrimental to the integrity of academic research, especially cancer literature. A recent study leveraged a BERT-based classifier trained on a paper's title and abstract to identify paper mill products in cancer research, achieving an accuracy of 91% on internal validation and 93% on external validation. This paper investigates whether incorporating a paper's full text, figure images and references alongside its abstract and title can enhance the performance of this detection pipeline.

Methods: A dataset was created combining known paper mill publications from Retraction Watch and PubPeer with control papers extracted from high impact journals and a stratified sample of countries, selected to account for linguistic biases. Every paper's title, abstract, figure images, citations and full text (Methods, Results and figure captions) were gathered. Using this dataset, four signals were developed: a BERT-based classifier trained on title and abstracts (mirroring the original study), a BERT-based classifier extending this by integrating full text, a citation signal measuring references to known paper mill articles and a CLIP-based figure image classifier. These signals were all merged using logistic regression, creating a final fused model.

Results: The abstract-only model performed similarly to the original study's model, attaining accuracies of 93.4% and 91.5% on internal and external validation respectively. The full text model showed some improvement upon this, boosting internal validation accuracy to 95.5% and external validation accuracy to 92.1%. Compared to the text-based models, the image classifier was less reliable, yielding an accuracy of 83.6% on internal validation and 76.4% on external validation. The fused model demonstrated the best performance out of all the models, achieving a 98.2% accuracy on internal validation and 96.4% on external validation.

Conclusions: Supplementing title and abstract text with image, full-text and citation data can significantly boost the performance of machine learning-based paper mill detection. Paper mills are poisoning the credibility of academic literature and will likely continue to become larger and more evasive. Therefore, developing more comprehensive and reliable ways to detect their work is very important.

Background

Research paper mills are "fraudulent organizations that make money by writing fake manuscripts and offering authorship slots for sale to academic customers" [1]. These practices were first exposed at a large scale in 2013 when a Science investigative article discovered a Chinese brokerage company selling authorship credit to scientific manuscripts [2]. Since then, many more companies have been caught selling authorship rights to fraudulent research papers. In fact, a recent study estimates that 5.8% (107,800) of the 1.86 million biomedical publications released in 2023 were produced by paper mills [3]. This estimate is about 19 times greater than the 5,671 biomedical retractions made in 2023, demonstrating how many fraudulent research papers go unnoticed by publishers.

Undetected paper mill activity is harmful for numerous reasons. Fraudulent paper mill articles, especially in domains like oncology or pharmacology, can affect medical practices and put patients at significant risk [4]. Furthermore, the fabricated data included in these manuscripts diminishes public trust in peer-reviewed journals and contaminates the knowledge base around a topic, slowing progress in research [5, 6]. For example, a scientist could attempt to build upon a study fabricated by a paper mill company which would be a major waste of time and money. Moreover, paper mills are able to produce and submit an extremely large number of deceitful papers, overwhelming journal reviewers and editors [7].

Cancer research is suspected to be particularly harmed by paper mill fraud compared to other scientific domains [8]. Since oncologists experience extreme pressure to publish research, they are natural targets for paper mills to sell fake authorship rights to [9]. In addition, the experimental results of cancer research are relatively easy to fabricate, allowing paper mills to churn out many fraudulent cancer-related papers [8]. Many of these faulty studies are cited numerous times, and their inaccurate results end up becoming ingrained in the field [10].

To combat the rise of paper mills, scientists have tried building systems to detect fake research papers. Cabanac et al. manually collected a dictionary of “tortured-phrases” which are awkward phrasings commonly used by the paraphrasing tools that paper mills utilize [11]. Another group of researchers created an algorithm to detect groups of journals that cite each other at excessively high rates [12]. Recently, researchers have started using machine learning techniques to identify paper mill products [13–15].

While most detection methods have focused on identifying fraudulent articles across many different areas of study [11–15], Scancar et al. decided to focus on detecting paper mill publications solely in cancer-related research [16]. Using Retraction Watch, an open-source database of retracted research papers [17], they collected the abstracts and titles of 2,202 paper mill publications. They then inputted that data into a BERT (bidirectional encoder representations from transformers) based text classification model to train it to pinpoint manuscripts created by paper mills. Their model achieved an accuracy of 91% and when it was deployed on real cancer literature, it flagged 10% of the 2.6 million articles it screened as resembling paper mill publications.

However, their approach of solely using the abstract text and title leaves many other potential indicators on the table. Paper mills have been shown to recycle experimental protocols and invalid reagent sequences [18]. Patterns like these would only be seen in the paper’s full text rather than the abstract. Research has also shown that paper mill publications often use inaccurate or duplicated images [19] and cite each other at disproportionately high rates [12].

This study aims to build on the research of Scancar et al. by testing whether integrating citation data, figure images and specific parts of the full text (Methods, Results and figure captions) could improve the performance of their title and abstract-based pipeline. Each modality was first evaluated on its own. Then, all three modalities (text, image, citations) were combined into one machine learning pipeline, and its performance was compared to the original abstract-based approach.

Methods

Paper Selection

To gather a dataset of paper mill and control publications, this study followed a methodology very similar to the one described by Scancar et al. [16]. Two sources of paper mill publications were used: manuscripts labeled as paper mill products in the Retraction Watch database [17] and a list of papers flagged as paper mill articles by PubPeer integrity experts [20]. PubPeer is an open-source database that allows people in the scientific community to flag research integrity violations [21]. “Spreadsheet of spreadsheets” contains a list of paper mill publications reported by PubPeer users [20]. Once the data was collected, it was filtered to only include cancer-related papers by checking whether the title or abstract contained one of the original study’s 39 cancer-related keywords [16]. If an article appeared in both the Retraction Watch and PubPeer datasets, it was removed from the PubPeer dataset. The data from the Retraction Watch database was used for model training and internal validation while the data from the PubPeer users was solely used for external validation. After filtering, there were 2,477 publications in the model training set and 3,000 publications in the external validation set.

To help the model understand what a genuine cancer-related paper looks like, a control dataset for model training was built. Since it is almost impossible to validate that large samples of research papers aren’t fraudulent, the controls were selected from two groups that are highly unlikely to contain paper mill products: publications from high impact journals and publications from countries that are underrepresented in the Retraction Watch dataset (based on the country of the first author’s institution). For the first group, I used the same four journals that the original paper selected (*Cell*, *Cancer Cell*, *Molecular Cell*, and *The EMBO Journal*). Since collecting full-text and image data is much more difficult than gathering abstracts alone, I added two more well-accredited journals (*Oncogene* and *Nature Medicine*) to obtain a large enough sample. Due to the large portion of Chinese-authored articles in the paper mill dataset, 664 articles (8.7%) from Chinese institutions published in the six high-impact journals were added to the control to account for differences in the ways Chinese authors write. 3,225 (42.3%) top cited Taiwanese cancer-related papers were added to the control using OpenAlex [22] because Taiwanese people also speak Mandarin, but Taiwan is underrepresented in the Retraction Watch database. Moreover, 854 (11.2%) cancer-related papers from Swedish, Norwegian and Finnish institutions were added to the control since none of these countries had a single paper mill retraction in the Retraction Watch database. The remaining 37.8% of the control came from a random sample of cancer research publications from the six high-impact journals excluding publications from China, Sweden, Norway, Finland and Taiwan. All articles in both the paper mill and control datasets were filtered to only include articles published from 1999–2026. The control dataset for model training contained 7,623 articles. For external validation, I supplemented the 3,000 suspected paper mill articles from PubPeer with random samples of 1,500 papers from the high-impact journals mentioned above and 1,500 papers from institutions in Norway, Finland and Sweden.

For both the control and the paper mill datasets, the following data was collected using OpenAlex and PubMed: journal name, title, abstract, first author's country and publication year.

Full text extraction

To obtain each paper's full body of text, three approaches were tried. The first approach was querying PubMed Central's full text archive, which returns an XML file of the article's full text [23]. If the article wasn't in the PubMed Central database, I tried acquiring its full text from the Europe PMC database, which also returns an XML file of the complete text [24]. If that failed, the publication's full PDF was collected and the python library PyMuPDF was used to extract its text. Once the paper's full text was obtained, only its figure captions, Methods and Results were kept. This is because these sections are most likely to include documented paper mill irregularities like recycled experimental protocols, inaccurate reagent sequences, and fabricated data [18]. The Introduction and Discussion sections are often written in a formulaic manner and likely don't provide much insight into whether an article is fraudulent or not.

Image extraction

Figures were acquired with a similar fallback system as the one used to extract the full text. First, Europe PMC's supplementary files service was queried using each paper's PubMed Central Identification number [24]. If the article wasn't found in that database, I queried NCBI's PMC Open Access Subset to obtain its figure images [25]. For papers that weren't included in either archive, the PDF was downloaded, and its images were extracted using PyMuPDF. Since these databases included images of mini-icons, lines and symbols, images with an aspect ratio of greater than 8:1 and images with a dimension less than 100 pixels were removed. Furthermore, when publishers included duplicates of the same image with different resolutions, only the higher quality image was retained. If none of these three methodologies could retrieve the images of a paper, it was not included in the training dataset for the image classifier.

Citation Data Collection

Using Retraction Watch and PubPeer, a list of 27,216 known paper mill publications was compiled. Then, the full reference list of each article in the model training and external validation datasets was gathered via OpenAlex. The fraction of each paper's references that cite known paper mill articles was calculated using these datasets. Prior work has shown that paper mill publications cite each other at disproportionately high rates [12]. Thus, the proportion of a paper's references that point to paper mill articles could be a useful signal for paper mill identification.

Model Training

The detection of paper mill publications was treated as a binary classification problem where each paper was either labeled as a paper mill paper or an authentic control paper. Five machine learning models were constructed for this task: an abstract-only baseline, a full-text classifier, an image classifier, and two fused models that combine textual and image signals with the citation data.

Abstract-based Classifier

To begin, a model solely trained on a paper's title and abstract was built to replicate the original study's work and obtain a baseline to assess the added value of the other components. The methodology for training this

model mimics Scancar et al. [16]. BERT (Bidirectional Encoder Representations from Transformers) was selected to match the model used in the original study. It is well suited for this task because its bidirectional architecture helps it capture the sophisticated details in scientific writing. The public BERT base uncased model was fine-tuned using this study's paper mill dataset with a new classification layer specialized for this task. The title and abstract of each paper were split into individual sentences because BERT has an input limit of 512 tokens. Every sentence was labeled as either paper mill or authentic and was fed into the BERT model for training. The model's predictions at the sentence level were aggregated to determine the model's assessment of the full article.

To fine-tune the model, the 2,477 cancer-related paper mill articles in the dataset were combined with 2,477 randomly selected control cancer-related articles. These publications were split into 70% for training (3,468), 17.5% for optimization (866 papers), and 12.5% for internal validation (620 papers), each evenly divided between paper mill and control. A random sample of 500 suspected paper mill papers from PubPeer and 500 control papers were used for external validation. Instead of using a simple probability threshold of 0.5, a cutoff was selected on the optimization step to prioritize specificity and avoid false accusations of fraud.

The model was trained on a single Apple Silicon machine using the Hugging Face Transformers library for BERT fine-tuning. Because of limited access to computing power compared to the original study, the BERT model's training process was made simpler and less computationally intensive. Rather than running hyperparameter optimization over 150 trials, the model was fine-tuned using a fixed configuration of 3 epochs, a learning rate of $2e-5$ and a batch size of 32.

Full Text Classifier

A second BERT model was trained with the exact same methodology as above, except each paper's Methods, Results and figure captions were inputted along with their abstract and title. The title and abstracts of papers whose full text couldn't be acquired were still included in the model training process. Thus, the exact same training, optimization, internal validation and external validation sets were used. This allows us to directly assess how including the full text influences model performance since the only factor that was changed was the additional text.

Image Classifier

Each paper's figure images were initially embedded using a pre-trained CLIP model. These embeddings were averaged out, resulting in a single vector per paper labeled with whether the paper was fraudulent or not. A logistic regression classifier, implemented via scikit-learn, was then trained to predict whether a paper was fraudulent or not based on its vector. The pool inputted into the model included all 1,711 paper mill papers with an image extracted and an equal amount of control papers with images that were drawn to match the original country-based stratification design. 70% of this dataset was used for model training, 17.5% was used for optimization and 12.5% was used for internal validation. Similar to the text classifiers, a cutoff was selected to prioritize specificity and mitigate false positives. The image data from 404 paper mill papers and 404 control papers was used for external validation.

Fusion Model

A fusion model was built to combine all three modalities (text, images and citations) into one classifier. This was implemented using a logistic regression model (created with scikit-learn) trained on each paper's text classifier probability, image classifier probability and its paper mill citation rate. An approach that trained a logistic regression model on raw features (BERT and CLIP embeddings) instead of the individual text and image classifier output probabilities was also tried but the accuracy was similar, so the simpler approach of using each model's output probability was chosen. Training the fusion model requires all three modalities to be available for each paper in the training data. Thus, the dataset used to fit the model was shrunk to solely include papers that had image, full-text and citation data. There were initially 3,764 papers that fit this criteria but the amount decreased after down sampling to get an equal amount of control and paper mill papers. The dataset was split 70% / 17.5% / 12.5% into training, optimization and internal validation. As with the other classifiers, a probability threshold was selected to minimize false accusations of fraud. A pool of 364 control papers and 364 PubPeer-sourced paper mill papers was utilized for external validation. Two fusions models were tested: one that used the abstract-only model's score and one that used the full-text model's score.

Model Evaluation

Each model's performance was evaluated using accuracy, sensitivity and specificity, which are the same metrics the original study used. These metrics were reported for both the internal and external validation datasets. Just like the original paper, accuracy, sensitivity and specificity are reported as point estimates from one final model because they were calculated from fixed validation sets.

Results

Table 1 shows model performance metrics for each classifier. The abstract-only classifier achieved an accuracy of 93.39% on internal validation (n = 620, sensitivity 91.61%, specificity 95.16%) and 91.50% on external validation (n = 1,000, sensitivity 83.80%, specificity 99.20%). This roughly aligns with the original study's reported accuracy of 91% on internal validation and 93% on external validation. It's notable that in the original study, the model performed better on the external validation set while in this study, the model performed better on the internal validation. The full-text classifier was trained on the exact same data and achieved 95.48% accuracy on internal validation (n = 620, sensitivity 93.87%, specificity 97.10%) and 92.10% on external validation (n = 1,000, sensitivity 85.20%, specificity 99.00%). Thus, the full-text showed an improvement over the abstract-only classifier in both the internal and external validation sets. The image classifier performed worse than the text-based classifiers, reaching an accuracy of 83.64% on internal validation (n = 428, sensitivity 72.43%, specificity 94.86%) and 76.36% on external validation (n = 808, sensitivity 54.70%, specificity 98.02%). Its sensitivity was consistently much lower than its specificity, meaning that the classifier correctly identified control papers far more frequently than it correctly identified paper mill papers.

Table 1
Model performance summary.

Model	Split	n	Accuracy	Sensitivity	Specificity
Abstract-only text	Internal validation	620	93.39%	91.61%	95.16%
Abstract-only text	External validation	1000	91.50%	83.80%	99.20%
Full-text	Internal validation	620	95.48%	93.87%	97.10%
Full-text	External validation	1000	92.10%	85.20%	99.00%
Image	Internal validation	428	83.64%	72.43%	94.86%
Image	External validation	808	76.36%	54.70%	98.02%
Fusion (abstract+citation+image)	Internal validation	340	97.65%	99.41%	95.88%
Fusion (abstract+citation+image)	External validation	728	95.19%	92.31%	98.08%
Fusion (full text+citation+image)	Internal validation	340	98.24%	100.00%	96.47%
Fusion (full text+citation+image)	External validation	728	96.43%	95.33%	97.53%

The fused model that integrated the abstract-only classifier's score with citation and image scores attained an accuracy of 97.65% on internal validation (n = 340, sensitivity 99.41%, specificity 95.88%) and 95.19% on external validation (n = 728, sensitivity 92.31%, specificity 98.08%). The fused model that replaced the abstract-only model score with the full-text score achieved a higher accuracy on both validation sets. It reached an accuracy of 98.24% on internal validation (n = 340, sensitivity 100%, specificity 96.47%) and 96.43% on external validation (n = 728, sensitivity 95.33%, specificity 97.53%). Both fusion variants attained better accuracies compared to the models that relied on a single modality.

Table 2 displays the fused logistic regression model's coefficients. In both variants of the models, the text feature has the highest weighted coefficient (7.51 and 8.76), followed by the image feature (5.36 and 4.11). The citation feature was weighted very minimally with coefficients of 0.38 and 0.35.

Table 2
Logistic regression coefficients for each fusion variant.

Fusion variant	Text coefficient	Citation coefficient	Image coefficient
Abstract + citation + image	7.51	0.38	5.36
Full text + citation + image	8.76	0.35	4.11

Discussion

The primary takeaway from this experiment is that combining title and abstract data with more comprehensive textual and visual content enhances paper mill detection. In general, the classifier's performance improved as more modalities were integrated, although the degree of that improvement varied depending on the modalities added. When tested and trained on the same data, the full-text classifier outperformed the abstract-only classifier on both validation datasets. Furthermore, the fusion models that merged textual inputs with figure images and citation data attained higher accuracies than the single-modality classifiers. The full-text based fusion variant had the best performance of all five tested models, achieving the highest accuracy on both validation sets and showing significantly improved performance compared to the original paper's abstract-only approach [16].

The image-only classifier was much less reliable compared to the text-based models. It had the worst performance out of all five tested models, yielding an accuracy of 83.64% on internal validation and 76.36% on external validation. This indicates that a paper's textual content is a stronger and more reliable standalone signal for paper mill detection compared to figure images. In addition, the model experienced a decrease in accuracy between internal and external validation, primarily caused by a drop in sensitivity from 72.43% to 54.70%. This discrepancy between performance in internal and external validation suggests that the model was slightly overfitted towards patterns in the training data. Since a large portion of this study's controls came from six journals, the model may have learned specific patterns in the ways these journals format their images rather than genuine signals of image manipulation. Thus, a larger, more diverse training dataset could yield better results.

The fusion model's coefficients provide more insight into how valuable each signal is in detecting fraudulent publications. The model learned to give text the highest weight, with images contributing a smaller but still meaningful role, reinforcing the earlier discovery that a paper's text is a more valuable standalone signal for paper mill identification than figure images. A paper's fraudulent citation rate was given a very small weighting, suggesting that it provides little insight beyond what the image and textual signals already capture.

Limitations

This study's dataset of paper mill publications has limitations. Papers labeled as paper mills in the Retraction Watch database come from scientists manually interpreting a paper's retraction notice. Since there are no systematic rules for identifying a paper mill publication, this could have led to some inconsistency in the database. Moreover, the external validation set of PubPeer articles is only based on reports of research integrity violations rather than confirmed evidence of paper mill origin. It is also possible that many paper mill articles haven't been uncovered yet. In that case, this study's dataset might only represent a certain subsection of all paper mill papers.

The control dataset also has potential faults. Since it's very difficult to confirm that a paper isn't fraudulent, controls were drawn from groups that are very unlikely to have paper mill origin. Nevertheless, this doesn't guarantee that every paper in the dataset is authentic. Furthermore, the choice to include many articles from high impact journals could have created a high contrast between the paper mill and control datasets, mitigating the model's ability to identify more subtle cases of fraud.

For the image detection model, the images came from three different sources: Europe PMC's supplementary-files service, NCBI's PMC open-access archive, and direct extraction from a paper's own PDF. Extracting the image from the PDF could cause it to have different characteristics like lower resolution compared to images extracted directly from archives. This was accounted for by ensuring that the control and paper mill datasets had similar amounts of images from each source. However, there's still a chance that the model learned to pick up on the source-based image differences rather than the image content itself.

The citation feature is also limited because it relies on the dataset of paper mill articles collected from Retraction Watch and PubPeer. There are likely many more paper mill articles that have gone undetected and don't appear in these datasets. Therefore, the paper mill citation rate calculated for each paper underestimates the true rate at which these papers cite fraudulent articles. This likely contributed to the citation component of the fusion model being weighted so minimally.

Conclusions

Overall, this study demonstrates that supplementing a paper's title and abstract with figure images and full-text content significantly improves machine learning-based paper mill detection for cancer-related research papers. The highest-performing model, which combined full-text, citation and image data, reached 98.24% accuracy on internal validation and 96.43% accuracy on external validation, which is a meaningful improvement over the original study's abstract-only model (91% and 93% accuracy respectively). The results also indicate that a paper's text is the most valuable signal for paper mill detection, with figure images being a useful but less reliable predictor. Citation rates were weighted very minimally by the fusion model, though this can be attributed to a lack of adequate citation data. These findings are relevant because, as paper mill activity continues to grow, reliable detection methods are going to be important for protecting the integrity of academic research. And with paper mills becoming more sophisticated, detection systems that consider multiple modalities rather than one will become increasingly vital.

Abbreviations

BERT

Bidirectional Encoder Representations from Transformers

CLIP

Contrastive Language-Image Pre-training

DOI

Digital Object Identifier

NCBI

National Center for Biotechnology Information

PDF

Portable Document Format

PMC

PubMed Central

PMID
PubMed Identifier
XML
Extensible Markup Language

Declarations

Ethics approval and consent to participate:

Not Applicable

Consent for publication:

Not Applicable

Competing interests:

The author declares that he has no competing interests

Authors' contributions

AM designed the data collection pipeline, trained the machine learning models, examined the results and wrote the manuscript.

Funding:

Not applicable

Author Contribution

AM designed the data collection pipeline, trained the machine learning models, examined the results and wrote the manuscript.

Acknowledgements:

Not applicable

Authors' information: Not applicable

Data Availability

The dataset of all paper mill and control papers analyzed in this study, along with the source code for data collection, model training and model evaluation is available at the Github repository: <https://github.com/ArjunMukherjee1/Multimodal-Paper-Mill-Detection>. Due to copyright constraints, the abstract, full text and figure image data cannot be published, but the repository contains a script to reproduce this data from each article's DOI.

References

1. Parker L, Boughton S, Bero L, Byrne JA. Paper mill challenges: past, present, and future. *J Clin Epidemiol*. 2024;176:111549. 10.1016/j.jclinepi.2024.111549.
2. Hvistendahl M. China's publication bazaar. *Science*. 2013;342(6162):1035–9. 10.1126/science.342.6162.1035.
3. Sabel BA, Knaack E, Gigerenzer G, Bilc MI. Fake publications in biomedical science: red-flagging method indicates mass production. *Naunyn Schmiedebergs Arch Pharmacol*. 2026;399(2):2943–55. 10.1007/s00210-025-04275-9.
4. Cavallo J. How the proliferation of fraudulent scientific papers is threatening the integrity of cancer research. *The ASCO Post*. 2025 Nov 25. Available from: <https://ascopost.com/issues/november-25-2025/how-the-proliferation-of-fraudulent-scientific-papers-is-threatening-the-integrity-of-cancer-research/>. Accessed 25 July 2026.
5. Jian X, Yaeger LH. Combating fake science in the age of generative artificial intelligence: a biomedical perspective. *Mo Med*. 2025;122(3):162–8.
6. Mainous AG 3. Papermills as another challenge to research integrity and trust in science. *Front Med (Lausanne)*. 2025;12:1557024. 10.3389/fmed.2025.1557024.
7. Basu M. Paper mill cancer studies get double the number of citations as genuine papers. *Nature*. 2026;655:552–3. 10.1038/d41586-026-01908-8.
8. Byrne JA, Park Y, Richardson RAK, Pathmendra P, Sun M, Stoeger T. Protection of the human gene research literature from contract cheating organizations known as research paper mills. *Nucleic Acids Res*. 2022;50(21):12058–70. 10.1093/nar/gkac1139.
9. Lambertini M. The effect of 'publish or perish' on young academic oncologists. *ESMO Dly Report*. 2022 Sep 5.
10. Bozzo A, Bali K, Evaniew N, Ghert M. Retractions in cancer research: a systematic survey. *Res Integr Peer Rev*. 2017;2:5. 10.1186/s41073-017-0031-1.
11. Cabanac G, Labbé C, Magazinov A. Tortured phrases: a dubious writing style emerging in science. *arXiv*. 2021;2107.06751.
12. Kojaku S, Livan G, Masuda N. Detecting anomalous citation groups in journal networks. *Sci Rep*. 2021;11:14524. 10.1038/s41598-021-93572-3.
13. Lambertini M, ESMO Daily Reporter. The effect of 'publish or perish' on young academic oncologists. 2022 Sep 5. Available from: <https://dailyreporter.esmo.org/esmo-congress-2022/opinions/the-effect-of-publish-or-perish-on-young-academic-oncologists>. Accessed 28 July 2026.

14. Fletcher AHA, Stevenson M. Predicting retracted research: a dataset and machine learning approaches. *Res Integr Peer Rev.* 2025;10:9. 10.1186/s41073-025-00168-w.
15. Chen L, Fu S, Freedman G, Zor C, Martin G, Kinross J et al. Pub-Guard-LLM: detecting fraudulent biomedical articles with reliable explanations. *arXiv.* 2025;2502.15429.
16. Scancar B, Byrne JA, Causeur D, Barnett AG. Machine learning based screening of potential paper mill publications in cancer research: methodological and cross sectional study. *BMJ.* 2026;392:e087581. 10.1136/bmj-2025-087581.
17. The Retraction Watch Database [Internet]. New York: The Center for Scientific Integrity. Available from: <http://retractiondatabase.org/>. Accessed 28 July 2026.
18. Park Y, West RA, Pathmendra P, Favier B, Stoeger T, Capes-Davis A, et al. Identification of human gene research articles with wrongly identified nucleotide sequences. *Life Sci Alliance.* 2022;5(4):e202101203. 10.26508/lsa.202101203.
19. Bik EM, Casadevall A, Fang FC. The prevalence of inappropriate image duplication in biomedical research publications. *mBio.* 2016;7(3):e00809–16. 10.1128/mBio.00809-16.
20. Hoya Camphorifolia [PubPeer user]. Spreadsheet of spreadsheets [Internet]. Available from: <https://docs.google.com/spreadsheets/d/1zKxfaqug4ZhwHyGzslF38pFyC8xtU8lzmmOFMGYITDI/edit?gid=1473413779#gid=1473413779>. Accessed 27 July 2026.
21. PubPeer. the online journal club [Internet]. Available from: <https://PubPeer.com/>. Accessed 27 July 2026.
22. Priem J, Piwowar H, Orr R. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv.* 2022;2205.01833.
23. National Center for Biotechnology Information (NCBI). [Internet]. Bethesda (MD): National Library of Medicine (US). Available from: <https://www.ncbi.nlm.nih.gov/>. Accessed 4 August 2026.
24. The Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res.* 2015;43(D1):D1042–8. 10.1093/nar/gku1061.
25. National Center for Biotechnology Information (NCBI). PMC Open Access Subset [Internet]. Bethesda (MD): National Library of Medicine (US). Available from: <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>. Accessed 4 August 2026.