29 Retraction Analysis

R. Heller · Date: 2026-05-09 · Last changed: 2026-10-31

29.1 Learning objectives

After completing this chapter, you will be able to:

- Identify retracted works in OpenAlex using type and metadata filters
- Analyse retraction rates by year, journal, and field
- Examine citation patterns before and after retraction
- Discuss the limitations of bibliometric approaches to research integrity
- Explain why retracted papers continue to accumulate citations

29.2 Setup

```
library(tidyverse)
library(openalexR)
library(glue)
library(gt)

set.seed(20260509)

source(here::here("R", "api_helpers.R"))
source(here::here("R", "utils.R"))
source(here::here("R", "sci_palette.R"))
```

29.3 Conceptual background

Retraction is the formal mechanism by which the scholarly record corrects itself. A paper is retracted when it is found to contain fabricated data, serious errors, plagiarism, or ethical violations. In theory, retraction should halt the influence of flawed work. In practice, retracted papers continue to be cited — sometimes for years or decades after retraction.

The growth in retractions has been dramatic: from a few dozen per year in the 1990s to over 10,000 in recent years. This increase reflects both more misconduct detection (plagiarism software, image forensics) and more publications overall. Retraction rates vary by field: biomedical sciences have the highest absolute numbers, partly because of larger publication volumes and greater scrutiny.

Citation persistence is the most studied bibliometric aspect of retractions. Studies consistently find that retracted papers continue to accumulate positive citations (i.e., citations that treat the retracted work as valid) long after retraction. This occurs because: (1) citing authors may not be aware of the retraction; (2) database retraction notices are inconsistently displayed; (3) citations in already-published papers are not retroactively removed.

OpenAlex marks retracted works with a specific type or retraction flag, though coverage depends on publisher reporting. The Retraction Watch database provides the most comprehensive curated list of retractions but is not directly integrated into most bibliometric databases.

Bibliometric analysis of retractions must be approached carefully. The goal is to improve the self-correction mechanism, not to stigmatise individuals or institutions.

29.4 Worked example

29.4.1 Finding retracted works

```
retracted ← fetch_openalex(
  entity = "works",
  is_retracted = "true",
  from_publication_date = "2010-01-01",
  to_publication_date = "2023-12-31",
  options = list(sample = 200, seed = 42)
)

cat(glue("Retracted works retrieved: {nrow(retracted)}\n"))
```

```
#> Retracted works retrieved: 200
```

```
retraction_years ← retracted ▷
  mutate(year = year(publication_date)) ▷
  count(year, sort = FALSE)

retraction_years ▷ gt()
```

year	n
2010	9
2011	13
2012	3
2013	7
2014	12
2015	6
2016	6
2017	10
2018	10
2019	12
2020	19

year	n
2021	25
2022	41
2023	27

29.4.2 Citation counts of retracted papers

```
ggplot(retracted, aes(x = cited_by_count)) +
  geom_histogram(binwidth = 5, fill = palette_sci(1), colour = "white") +
  labs(x = "Citations received", y = "Number of retracted papers") +
  theme_sci()
```

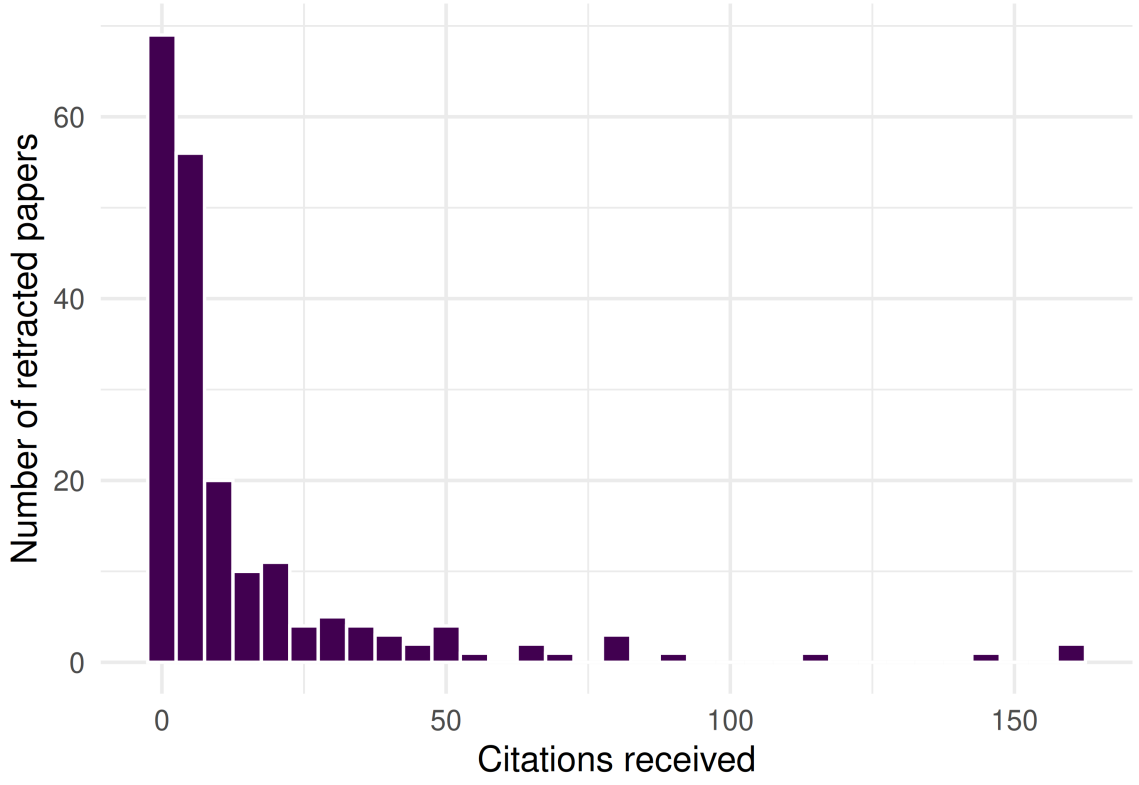


Figure 29.1: Citation count distribution of retracted papers.

```
cite_stats ← retracted ▷
  summarise(
    n = n(),
    mean_cites = round(mean(cited_by_count), 1),
    median_cites = median(cited_by_count),
    max_cites = max(cited_by_count),
    pct_uncited = scales::percent(mean(cited_by_count == 0))
  )

cite_stats ▷ gt()
```

n	mean_cites	median_cites	max_cites	pct_uncited
200	14.2	5	161	20%

29.4.3 **Retracted papers by source**

```
retracted ▷
  filter(!is.na(source_display_name)) ▷
  count(source_display_name, sort = TRUE) ▷
  head(10) ▷
  mutate(source_display_name = fct_reorder(source_display_name, n)) ▷
  ggplot(aes(x = n, y = source_display_name)) +
  geom_col(fill = palette_sci(1)) +
  labs(x = "Retracted papers", y = NULL) +
  theme_sci()
```

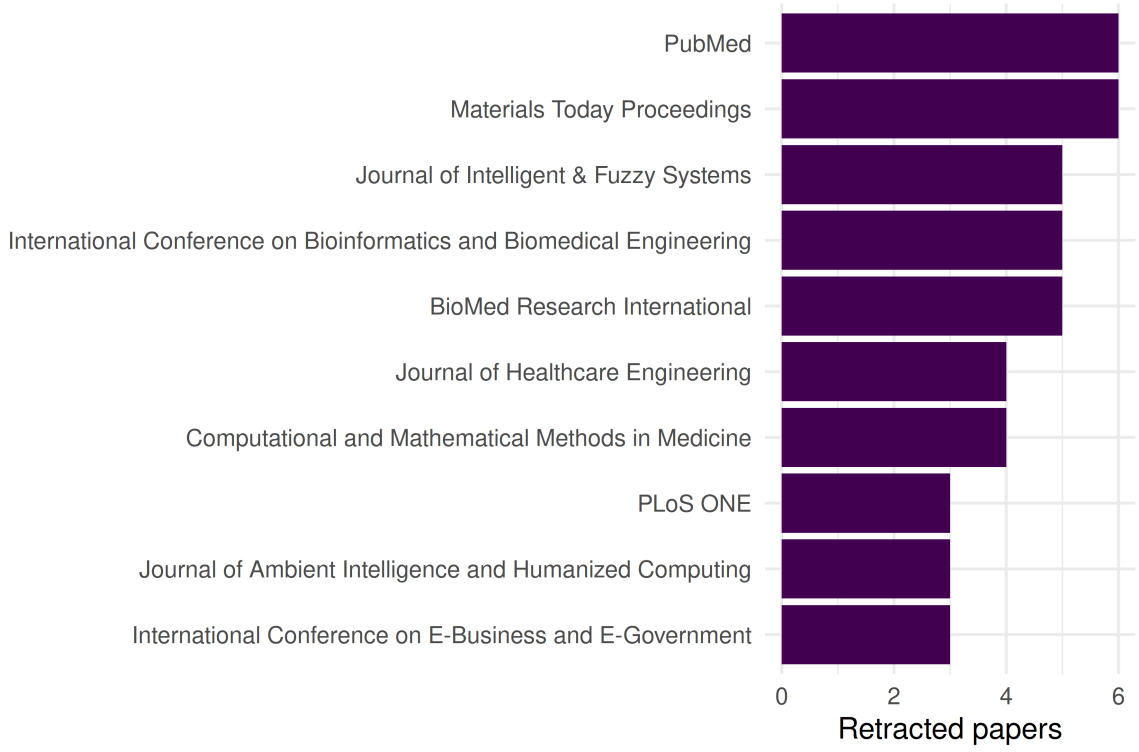


Figure 29.2: Top journals with the most retracted papers in the sample.

29.4.4 Citation persistence over time

```
retracted |>
  mutate(age = 2024 - year(publication_date)) |>
  filter(age >= 1, age <= 12) |>
  group_by(age) |>
  summarise(mean_cites = mean(cited_by_count), n = n(), .groups = "drop") |>
  ggplot(aes(x = age, y = mean_cites)) +
  geom_line(linewidth = 1, colour = palette_sci(1)) +
  geom_point(size = 2, colour = palette_sci(1)) +
  labs(x = "Years since publication", y = "Mean citations") +
  theme_sci()
```

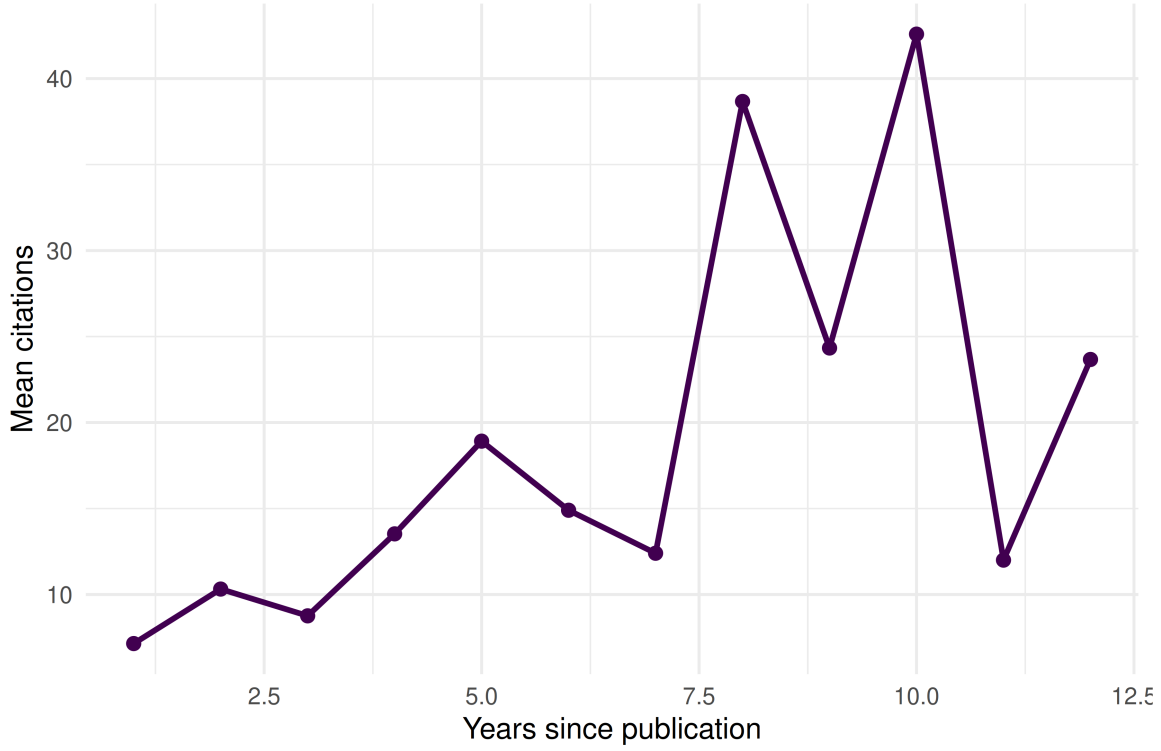


Figure 29.3: Mean citations by publication age for retracted papers.

29.5 Diagnostics and interpretation

- **Coverage**: OpenAlex retraction coverage depends on publisher metadata. Not all retractions are flagged. Cross-check with Retraction Watch for comprehensive coverage.
- **Retraction vs. publication date**: The publication date is when the paper was published, not when it was retracted. The gap between these dates varies from months to decades.
- **Positive vs. negative citations**: Citation counts do not distinguish between "this paper found X" (positive, problematic) and "this paper was retracted for Y" (negative, appropriate). Manual checking of citing contexts is needed.

- **Denominator**: Retraction *rates* require the total number of publications as denominator. Raw retraction counts favour large journals.

29.6 Limitations and responsible use

29.7 Limitations and responsible use

- **Retractions are not all equal.** Some retractions reflect honest error, others fraud. The reason for retraction matters for interpretation but is often not coded in bibliometric databases.
- **Stigmatisation risk.** Listing authors or institutions with high retraction counts can be stigmatising. Report aggregate patterns, not individual-level data, unless specifically studying misconduct.
- **Database lag.** Retraction notices may take months or years to propagate through databases. Recent retraction data is incomplete.
- **Self-correction is broader than retraction.** Corrections, expressions of concern, and errata also contribute to self-correction but are harder to track bibliometrically.
- **Do not equate retractions with research quality.** Fields with more scrutiny detect more problems. A high retraction rate may indicate effective oversight, not worse science (Hicks et al. 2015).

29.8 Common pitfalls

29.9 Common pitfalls

- **Counting retractions without denominators.** A journal with 10 retractions from 50,000 papers has a lower retraction rate than one with 5 from 1,000.
- **Assuming all post-retraction citations are problematic.** Some citing papers may specifically discuss the retraction or cite the paper for methodology that remains valid despite data issues.
- **Ignoring retraction reasons.** Honest errors should be distinguished from misconduct in any analysis of research integrity.
- **Using retraction analysis for individual evaluation.** Authors of retracted papers may be victims of co-author misconduct. Never use retraction counts as performance indicators.

29.10 Exercises

1. **Retraction rate by field.** Fetch retracted works from two different fields. Compare the retraction rate (retractions per 1,000 publications). Account for corpus size.
2. **Citation trajectory.** For a retracted paper with many citations, examine when the citations accumulated. Were most citations received before or after the retraction?
3. **Retraction-citation correlation.** Is there a relationship between a paper's citation count at the time of retraction and the number of post-retraction citations it receives?

29.11 Solutions

Solutions are provided in E.

29.12 Further reading

- Fang et al. (2012) — Large-scale analysis of retraction patterns and their causes.
- Hicks et al. (2015) — The Leiden Manifesto: responsible use of metrics in the context of research integrity.
- Garfield (1955) — Citation indexing: the infrastructure that makes retraction tracking possible.

29.13 Session info

```
sessionInfo()
```

```
#> R version 4.6.1 (2026-06-24)
#> Platform: x86_64-pc-linux-gnu
#> Running under: Ubuntu 24.04.5 LTS
#>
#> Matrix products: default
#> BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
#> LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.26.so;  LAPACK
version 3.12.0
#>
#> locale:
#>  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
#>  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
#>  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
#> [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
#>
#> time zone: UTC
#> tzcode source: system (glibc)
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#>  [1] uwot_0.2.5                Matrix_1.7-5
#>  [3] word2vec_0.4.1            stm_1.3.10
#>  [5] topicmodels_0.2-17        quanteda.textstats_0.97.2
#>  [7] visNetwork_2.1.4          ggraph_2.2.2
#>  [9] tidygraph_1.3.1           igraph_2.3.4
#> [11] quanteda_4.5.1            pdftools_3.9.1
#> [13] arrow_26.0.1.1            bibliometrix_5.5.0
#> [15] RefManageR_1.4.0          bib2df_1.1.2.0
#> [17] rcrossref_1.2.1           gt_1.4.0
#> [19] tidytext_0.4.3            glue_1.8.1
#> [21] openalexR_3.1.2           lubridate_1.9.5
#> [23] forcats_1.0.1             stringr_1.6.0
#> [25] dplyr_1.2.1               purrr_1.2.2
#> [27] readr_2.2.0               tidyr_1.3.2
#> [29] tibble_3.3.1              ggplot2_4.0.4
#> [31] tidyverse_2.0.0
#>
#> loaded via a namespace (and not attached):
#>   [1] bibtex_0.5.3           RColorBrewer_1.1-3     rstudioapi_0.19.0
#>   [4] jsonlite_2.0.0         magrittr_2.0.5         modeltools_0.2-25
#>   [7] farver_2.1.2           rmarkdown_2.32         fs_2.1.0
#>  [10] vctrs_0.7.3            memoise_2.0.1          askpass_1.2.1
#>  [13] base64enc_0.1-6        htmltools_0.5.9        contentanalysis_1.1.1
#>  [16] curl_8.1.0             janeaustenr_1.0.0      cellranger_1.1.0
#>  [19] sass_0.5.0             bslib_0.12.0           htmlwidgets_1.6.4
#>  [22] tokenizers_0.3.0       plyr_1.8.9             httr2_1.3.0
#>  [25] plotly_4.12.1          cachem_1.1.0           dimensionsR_0.0.3
#>  [28] mime_0.13              lifecycle_1.0.5        pkgconfig_2.0.3
#>  [31] R6_2.6.1               fastmap_1.2.0          shiny_1.15.0
#>  [34] digest_0.6.39          patchwork_1.3.2        shinycssloaders_1.1.0
#>  [37] rprojroot_2.1.1        RSpectra_0.16-2        SnowballC_0.7.1
#>  [40] labeling_0.4.3         urltools_1.7.3.1       timechange_0.4.0
#>  [43] mgcv_1.9-4             polyclip_1.10-7        httr_1.4.9
#>  [46] compiler_4.6.1         here_1.0.2             bit64_4.8.6
#>  [49] withr_3.0.3            S7_1.0.0               viridis_0.6.5
#>  [52] ggforce_0.5.0          MASS_7.3-65            bibliometrixData_0.3.0
#>  [55] tools_4.6.1            otel_0.3.0             stopwords_2.3
#>  [58] zip_3.0.2              httpuv_1.6.17          rentrez_1.2.4
#>  [61] nlme_3.1-169           promises_1.5.0         grid_4.6.1
#>  [64] stringdist_0.9.17      reshape2_1.4.5         generics_0.1.4
#>  [67] gtable_0.3.6           tzdb_0.5.0             rscopus_0.9.0
```

```
#>  [70] ca_0.72                data.table_1.18.6.1    hms_1.1.4
#>  [73] xml2_1.6.0             utf8_1.2.6             ggrepel_0.9.8
#>  [76] pillar_1.11.1          nsyllable_1.0.1        vroom_1.7.1
#>  [79] later_1.4.8            splines_4.6.1          tweenr_2.0.3
#>  [82] lattice_0.22-9         FNN_1.1.4.1            bit_4.6.0
#>  [85] tidyselect_1.2.1       tm_0.7-20              miniUI_0.1.2
#>  [88] downlit_0.4.5          knitr_1.52             gridExtra_2.3.1
#>  [91] NLP_0.3-3              bookdown_0.48          stats4_4.6.1
#>  [94] crul_1.6.0             xfun_0.62              graphlayouts_1.2.5
#>  [97] matrixStats_1.5.0      DT_0.34.0              humaniformat_0.6.0
#> [100] stringi_1.8.9          qpdf_1.4.1             yaml_2.3.12
#> [103] evaluate_1.0.5         httpcode_0.3.0         cli_3.6.6
#> [106] xtable_1.8-8           jquerylib_0.1.4        Rcpp_1.1.2
#> [109] readxl_1.6.0           triebeard_0.4.1        XML_3.99-0.25
#> [112] parallel_4.6.1         assertthat_0.2.1       pubmedR_1.1.0
#> [115] slam_0.1-56            viridisLite_0.4.3      scales_1.4.0
#> [118] crayon_1.5.3           openxlsx_4.2.9         rlang_1.3.0
#> [121] fastmatch_1.1-8
```