Online Resource 1

What delisting-based exposure measures cannot see: country rankings, event artifacts, and the bibliographic visibility of Libya's national journals

Abdelmuhsen M. Abusneina. Department of Molecular Diagnostics, Faculty of Biomedical Sciences, University of Benghazi, Fourth Ring Road, Benghazi, Libya. abdelmuhsen.abusneina@uob.edu.ly

Scientometrics. Online Resource 1, 7 October 2026.

S1 Corpus and sampling frame

Fig. S1 traces the publication corpus from retrieval to the five venue tiers and the ranked states (panel a), and the frame of the visibility probe from the rows it returned to the journal rows analyzed (panel b). Counts are those reported in Sections 2.1, 2.3, and 2.7 of the main text.

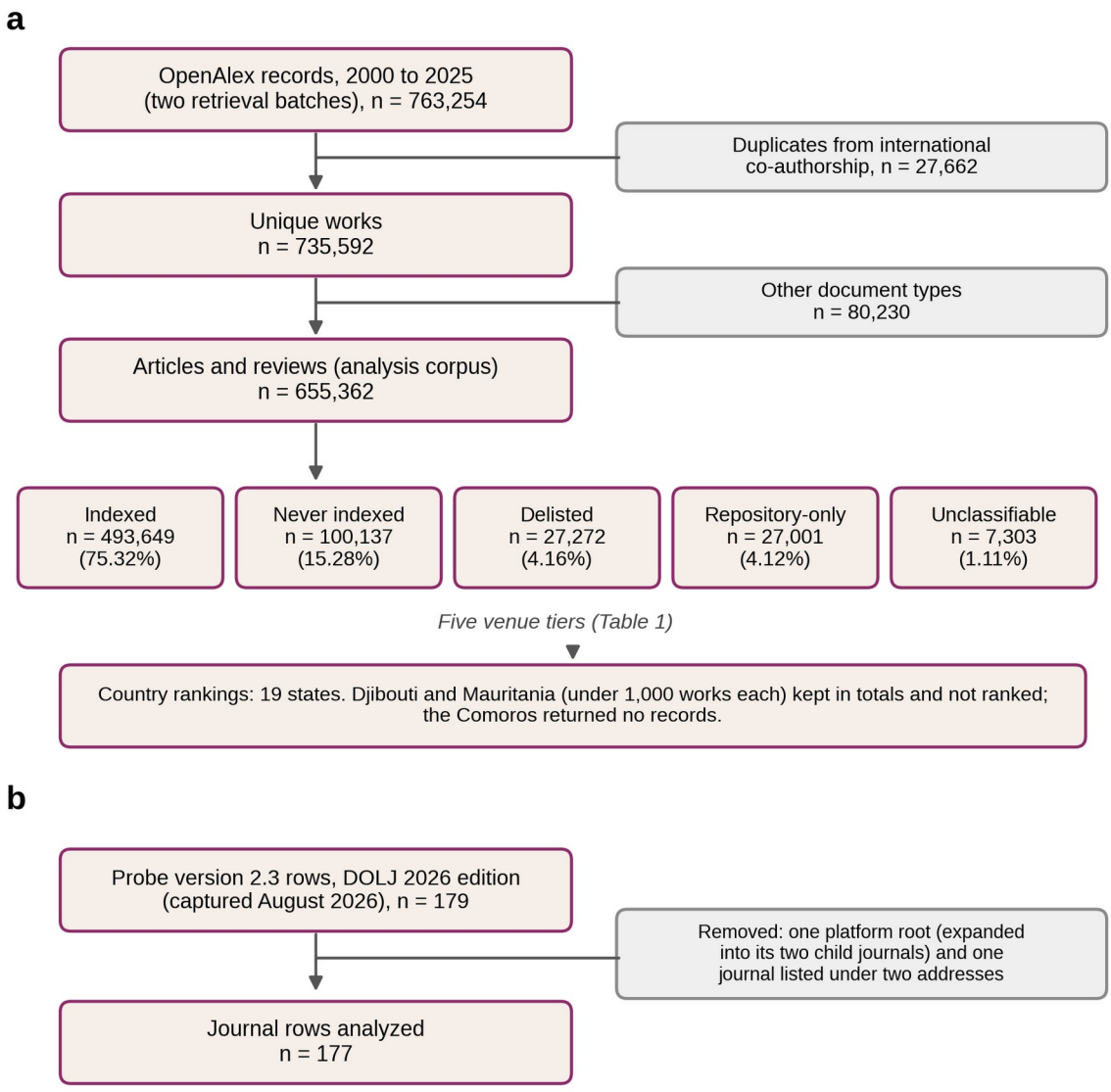


Fig. S1 Flow of records. (a) Publication corpus: OpenAlex records for 2000 to 2025 with at least one author in a League of Arab States member and a primary topic in health sciences, deduplicated, restricted to articles and

reviews, and assigned to five venue tiers. (b) Visibility probe frame: rows returned by probe version 2.3 for the addresses listed in the 2026 edition of the Directory of Online Libyan Journals, and the journal rows analyzed

S2 Study-design classification and its validation

S2.1 Classification rules

A rule-based classifier assigned each article or review with an abstract of at least 120 characters to one of five classes from three regular-expression lists, applied without regard to case. The high-ceiling list covers laboratory, animal, molecular, trial, cohort, and computational terms (for example in vitro, mice, PCR, ELISA, randomized controlled trial, prospective, in silico, molecular docking). The low-ceiling list covers record-based and survey terms (for example cross-sectional, questionnaire, retrospective, medical records, case report, case-control, interview). The synthesis list covers review terms (for example systematic review, meta-analysis, scoping review, bibliometric). The classes follow this order of precedence: synthesis when a synthesis term matched and no high-ceiling term did; high when only high-ceiling terms matched; low when only low-ceiling terms matched; mixed when both matched; and unresolved otherwise. The rule file is deposited with the data.

Applied to 655,362 articles and reviews, the rules gave the distribution in Table S1. Rerunning the deposited rule file on the deposited abstracts reproduced every share in the table.

Table S1 Distribution of rule-based design classes

Class	Articles and reviews, %
High resource ceiling	19.3
Low resource ceiling	21.4
Mixed	6.5
Synthesis	4.2
Unresolved	27.2
No usable abstract	21.5
Total (n = 655,362)	100.0

Among works with a usable abstract, the rules assigned 65.4% to a class other than unresolved. Shares sum to 100.1% because of rounding.

S2.2 Validation design

Sample. We drew a stratified random sample of 200 articles and reviews with abstracts of at least 40 words (seed 20260929). Within each region (Libya; the other states), we drew 40 abstracts that the rules labeled high, 40 labeled low, and 20 that the rules did not assign to either class. A Libyan co-authored work counted as Libyan. Region and rule label were withheld from every labeler; each labeler saw only an identifier, the title, and the abstract, in random order.

Codebook. The codebook defined the construct as whether a study could be carried out with records, a questionnaire, and a clinic, or needed further infrastructure. It offered five labels: high (trials, prospective cohorts or follow-up, laboratory work beyond routine diagnostics, imaging cohorts), low (record reviews, analyses of routine hospital or laboratory records, questionnaire surveys, case reports and series, audits, qualitative interviews), computational (computer analysis of existing or public data only), synthesis, and cannot tell. Studies with a low-ceiling framework and laboratory measurements beyond routine diagnostics were to be labeled high. The codebook was fixed before any labeler saw an abstract.

Labelers. An AI model (Claude, Anthropic) labeled all 200 abstracts. Its labels were saved and fixed with a SHA-256 checksum before the author began labeling. The author labeled a random subset of 50 (seed 20260930; 25 Libyan and 25 from the other states, balanced across the rule strata), with a confidence rating of sure, unsure, or outside my field. Neither labeler saw the other's labels or the key until both sets were complete.

Decision rule. Before labeling we fixed the following rule. A three-class Cohen's kappa (high, low, other) of at least 0.61 with precision of at least 80% for both rule classes would keep the analysis in the main text; a kappa from 0.41 to 0.60 would move it to supplementary material with a caveat; a kappa below 0.41 would remove the design claims. The main analysis counts computational labels as other, and a sensitivity analysis counts them as high, as the rules do.

Statistics. Confidence intervals for kappa are percentile intervals from 2,000 bootstrap resamples of abstracts; intervals for precision are Wilson score intervals.

S2.3 Results

Table S2 Agreement between the rules and independent labels

Comparison	n	Three-class kappa (95% CI)	Precision, rule high, % (95% CI)	Precision, rule low, % (95% CI)	Not assigned by rules, placed high or low, % (95% CI)
Rules vs. author	50	0.43 (0.22 to 0.62)	75 (53 to 89)	70 (48 to 85)	70 (40 to 89)
Rules vs. AI labels	200	0.55 (0.45 to 0.64)	71 (61 to 80)	91 (83 to 96)	68 (52 to 80)
Rules vs. AI labels, Libya	100	0.48 (0.35 to 0.61)	57 (42 to 71)	95 (83 to 99)	70 (48 to 85)
Rules vs. AI labels, other states	100	0.62 (0.49 to 0.74)	85 (71 to 93)	88 (74 to 95)	65 (43 to 82)
Author vs. AI labels	50	0.84 (0.68 to 0.97)			

Precision is the share of abstracts in a rule class that the labeler placed in the same class. Five-class kappa between the author and the AI labels was 0.84 (0.70 to 0.97); the two agreed on 45 of 50 abstracts. Restricted to the 39 abstracts the author rated sure, kappa between the rules and the author was 0.51 (0.27 to 0.72). Counting computational labels as high gave kappa 0.46 (0.25 to 0.65) for the rules against the author and 0.54 (0.45 to 0.64) against the AI labels. Denominators: 20, 20, and 10 abstracts for the author comparison; 80, 80, and 40 for all AI labels; half of each for each region.

Both comparisons with the rules fall in the band of 0.41 to 0.60 that we fixed in advance for supplementary reporting. The two sets of independent labels agreed closely with each other. All five disagreements between the author and the AI labels concerned abstracts that reported routine laboratory tests (for example serum uric acid, a lipid profile, or antibiotic susceptibility of clinical isolates): the author labeled them high and the AI labels low. That pattern points to the codebook's phrase "beyond routine diagnostics", which the two labelers read differently.

The rules erred most where the main text needed them most (Table S3). Of the 40 Libyan abstracts that the rules labeled high, the AI labels placed 9 among syntheses and 6 among low-ceiling studies. Reviews that mention laboratory or animal work satisfied the high-ceiling list, and the precedence order then let the high label override the synthesis label.

Table S3 Rule labels against AI labels by region (n = 200)

Region	Rule label	High	Low	Synthesis	Computational	Cannot tell
Libya	High	23	6	9	2	0
	Low	1	38	1	0	0

	Not assigned	7	7	5	1	0
Other states	High	34	2	3	0	1
	Low	2	35	0	2	1
	Not assigned	4	9	6	1	0

Each row sums to 40 (rule high or low) or 20 (not assigned).

S2.4 Descriptive results from the classification

We report the following figures because earlier drafts cited them; the validation above limits what they can show, and none enters the main text. All were computed before the common ISSN crosswalk. Among Libyan articles that the rules classified, the low-ceiling share was 55.3% in 2000 to 2010, 53.6% in 2011 to 2014, 51.9% in 2015 to 2019, and 64.6% in 2020 to 2025. Within several countries, low-ceiling work appeared in never-indexed venues more often than high-ceiling work: Morocco 36.7% against 12.7%, Tunisia 10.8% against 4.8%, Saudi Arabia 13.7% against 6.4%, and Jordan 9.6% against 5.1%. In Libya the corresponding figures were 42.0% for low-ceiling and 34.3% for high-ceiling work.

The Libyan contrast is the least reliable of these. With rule-high precision near 57% for Libyan abstracts and many reviews in the rule-high class, the Libyan high-ceiling group mixes primary studies with reviews, and the comparison cannot show whether study design explains venue choice in Libya.

S2.5 What a usable classification would require

Two changes follow from the disagreements: synthesis terms should take precedence over high-ceiling terms, and the codebook and rules should state how routine laboratory tests are treated. A revised classifier would need validation on a fresh sample, since tuning the rules on these 200 abstracts and testing them on the same abstracts would overstate agreement.

S3 Country fixed-effects model

Purpose. Section 3.4 compares Libya’s rise in the never-indexed share with the rise in other groups of states. A reviewer may ask whether the comparison survives fixed differences between states and shocks that hit all states in a period, and whether the large output of Egypt and Saudi Arabia drives it. The model below addresses both. It is descriptive and supports no causal reading.

Model. For each of the 19 ranked states and four periods (2000 to 2010, 2011 to 2014, 2015 to 2019, 2020 to 2025) we computed the percentage of the state’s articles and reviews in never-indexed venues (76 cells; a work with authors in several states counts in each). We fitted share = state effect + period effect + sum over later periods of (Libya effect in that period), by weighted least squares with weights equal to the number of works in the cell. Each Libya effect is the difference, in percentage points, between Libya’s change since 2000 to 2010 and the weighted change in the other states. Because 19 states are too few for cluster-robust standard errors, we judged Libya by a placebo test: we repeated the estimation giving each state in turn the same treatment and counted how many states had an estimate at least as large as Libya’s. With 19 states the smallest attainable p value is 1/19 = 0.053 (1/17 = 0.059 when Egypt and Saudi Arabia are dropped).

Table S4 Libya’s country-fixed-effects estimates for the never-indexed share, relative to 2000 to 2010

Specification	Period	Libya effect, percentage points	Median of other states	Libya rank (p)
Weighted by works, 19 states	2011 to 2014	8.6	0.0	1 of 19 (0.053)

	2015 to 2019	24.3	−1.6	1 of 19 (0.053)
	2020 to 2025	24.0	−1.6	1 of 19 (0.053)
Unweighted, 19 states	2011 to 2014	8.6	−0.4	1 of 19 (0.053)
	2015 to 2019	24.4	−2.8	1 of 19 (0.053)
	2020 to 2025	22.5	−4.3	1 of 19 (0.053)
Weighted, without Egypt and Saudi Arabia	2011 to 2014	9.7	0.6	1 of 17 (0.059)
	2015 to 2019	27.0	1.1	1 of 17 (0.059)
	2020 to 2025	23.3	−2.5	1 of 17 (0.059)
Unweighted, without Egypt and Saudi Arabia	2011 to 2014	8.7	−0.9	1 of 17 (0.059)
	2015 to 2019	24.6	−2.7	1 of 17 (0.059)
	2020 to 2025	22.2	−4.8	1 of 17 (0.059)

Libya's estimate is the largest of all states in every period and specification. In the weighted model for 2020 to 2025, the next largest estimates are Somalia (18.4 points), Morocco (17.5), and Yemen (10.8), and the median of the other 18 states is −1.6. The estimates for 2015 to 2019 and 2020 to 2025 are close to each other (24.3 and 24.0), which matches the plateau in Table 4. Libya's cells are small in 2000 to 2010 (451 works) and 2011 to 2014 (397), so the early estimates carry more sampling noise than the later ones (1,122 and 4,070 works).

Limits. The placebo p value reflects only how many states resemble Libya; it ignores sampling error within cells and the dependence between states created by shared co-authored works. Any state-specific event in the same years, such as the expansion of domestic journals or changes in OpenAlex coverage, is part of the Libya effect. The files fe_panel_country_period.csv and fe_model.py reproduce every number in this section and are deposited with the data.

S4 Venue tiers by state

Table S5 gives the five-tier distribution of each state's articles and reviews. The unclassifiable share ranges from 0.51% (Djibouti) to 6.34% (Algeria), and the repository-only share from 2.65% (Egypt) to 18.78% (Djibouti). Recomputing the four exposure measures over ISSN-bearing works only moves no ranked state by more than one place (Section 3.1 of the main text).

Table S5 Five-tier distribution of articles and reviews by state, % (2000 to 2025)

State	n	Indexed	Never indexed	Delisted	Repository-only	Unclassifiable
Algeria	17,883	75.43	11.80	1.58	4.84	6.34
Bahrain	6,421	72.37	14.76	8.86	3.24	0.76
Djibouti	197	68.02	10.15	2.54	18.78	0.51
Egypt	197,867	73.55	19.39	3.89	2.65	0.53
Iraq	51,453	70.59	20.80	2.66	5.18	0.76
Jordan	28,021	83.81	8.84	3.16	2.99	1.21
Kuwait	16,089	77.53	12.87	4.19	4.52	0.89
Lebanon	27,340	85.94	6.81	2.43	4.28	0.54
Libya	6,040	54.16	33.43	4.87	6.13	1.42
Mauritania	514	78.99	10.31	5.25	4.28	1.17
Morocco	55,104	69.06	18.45	4.34	6.17	1.98
Oman	12,597	81.04	9.60	2.80	5.72	0.84
Palestine	7,670	76.04	13.83	5.36	3.87	0.90
Qatar	32,942	80.06	10.31	4.62	3.84	1.17
Saudi Arabia	152,552	78.35	11.12	6.27	3.54	0.71
Somalia	4,765	56.35	31.37	6.02	4.97	1.28
Sudan	12,496	64.21	22.98	6.20	4.83	1.78
Syria	6,412	80.12	10.73	4.48	4.12	0.56
Tunisia	50,678	84.72	7.44	0.74	5.62	1.47
United Arab Emirates	34,505	78.69	11.56	5.89	2.73	1.13
Yemen	6,485	67.12	22.79	3.62	4.18	2.28

Rows sum to 100% within rounding. n is the country-level count: a work with authors in several states counts once for each state. Djibouti and Mauritania have fewer than 1,000 works and are not ranked in the main text; the Comoros returned no records.

S5 Ranks of all states under each DOAJ measure variant

Table S6 gives the rank of every ranked state under the six variants of Table 3 in the main text, so that each rank movement quoted in Section 3.3 can be checked. Rank 1 is the highest share of a state's articles and reviews in flagged venues.

Table S6 Ranks of 19 states under the six DOAJ measure variants of Table 3

State	1 Current log	2 Archived, ever removed	3 As 2, purge included	4 Archived, status at year (strict)	5 Complete, status at year (strict)	6 Complete, status at year (loose)
Algeria	16	19	19	19	19	19
Bahrain	5	3	3	2	2	2
Egypt	9	15	16	16	17	16
Iraq	3	12	12	10	11	12
Jordan	4	10	11	9	8	8
Kuwait	11	11	4	11	10	11
Lebanon	6	17	17	14	14	15
Libya	18	16	13	17	16	17
Morocco	19	5	7	5	7	7
Oman	12	13	15	15	15	14
Palestine	13	6	8	4	4	4
Qatar	10	14	14	12	12	13
Saudi Arabia	1	2	2	3	3	3
Somalia	14	9	6	7	9	10
Sudan	8	4	5	6	5	5
Syria	15	1	1	1	1	1
Tunisia	17	18	18	18	18	18
United Arab Emirates	2	8	10	8	6	6
Yemen	7	7	9	13	13	9

Columns follow the rows of Table 3. Archived = the log to January 2024; complete = the merged log to 2 September 2026. Strict and loose differ only for journals restored to DOAJ with an unknown readmission date.