Supplementary Information for: **Publication time valid prediction of citation risk outcomes in a bounded clinical specialty literature corpus**

Sunny Chung, MD; Yale School of Medicine Section of Digestive Diseases, New Haven, CT 06511

Charles Kahi, MD; Indiana University School of Medicine, Roudebush VAMC, Indianapolis, IN 46202

Siddharth Singh, MD; Mayo Clinic, Phoenix, Arizona, 85054

Corresponding Author: Sunny Chung (sunny.chung@yale.edu)

Journal: *Scientometrics*

Online Resource 1. Temporal-validity and leakage-control safeguards

Temporal validity governed both feature construction and model evaluation. Predictors were restricted to information that would have been available on or before the source paper's publication date, or publication year when the underlying metadata were available only at annual resolution. This restriction applied to bibliographic metadata, reference-derived features, cited-work maturity measures, semantic features, author-history variables, and all preprocessing steps that transformed those variables.

Several safeguards were implemented before model comparison.

First, source-paper rows were preserved during author-feature merges. Author-layer features were joined to the citation-layer table using a source-paper-preserving merge rather than an inner merge. This prevented papers from being silently dropped when author information was incomplete or unavailable. Missing author-history information was carried forward as missingness where appropriate, rather than being indiscriminately zero-filled. Structural zeros were used only when the absence of a countable prior event had a substantive zero interpretation.

Second, same-year prominence-derived features were removed from the predictive branch. These variables could reflect information accumulated during the same publication year as the source paper and were therefore not treated as publication-time-available signals. Their removal prevented models from using contemporaneous visibility information that would not have been available at the intended prediction time.

Third, MeSH terms and OpenAlex-derived keyword fields were excluded from the main predictive branch. These fields are not reliably available at publication time in the intended use setting and may be assigned, revised, or enriched after publication. They were therefore excluded from the leakage-clean modeling path even if available retrospectively.

Fourth, citation-outcome construction used explicit provenance rules. Three-year citation outcomes were taken from source journal citation records when available. Two-year citation windows were derived from yearly citation trajectories. Outcome construction drew first from original source journal datasets and then, when necessary, from cited-work metadata elements. This allowed recovery of yearly citation trajectories for 2-year endpoints while maintaining explicit source provenance.

Fifth, semantic dimensionality reduction was fold-local. For semantic stages, dimensionality reduction was estimated using only the training years within each temporal fold and then applied unchanged to that fold's validation and held-out test years. Validation-year and test-year papers therefore did not influence the semantic coordinate system used to evaluate them.

Together, these safeguards were intended to preserve the interpretation of the study as publication-time-valid prediction rather than retrospective citation ranking.

Online Resource 2. Journal E and Journal G clinical-corpus construction and audit

Journal E and Journal G were handled through a shared clinical-corpus preprocessing branch before being merged into the common prediction corpus. This additional step was required because the shared source export included both clinical gastroenterology articles and basic science or translational research articles. The goal of this preprocessing branch was to construct a clinical-research source that was comparable to the other clinical journals while preserving Journal E and Journal G as separate analytic units.

Records were first assigned to publication-type lanes, including original articles, reviews, letters, research letters, case reports, front matter, conference material, and excluded nonresearch content. MEDLINE status, human and animal MeSH indicators, clinical-disease MeSH indicators, and publication-type metadata were then used to separate clinically relevant records from basic/translational material, front matter, correspondence, conference abstracts, and case reports.

Within the Journal E/G preprocessing branch, deterministic publication-type, MeSH, and title rules resolved clear cases first. Records remaining unresolved after deterministic triage were classified using a prompt-constrained large language model classifier, not a trained logistic-regression or random-forest model. A total of 3,020 records were submitted to the LLM, of which 1,669 were assigned include, 1,334 exclude, and 17 uncertain before audit. The classifier used OpenAI GPT-5.2 alias (gpt-5.2, temperature set to 0, run on February 2, 2026), and structured JSON output. For each unresolved record, the prompt supplied the journal, publication year, title, PubMed abstract when available, grey-area referral reason, PubMed publication types, MeSH terms, OpenAlex type, and DOI. No citation counts, reference-derived features, model predictors, or downstream citation-outcome labels were supplied to the classifier. The model was constrained to return one of three triage decisions: include, exclude, or uncertain. Included records were further constrained to clinical original research, clinical review, research letter with original data; excluded records were constrained to conference abstract/poster, case report/vignette, correspondence/front matter, editorial/commentary, correction/erratum/retraction, non-article metadata/heading, or other nonresearch. The output schema also required a brief rationale, cues used from the title/abstract, confidence, and flags for human-derived data and ML/AI/digital-health content. LLM-resolved labels were not treated as unreviewed ground truth: all uncertain records and a stratified sample of 206 LLM- and rule-resolved records were manually audited, and audited disagreements were corrected before final corpus construction. The final corpus contained 35 manual overrides. Because the LLM classifier was a proprietary hosted model, exact future regeneration of the model's raw decisions cannot be guaranteed. To support reproducibility, we preserved the prompt template, input fields, structured output schema, model alias, model outputs, audit file, and

manual correction table. The LLM step should be interpreted as a documented corpus-construction aid with manual audit, not as an independently reproducible gold-standard classifier.

The resulting shared clinical source contained 2,205 records, including clinical original articles, clinically anchored reviews, and research letters with original data. The downstream prediction corpus then applied the same article-only, publication-year, and guideline-title filters used for all journals. After those common filters, 1,531 papers entered the final analytic corpus from this branch: 955 in Journal E and 576 in Journal G.

The curation branch was audited by stratified manual review of 206 records spanning rule-classified, classifier-resolved, and uncertain strata. Agreement was 83.0% for the include-versus-exclude decision and 83.5% overall in the completed audit file. Thirty-five audited disagreements were resolved before the final clinical source was used for feature generation or modeling. The audit should be interpreted as a quality-control check for construction of a clinically bounded source, not as a gold-standard classification study of all publication-type lanes.

Online Resource 3. Source-record eligibility exclusions

Two source-record exclusions were applied before article-level eligibility filters. These exclusions were prespecified after review of source-record patterns that carried article-type metadata but did not represent original research articles or followed a different citation logic from the target corpus.

The first exclusion removed guideline-like records by title. Guideline, consensus, practice-update, recommendation, and position-statement articles were excluded because they are often cited through a different mechanism than original research articles. The case-insensitive title rule was:

\b(guideline|guidance|consensus|position statement|recommendation|practice guideline|clinical guideline|practice update|best practice)\b

The second exclusion removed journal-specific abstract or supplement records that carried an article publication-type tag in OpenAlex but did not represent original research articles. For Journal A, records whose DOI matched the following pattern were excluded:

10.14309/00000434-\d{9}-\d+

For Journal D, records whose DOI matched the following 2021 publisher abstract/supplement pattern were excluded:

10.1055/s-0041-172[459]

These DOI-pattern exclusions removed 5,210 source records in total: 4,475 from Journal A and 735 from Journal D. The same title and DOI rules were applied in the citation-layer feature generator and in the workflow used to generate manuscript tables and figures. This ensured that the analytic corpus, manuscript tables, flow diagrams, and modeling inputs used the same source-record eligibility definition.

Online Resource 4: Model-family definitions

Manuscript family	Feature families included	Primary manuscript role
Non-semantic citation/reference/context baseline	Metadata and context; reference composition; cited-knowledge maturity; ancestry and novelty structure	Non-semantic benchmark and reference point for all additive-gain comparisons
Whole-document semantic model	Non-semantic baseline plus whole-document SPECTER2 embedding of title and abstract	Retained primary family for ≤3 citations in 2 years; comparator family against which structure-aware additions are evaluated
Structure-aware contextual model	Nonsemantic baseline plus publication-time reference-context distributional features derived from the source paper's cited literature, using distributions over cited source/journal identifiers compared with year-specific training-fold low- and not-low-citation reference-context centroids	Retained family for >20 citations in 2 years
Structure-aware content model	Nonsemantic baseline plus role-segmented SPECTER2 embeddings of source-paper title/abstract text	Retained family for ≤3 citations in 3 years
Structure-aware content + context model	Nonsemantic baseline plus role-segmented SPECTER2 source-text embeddings and publication-time reference-context distributional features	Retained family for the stringent 0 citations in 3 years endpoint

Contextual feature families did not embed cited-reference text directly. In this supplement, "contextual" refers to publication-time reference-context distributional features derived from cited source/journal identifiers.

Online Resource 5: Pooled versus Journal C-local comparison

Endpoint	Journal C eligible n	Journal C positive n	Journal C prevalence	Pooled PR-AUC	Pooled F1 (95% CI)	Pooled Precision@10%	Best interpretation by F1	Best local PR-AUC	Best local F1 (95% CI)	Best local Precision@10%
≤3 citations in 2 years	540	198	36.7%	0.696	0.645 (0.584-0.699)	0.815	Pooled default	0.696	0.645 (0.584-0.699)	0.815
0 citations in 3 years	545	9	1.7%	0.205	0.326 (0.143-0.491)	0.145	Pooled default	0.205	0.326 (0.143-0.491)	0.145
≤3 citations in 3 years	545	66	12.1%	0.474	0.446 (0.333-0.550)	0.491	Pooled default	0.474	0.446 (0.333-0.550)	0.491
>20 citations in 2 years	540	43	8.0%	0.240	0.247 (0.130-0.375)	0.259	Journal-local threshold	0.240	0.286 (0.182-0.393)	0.259

Journal C-local operating-point audit comparing pooled-default performance with local thresholding/calibration choices for the journal-local held-out rows. The table reports the local denominator, positive-label count, prevalence, pooled and local performance metrics, and nonparametric bootstrap intervals for the F1 values used to select the best local interpretation. Note. Best interpretation is selected by held-out F1 within the Journal C analysis; ranking metrics remain unchanged when only the threshold changes. Precision@10% denotes precision among the top-ranked decile of held-out predictions. F1 intervals are nonparametric bootstrap 95% intervals over the journal-local held-out rows; PR-AUC and Precision@10% are reported as point estimates. Eligible n differs between 2-year and 3-year endpoints because 2-year endpoints required recovered yearly citation trajectories, whereas 3-year endpoints used precomputed three-year citation outcomes available in the source citation records.

Online Resource 6. Metrics, bootstrap uncertainty, and endpoint-specific reporting selection

Probability thresholds for binary classification were chosen within each temporal fold on that fold's validation year by maximizing validation F1. The selected thresholds were then applied without re-selection to the held-out test year. This separated threshold selection from test-set evaluation.

Reported pooled held-out metrics were computed over the concatenation of the 2021 and 2022 test rows from the two temporal folds. In the reference-observed cohort, this corresponded to 2,851 held-out rows for 3-year endpoints and 2,626 held-out rows for 2-year endpoints. The 2-year evaluation basis was smaller because 2-year endpoints required recovered yearly citation trajectories.

Model performance was summarized using PR-AUC, F1, and precision@10%. PR-AUC was used because several endpoints were imbalanced and because ranking performance among positive cases was central to interpretation. F1 summarized thresholded classification at the validation-selected operating point. Precision@10% was defined as precision among the top-ranked 10% of held-out predictions for the endpoint under evaluation and was included to summarize ranking behavior at a consistent review depth. We used a top-decile threshold rather than a fixed k because held-out denominators differed across 2-year and 3-year endpoints.

Uncertainty intervals were computed from archived held-out predictions rather than from full retraining resamples. This pragmatic choice holds trained models fixed and resamples only held-out evaluation rows. The intervals therefore reflect evaluation-row sampling variance but understate variance attributable to model fitting, feature generation, and hyperparameter selection.

For each endpoint, we computed two complementary fold-stratified bootstrap analyses using 2,000 resamples. The first analysis estimated per-model pooled fold-mean intervals for PR-AUC, F1, and precision@10%. Within each temporal fold, positive and negative examples were resampled with replacement, preserving fold and label structure. Metrics were then recomputed on the resampled held-out rows and summarized across bootstrap replicates.

The second analysis estimated paired retained-minus-comparator metric differences. For each bootstrap replicate, the same held-out rows were resampled for both the retained model family and the whole-document semantic comparator. This paired design isolated the model-pair difference from fold and label noise. Paired intervals are reported in Online Resource 11 and were used to support cautious interpretation of endpoint-specific gains, especially when per-model intervals overlapped.

For compact endpoint-level reporting, we designated an endpoint-specific retained family using a prespecified multi-metric reporting heuristic. A structure-aware feature family was retained over the whole-document semantic comparator only when PR-AUC improved

by at least 0.005, thresholded F1 did not decrease by more than 0.002, and precision@10% did not decrease by more than 0.01. These thresholds were intended as minimal practical-change filters to prevent highlighting a structure-aware variant on the basis of a single unstable metric. The heuristic was not treated as a formal hypothesis test. All comparator and retained-family metrics are reported, and paired bootstrap intervals provide the inferential comparison between retained and comparator families.

Online Resource 7. Journal-local validity diagnostic and calibration audit

Primary model performance was summarized in pooled cross-journal analyses. We additionally conducted a prespecified journal-local diagnostic for one within-corpus journal, Journal C. The purpose of this analysis was to evaluate whether pooled multi-journal performance translated into local validity for a specific journal context. It was not intended as a deployment claim, and no journal-specific model was trained for the main analysis.

The endpoint-specific pooled models were applied unchanged to held-out Journal C rows. The audit then compared the pooled default interpretation with local reinterpretations of the same prediction scores. The pooled default used the model probabilities and thresholds selected in the pooled modeling workflow. Journal-local thresholding kept the same prediction scores but selected a Journal C-specific F1 threshold using a cross-fit procedure: one held-out fold was used to estimate the local threshold, and the other held-out fold was used for evaluation, with the roles then reversed.

Journal-local recalibration used the same cross-fit structure. A logistic Platt recalibration model was fit on the opposite held-out fold and evaluated on the current fold. The recalibrated probabilities and fold-specific threshold were then used to compute thresholded classification metrics. Because both local thresholding and local recalibration used only held-out Journal C rows after the pooled model had already been trained, this audit assessed local reinterpretation of pooled predictions rather than construction of a new local model.

Threshold-only reinterpretation does not change the ordering of papers. Therefore, PR-AUC and precision@10% remain unchanged when only the threshold moves. F1 can change because the positive/negative decision boundary changes. Recalibration can change probability calibration and Brier score, but it does not necessarily improve ranking or thresholded classification. For this reason, the journal-local analysis reported PR-AUC, F1, precision@10%, Brier score, calibration slope, and the selected operating point as distinct diagnostic quantities.

The Journal C analysis should be read as a local-validity and operating-point diagnostic. It does not show that the pooled models are ready for journal-specific deployment. Any operational use would require prespecified local validation, calibration assessment, governance, and human oversight beyond the diagnostic analysis reported here.

Online Resource 8: Masked Journal Characterization

Journal	Corpus n	Ref-observed %	0 citations, 3y %	≤3 citations, 3y %	≤3 citations, 2y %	>20 citations, 2y %	Median 3y citations	Median refs	PubMed abstract %	Source branch
Journal A	1442	77.1%	9.4%	30.6%	46.4%	5.6%	9	29.5	60.5%	Direct OpenAlex source
Journal B	991	99.5%	0.8%	2.5%	14.9%	8.1%	18	36	96.8%	Direct OpenAlex source
Journal C	2215	79.6%	14.9%	31.8%	46.6%	6.5%	10	25	63.9%	Direct OpenAlex source
Journal D	2366	92.9%	26.2%	65.9%	76.3%	1.6%	2	5	22.0%	Direct OpenAlex source
Journal E	955	94.1%	1.7%	7.9%	14.8%	37.2%	35	36	72.6%	Curated Journal E/G branch
Journal F	879	99.5%	2.4%	11.4%	38.2%	5.2%	14	37	90.7%	Direct OpenAlex source
Journal G	576	100.0%	0.2%	0.9%	7.5%	39.8%	42	43	96.2%	Curated Journal E/G branch

Endpoint prevalences are shown by masked journal to make corpus heterogeneity assessable without disclosing journal identities. Three-year endpoint prevalences use papers with 3-year citation outcomes; two-year endpoint prevalences use papers with recovered yearly citation trajectories. PubMed abstract coverage indicates the share of source papers with PubMed-enriched abstracts.

Online Resource 9: Curated-branch prevalence sensitivity

Analysis set	Corpus n	Ref-observed n	3y available n	0 citations, 3y n (%)	≤3 citations, 3y n (%)	2y available n	≤3 citations, 2y n (%)	>20 citations, 2y n (%)
All journals	9424	8409	9424	1132 (12.0%)	2911 (30.9%)	8711	3711 (42.6%)	948 (10.9%)
Excluding curated Journal E/G branch	7893	6934	7893	1115 (14.1%)	2831 (35.9%)	7181	3527 (49.1%)	364 (5.1%)
Curated Journal E/G branch only	1531	1475	1531	17 (1.1%)	80 (5.2%)	1530	184 (12.0%)	584 (38.2%)

Sensitivity analysis comparing endpoint prevalence in the full seven-journal corpus, the corpus excluding the curated Journal E/G branch, and the curated Journal E/G branch alone. Three-year endpoints use papers with 3-year citation outcomes; two-year endpoints use papers with recovered yearly citation trajectories.

Pre-pandemic publication sensitivity: For the primary ≤3 citations within 2 years endpoint in the reference-observed cohort, prevalence was 37.6% among 2017–2018 publications whose 2-year citation windows ended before 2020, 37.2% among 2017–2019 pre-pandemic publication years, and 38.6% among 2020–2022 publications. This descriptive sensitivity did not retrain a separate pre-pandemic temporal model.

Online Resource 10: Primary endpoint calibration and operating-point diagnostics

Role	Family	n (pos)	Brier	Cal intercept	Cal slope	Mean p (+)	Mean p (-)	Thresholds (fold 1; fold 2)
Retained primary family	Whole-document semantic	2,626 (1,049)	0.145	0.149	0.890	0.628	0.200	0.270; 0.390

Post hoc primary-endpoint operating-point diagnostics for the retained whole-document semantic model for ≤3 citations within 2 years. Brier score is the mean of fold-specific held-out Brier scores. Calibration intercept and slope were fit post hoc on pooled held-out logits after clipping probabilities to [1e-6, 1 - 1e-6]. Thresholds are validation-F1-selected probability cutoffs. These diagnostics should be interpreted as post hoc operating-point analyses rather than evidence of independent model superiority.

Online Resource 11: Bootstrap uncertainty and paired retained-versus-comparator differences

Endpoint	Retained family	Comparator family	Metric	Comparator value	Retained value (95% CI)	Retained minus comparator (95% CI)	Pr(delta > 0)
≤3 citations in 2 years	Whole-document semantic	Whole-document semantic	PR-AUC	0.828	0.828 (0.809-0.847)	0.000	n/a
≤3 citations in 2 years	Whole-document semantic	Whole-document semantic	F1	0.735	0.735 (0.716-0.754)	0.000	n/a
≤3 citations in 2 years	Whole-document semantic	Whole-document semantic	Precision@10%	0.962	0.962 (0.935-0.981)	0.000	n/a
0 citations in 3 years	Structure-aware content + context	Whole-document semantic	PR-AUC	0.472	0.472 (0.429-0.530)	0.000 (-0.029 to +0.032)	0.552
0 citations in 3 years	Structure-aware content + context	Whole-document semantic	F1	0.325	0.453 (0.413-0.494)	+0.128 (+0.087 to +0.170)	1.000
0 citations in 3 years	Structure-aware content + context	Whole-document semantic	Precision@10%	0.458	0.458 (0.417-0.510)	0.000 (-0.035 to +0.053)	0.623
≤1 citation in 3 years	Structure-aware content, robustness reference	Whole-document semantic	PR-AUC	0.661	0.664 (0.626-0.709)	+0.003 (-0.028 to +0.032)	0.577
≤1 citation in 3 years	Structure-aware content, robustness reference	Whole-document semantic	F1	0.694	0.714 (0.689-0.739)	+0.020 (+0.001 to +0.038)	0.980
≤1 citation in 3 years	Structure-aware content, robustness reference	Whole-document semantic	Precision@10%	0.710	0.704 (0.655-0.756)	-0.007 (-0.055 to +0.036)	0.358
≤3 citations in 3 years	Structure-aware content	Whole-document semantic	PR-AUC	0.860	0.866 (0.845-0.886)	+0.006 (-0.001 to +0.013)	0.967
≤3 citations in 3 years	Structure-aware content	Whole-document semantic	F1	0.808	0.811 (0.791-0.830)	+0.003 (-0.006 to +0.012)	0.727
≤3 citations in 3 years	Structure-aware content	Whole-document semantic	Precision@10%	0.941	0.948 (0.923-0.972)	+0.007 (-0.011 to +0.038)	0.860
>20 citations in 2 years	Structure-aware contextual	Whole-document semantic	PR-AUC	0.464	0.509 (0.455-0.567)	+0.045 (-0.007 to +0.089)	0.957
>20 citations in 2 years	Structure-aware contextual	Whole-document semantic	F1	0.455	0.472 (0.428-0.515)	+0.017 (-0.023 to +0.058)	0.796
>20 citations in 2 years	Structure-aware contextual	Whole-document semantic	Precision@10%	0.491	0.491 (0.438-0.537)	0.000 (-0.049 to +0.042)	0.436

Fold-stratified bootstrap 95% intervals are shown for retained endpoint-model performance and paired retained-minus-comparator metric differences. Retained-model intervals resample held-out evaluation rows with trained models fixed. Paired delta intervals resample the same held-out rows for the retained family and whole-document semantic comparator; positive deltas favor the retained family. For the primary ≤3 citations within 2 years endpoint, retained and comparator families are identical, so paired deltas are zero by construction. The ≤1 citation within 3 years endpoint is included as a robustness endpoint only: although the structure-aware content model had slightly higher PR-AUC than the whole-document semantic comparator, the paired PR-AUC difference was below the prespecified +0.005 reporting threshold and its interval spanned zero.