Online Resource 1

Supplementary evidentiary chronology, hypothesis map, protocol-accountability details, and frozen lexical operationalization

Article title: Stopping Is a Conditional Decision: Corpus Qualification, Bayesian Sequential Stopping, and the Limits of Early Termination in Scholarly Literature Screening

Journal: Scientometrics

Authors: Mohammad Pasha; Mohammed Ali Shaik
Affiliations: SR University, Ananthasagar, Warangal 506371, Telangana, India
Corresponding author: Mohammad Pasha; mohd.pasha@outlook.com; ORCID 0000-0002-0596-1682

S1. Evidentiary chronology

The project used staged preregistration and prospective freezes to distinguish confirmatory, prospectively frozen, exploratory, and post-outcome descriptive analyses. Local cryptographic freezes are reported as such and are not represented as independent OSF registrations.

Stage	Evidentiary status	Primary purpose	Data context	Key decision
v1.4 ARS-BSC core	Prospectively registered/frozen	Corpus qualification and gated BSC	Controlled simulations + YOLOv8 edge-device corpus	BSC allowed only after ARS qualification
v1.5 boundary extension	Prospectively frozen	Search for a valid/useful qualification boundary	384,000 Monte Carlo runs	No post-hoc lowering of validity/usefulness criteria
Semantic development	Exploratory/ development	Test conceptual redundancy as alternative stopping information	Wolters + four labelled reviews	Final clusters used only for evaluation
v1.6 readiness test	Prospectively cryptographically frozen	External test of early semantic readiness and gating	Four untouched labelled reviews	Setting B and outcome thresholds fixed
v1.7 horizon test	Prospectively cryptographically frozen	Compare readiness at m=5 vs m=10 relevant papers	Four new untouched labelled reviews	Primary estimand Delta AUC=AUC10-AUC5
Retrieval ordering	Post-confirmatory exploratory; not preregistered/frozen	Test whether better ordering rescues stopping	Same v1.7 panel	Random vs BM25 vs LSA vs BM25-to-LSA rerank; descriptive stress test only
v1.8 benchmark-qualified BSC validation	Prospectively locally cryptographically frozen before outcome inspection	Isolate BSC performance after removing benchmark-relative corpus loss	10 previously unconsumed labelled review corpora; 500 random permutations/review	Primary Beta(1,1), ρ=0.10; validity/usefulness criteria unchanged from v1.5

Table S1. Analysis chronology and evidentiary status.

S2. Glossary and hypothesis map

Term / hypothesis	Meaning / registered role	Evidence stage
ARS; A/C/K/R	Adaptive Research Scoping; Alignment, Coverage, Contamination (lower is better), and Context Representation. These jointly define corpus admissibility.	Core architecture
D; R*	D is the frozen candidate corpus. R* is a constructed reference set; empirical recall against it is reference-set-relative, not absolute field recall.	ARS / evaluation
BSC; R_n	Bayesian Sequential Stopping; R_n is the number of relevant records remaining inside frozen D after n screens.	Closed-corpus stopping
rho = c_s/c_m	Relative screening-cost to missed-relevant-record loss used by the Bellman decision rule; an operating preference, not a universal constant.	BSC decision rule
Cov80 / Cov95; SBC	Empirical predictive-interval coverage; simulation-based calibration was preregistered but not completed, so no SBC validation claim is made.	Calibration
H0-H5	Core tests: input dominance, diagnostic identifiability, matched-budget retrieval, contamination, prior/field-density effects, and small-N calibration.	v1.4-v1.5
H6	Semantic-readiness family: whether early redundancy/similarity predicts safe savings and whether a frozen readiness gate improves the joint objective.	v1.6
H7	Information-horizon family: whether waiting for more relevant records improves readiness discrimination without erasing work saving.	v1.7
H8a-H8d	Benchmark-qualified BSC validation: calibration, premature stopping/recall safety, and usefulness under C_benchmark = 1.	v1.8
Ranking vs stopping	Ranking benefit = earlier relevant evidence / better ordering. Stopping benefit = a defensible decision to terminate. The former does not imply the latter.	Cross-cutting interpretation

Table S2. Compact glossary and hypothesis map.

S3. Registered comparator accountability

The v1.4 freeze registered consecutive-irrelevance rules (k=10,20,50), SAFE where rank scores exist, Callaghan-type rules where ranker output exists, Chao-type estimators when at least 10 relevant records have been observed, an evaluation-only 80%-recall oracle, and broader knee/fixed-budget comparator classes. It also froze explicit NA rules: SAFE/Callaghan are NA on unranked strata and Chao is NA when fewer than 10 relevant records have been observed. The primary YOLOv8 path terminated UNQUALIFIED, so BSC and its screening-stop comparators were protocol-defined NA on that path rather than negative outcomes.

A complete registered-comparator benchmark was not archived with the preregistered core analyses. After the primary outcomes were known, we therefore completed only the unambiguous consecutive-irrelevance

k={10,20,50} rules and the evaluation-only 80%-recall oracle on the frozen v1.8 replay orders, explicitly as descriptive/non-confirmatory analyses. We do not claim empirical superiority of BSC over SAFE, Callaghan, Chao, knee, or fixed-budget procedures. Retrospectively implementing those underspecified or input-dependent variants would introduce new analytic degrees of freedom, so they remain reported as NA/unexecuted rather than reconstructed post hoc.

Registered comparator	Frozen applicability rule	Status in core / current manuscript	Interpretive consequence
Consecutive irrelevance k={10,20,50}	Unranked sequence allowed	Post-outcome descriptive completion on v1.8: k=10/20/50; not confirmatory	Reported only as accountability/descriptive context; no confirmatory superiority claim
SAFE / SAFE-like	Rank scores required	NA on unranked core strata; later BM25/LSA analysis was exploratory, not a retroactive SAFE test	No SAFE superiority claim
Callaghan-type	Ranker output required	NA on unranked core strata; no frozen later implementation archived	No Callaghan superiority claim
Chao-type	At least 10 relevant records observed	NA whenever requirement not met; no frozen later implementation archived	No Chao superiority claim
Oracle 80% recall	Evaluation only; truth required	Post-outcome descriptive completion on v1.8; evaluation only	Quantifies attainable work saving with unavailable future truth; not deployable
Knee / fixed budget	Registered comparator class	Exact post-freeze algorithm/budget specification was insufficient for unique retrospective reconstruction	Disclosed protocol incompleteness; not reconstructed after outcomes

Table S3. Registered comparator accountability.

Simulation-based calibration (SBC) was also preregistered as a primary BSC calibration outcome. No frozen SBC implementation/output was produced in the completed analysis chain, and SBC was therefore not completed as preregistered. This is reported as an explicit protocol deviation rather than inferred from interval coverage. The manuscript makes no SBC claim; empirical 80%/95% predictive-interval coverage and operating characteristics are reported separately. A post-outcome SBC implementation would not repair the preregistered omission and is not presented as confirmatory evidence.

S4. Freeze chronology and version-label interpretation

The v1.4 core protocol was prospectively registered. The v1.5 boundary analysis was prospectively frozen. The v1.6 semantic-readiness and v1.7 information-horizon analyses were prospectively locally cryptographically frozen before their new external analyses. Retrieval ordering was explicitly post-confirmatory exploratory. The v1.8 benchmark-qualified validation was prospectively locally cryptographically frozen before BSC outcome inspection on the selected panel. Version labels document evidentiary chronology; the main manuscript is organized by scientific question rather than development sequence.

S5. v1.8 benchmark-qualified panel

The frozen panel contained ten labelled review corpora: Appenzeller-Herzog_2019, Theobald_2021, Abgaz_2023, Ali_2024, Anmarkrud_2021, Aouad_2024, Bech_2019, Wijnen_2024, Williamson_2023, and Zakeri-Nasrabadi_2023a. Eligibility required an exact CSV, binary inclusion labels, 5-30 known

included records, N<=1000, and no selection based on BSC outcome. The supplied review corpora are drawn from the SYNERGY screening-data collection; benchmark qualification means C_benchmark=1 within the supplied labels, not complete field coverage.

Dataset citation: De Bruin, J., Ma, Y., Ferdinands, G., Teijema, J., & van de Schoot, R. (2023). SYNERGY - Open machine learning dataset on study selection in systematic reviews (Version 1) [Data set]. DataverseNL. https://doi.org/10.34894/HE6NAQ

S6. Interpretive guardrails for the decision model

The Bellman loss ratio ρ=c_s/c_m is an operating preference, not a recall target. The frozen BSC policy minimizes expected screening-plus-omission loss under the stated model and does not directly optimize the manuscript's external ≥0.80 recall criterion. The prespecified ρ sensitivity analysis is therefore interpreted as an empirical safety-efficiency frontier. No post-outcome recall-constrained stopping rule was introduced, because doing so would define a new optimization problem after inspection of the frozen outcomes.

The YOLOv8 R* is a fixed, constructed 30-item diagnostic reference standard rather than a probability sample of the complete field. Relative coverage against R* is therefore reported descriptively. A conventional binomial confidence interval around that fixed proportion would not quantify uncertainty about relevant studies omitted from R* itself.

The primary Beta-Binomial BSC assumes conditional exchangeability and provides a conjugate predictive model for the residual relevant-record count inside a frozen corpus. A hypergeometric or other finite-population formulation is a natural prospectively specified sensitivity analysis for future work; it was not retrofitted after the present outcomes were observed.

S7. Frozen lexical operationalization for the YOLOv8/edge-device demonstration

For the frozen YOLOv8/edge-device demonstration, title and abstract text were lower-cased and matched against fixed lexical dictionaries. QPFS(D)=|D|^{-1}∑_i I[YOLO_i AND OD_i AND (EDGE_i OR LIGHT_i)], whereas EdgeCoverage(D)=|D|^{-1}∑_i I[EDGE_i]. The frozen dictionaries were:

- **YOLO:** {yolov8, yolo v8, yolo}
- **Object detection:** {object detection, object detector, detection}
- **EDGE:** {edge, embedded, resource constrained, resource-constrained, jetson, raspberry pi, tensorrt, onnx, tflite, microcontroller, npu, accelerator, on-device, on device, edge device, edge computing, low-power, low power}
- **LIGHT:** {lightweight, tiny, nano, efficient, efficiency, compressed, compression, pruning, quantization, low-compute, low compute, small model}

These dictionaries are demonstration-specific operationalizations of the ARS dimensions. They are reported for reproducibility and are not proposed as universal measures of corpus quality.