
Proceedings Paper:
Fletcher, A. and Stevenson, M. (2026) Confidence-based stopping methods for systematic reviews. In: Moffat, A., Scholer, F., Bast, H., Najork, M. and Zhang, M., (eds.) SIGIR '26: Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '26: The 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, 20-24 Jul 2026, Melbourne, VIC, Australia. Association for Computing Machinery, pp. 3721-3726. ISBN: 9798400725999.

https://doi.org/10.1145/3805712.3809924

Confidence-Based Stopping Methods for Systematic Reviews

Aaron H.A. Fletcher
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
ahafletcher1@sheffield.ac.uk

Mark Stevenson
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
mark.stevenson@sheffield.ac.uk

Abstract

Technology Assisted Review stopping methods aim to ensure that no more documents are screened than necessary. Most existing approaches focus on achieving a target recall, which does not consider whether an information need has been met. This paper introduces two heuristic stopping methods that instead monitor whether screened documents contain enough information to make a decision. Evaluation on a standard dataset of Diagnostic Test Accuracy Systematic Reviews demonstrates that the proposed approaches substantially reduce the number of documents that need to be examined while, in the majority of cases, maintaining conclusions that are consistent with all evidence available.

CCS Concepts

• **Information systems** → **Retrieval effectiveness**; **Retrieval efficiency**.

Keywords

technology-assisted review, stopping rules, systematic reviews, diagnostic test accuracy

ACM Reference Format:
Aaron H.A. Fletcher and Mark Stevenson. 2026. Confidence-Based Stopping Methods for Systematic Reviews. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3805712.3809924

1 Introduction

Technology Assisted Review (TAR) develops methods to identify information when it would be impractical to examine an entire collection [30]. Key applications include the development of systematic reviews [16] and eDiscovery [15, 26, 31, 35]. Within TAR, stopping rules help reviewers determine when to stop examining documents, thereby reducing the workload by reviewing only the necessary documents. A wide range of stopping methods have been proposed, most of which base the decision to stop on having identified a desired portion of all the relevant documents within the collection, known as the *target recall* [25, 39]. However, basing the stopping decision on target recall does not consider whether a user's information need has been satisfied. Under certain circumstances, a user's information need may be satisfied with information contained within a single document, even when many other relevant documents in the collection have not been identified. In other cases, the information need cannot be satisfied even after screening the entire collection.

An alternative approach is to consider whether the user's information need has been satisfied, rather than whether target recall has been achieved. This approach is formalised in this work within the context of developing Systematic Reviews, specifically those that determine the accuracy of diagnostic tests, such as blood tests for viral infection. The accuracy of such tests is reported in Diagnostic Test Accuracy (DTA) studies using clinical metrics such as False Positive Rate (FPR) and Sensitivity.[1] However, the values reported in individual studies vary due to factors such as different sample sizes and populations tested. DTA Systematic Reviews aim to summarise all available data to produce the best available estimate of test accuracy. However, systematic reviews are expensive and time-consuming to carry out, making them impractical in many scenarios. Instead, an individual may want to understand whether the test is accurate enough for some purpose or not. This paper introduces two stopping algorithms for this scenario, each designed to determine whether the examined evidence is sufficient to answer the information need. The approaches monitor the evolving point estimates and uncertainty of the clinical metrics and stop when the accumulated evidence appears sufficient to support the thresholded decision.

These approaches are compared against existing stopping rules (most of which base their decisions on target recall) and found to improve the efficiency of the review process. They achieve comparable results to target recall-based methods while examining substantially fewer documents.

This paper contributes by: (i) introducing confidence-based stopping rules that focus on determining when an information need has been met, rather than when a target recall has been achieved; (ii) developing two such rules for a systematic review setting; (iii) evaluating those rules on a real-world dataset of DTA reviews.

2 Background

TAR stopping rules aim to cut screening effort while meeting an acceptability constraint [10, 23, 25, 39]. Most methods monitor progress toward a *target recall*, e.g. 70% of all relevant documents in the collection, and stop once enough relevant documents have been observed that the target is likely to have been met or even exceeded. This recall-focused approach has historical roots in legal information retrieval [5].

[1] The sensitivity metric is more commonly known as recall within Information Retrieval. We use the term "sensitivity" to refer to estimates of diagnostic test accuracy reported within scientific papers while reserving the term "recall" to its standard usage, i.e. proportion of relevant documents identified.

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809924

Approaches include observing the yield of relevant items over time [10, 39], estimating remaining relevant items via sampling or classification [6, 7, 11, 25, 27, 38], analysing rank scores for inflection points [13, 17] and reinforcement learning [3, 4]. Evaluation typically examines whether the method reliably attains the target recall [10] or how far achieved recall deviates from the target [25, 39]. Framing stopping around recall is attractive because it rests on the familiar notion of relevance, is straightforward to implement with standard test collections (e.g., CLEF e-Health [19–21], TREC Total Recall [15], TREC Legal [12], RCV1 [24]), and aligns with eDiscovery where parties are required to achieve a target recall [41].

However, recall assesses the proportion of relevant documents discovered and not whether the user's information need has been met. It also treats all relevant items as equally valuable [36, 37] which is not the case in systematic reviews where studies contribute unequally to conclusions; large, high-quality studies often dominate the evidence and those reporting experiments carried out over small samples add little marginal value. It has been demonstrated that conclusions may be clear without the need to examine all relevant documents [29]. Consequently, a recall threshold can force unnecessary screening even when additional evidence does not change the answer.

The present work differs from previous TAR stopping methods in that it does not target a recall threshold, but instead monitors whether the accumulated evidence is sufficient for a downstream decision.

Our approach has some parallels with work on the analysis of stopping behaviour in interactive search which demonstrated that individuals stop searching when information feels "good enough," balancing benefits and costs under cognitive and environmental constraints [2, 8, 14, 18, 32, 33, 40, 42] and economic models that treat stopping interactive search as an economic problem by maximising expected utility net of search costs [1]. However, our focus is on determining when to stop examining ranked lists of documents rather than interactive search.

3 Stopping Methods

Assume that a collection of documents has been ranked and that its documents are examined in that order. Let i be the ith relevant study encountered in the ranking. When retrieving documents, a point estimate, $\hat{p}_i$, for a metric of interest (e.g., sensitivity or FPR) can be derived from the cumulative information gathered so far. We further assume that the user wishes to know whether the value of this metric is greater or less than some threshold, θ. The goal is to find the smallest index, I^*, such that sufficient information has been acquired for the information need to have been met.

The methods proposed here instantiate a confidence-based perspective through two heuristic rules which monitor the evolving value of $\hat{p}_i$ and its uncertainty.

Harmonic Shrinkage. The first approach sets a boundary around the value of θ which is gradually narrowed as more estimates of $\hat{p}_i$ become available. The approach stops when all values of $\hat{p}_i$ lie outside this boundary, either all above or below it. The approach is formalised as:

$$I^* = \arg\min_i \left(\left(\forall j \leq i,\ \hat{p}_j \geq \theta + \frac{s(1-\theta)}{i} \right) \vee \left(\forall j \leq i,\ \hat{p}_j \leq \theta - \frac{s\theta}{i} \right) \right) \tag{1}$$

where s is a hyperparameter that controls the convergence rate of the decision boundary.

This approach will never stop if the estimates straddle the decision threshold for any pair of estimates, i.e. $\hat{p_k} > \theta$ and $\hat{p_l} < \theta$ for some $k, l \leq i$, potentially limiting its applicability when the true value is close to the decision threshold.

Confidence Boundary. Another approach is to construct confidence intervals around each $\hat{p}_i$ using estimated variance.

This approach addresses a limitation of the HARMONIC SHRINKAGE method, namely its sole reliance on point estimates and its failure to consider the variability of $\hat{p}_i$. Let $MOE_{0.95}$ be the margin of error corresponding to the 95% confidence interval calculated from the cumulative data observed so far. The CONFIDENCE BOUNDARY approach is defined as follows:

$$I^* = \arg\min_i \left((\hat{p}_i + MOE_{0.95} < \theta) \vee (\hat{p}_i - MOE_{0.95} > \theta) \right) \tag{2}$$

In other words, this approach stops when the estimate is at least one margin of error away from θ.

4 Experiments

Experiments evaluate the performance of the proposed stopping methods along two dimensions: their ability to reach the correct conclusion given the evidence while also minimising the number of documents that need to be examined.

4.1 Dataset

Evaluation is based on the DTA systematic reviews from the Cochrane Library available from the CLEF Technology-Assisted Review in Empirical Medicine exercises (CLEF 2017/2018/2019) [19–21] which have been widely used in previous work on TAR stopping. These data sets consist of scientific articles from the PubMed database, identified uniquely by a PubMed Identifier (PMID), labelled for relevance depending on whether or not they contain information about a diagnostic test's accuracy.

Each review addresses multiple *outcomes*, each corresponding to an evaluation of the diagnostic test within a specific application (e.g. to a particular age group), with relevant PMIDs containing information about one or more of the review's outcomes. In order for the data set to be useful to evaluate the approaches proposed in this paper we need information about the test accuracy for each outcome reported in each relevant PMID. Fortunately much of this is available in the LIMSI-Cochrane dataset [28] which contains diagnostic counts of test accuracy (i.e. True Positives, True Negatives, False Positives and False Negatives) for relevant documents in CLEF 2017 and 2018.

The LIMSI-Cochrane dataset was constructed using a combination of optical character recognition technology, manual verification and post-editing with a reported low extraction error rate (0.06–0.3%). Since LIMSI-Cochrane predates the CLEF 2019 reviews, the required information for studies within these reviews were extracted directly from source documents available on the Cochrane Library website using AWS Textract OCR technology and linked

to the corresponding PMID. In some cases, it was not possible to extract diagnostic counts, and such PMIDs were treated as being irrelevant (i.e. do not contain useful information for determining the outcome).

The evaluation dataset was constructed by considering each outcome individually. The set of PMIDs from which diagnostic counts for each outcome were available was identified and the outcome retained only if there were at least five such PMIDs. This produced a data set consisting of systematic review outcomes for which PMIDs identified as relevant also included diagnostic counts for that outcome. Because each systematic review may address multiple outcomes, the number of relevant PMIDs for an individual outcome is typically smaller than the number of relevant PMIDs for the full review. This procedure yielded a dataset containing 135 qualifying outcomes: 90 from CLEF 2017 reviews, 29 from CLEF 2018, and 16 from CLEF 2019.

For our evaluation, we also require a full-evidence reference value for the sensitivity and FPR for each outcome. A systematic review contains the best estimate of these values using all evidence available at the time of its creation so we use the value produced by considering all extracted relevant studies for an outcome as the full-evidence reference value. This may differ from the value reported in the original review if diagnostic counts could not be extracted from some PMIDs.

4.2 Implementation

The complete set of documents for each outcome is ranked by AutoTAR [9] using a reference implementation [25] with an initial sample of 100 documents. For each outcome, documents are then examined in ranked order and when each relevant document is encountered the sensitivity and FPR estimates are updated using the Reitsma bivariate meta-analytic model estimate [34] to produce the sequence of estimates $\{\hat{p}_i\}$ associated with each relevant document. Reitsma also computes the confidence interval ($MOE_{0.95}$) used by the Confidence Boundary approach.

Decision thresholds (θ) were set to 0.75 for sensitivity and 0.1 for FPR. These values were selected because of their clinical plausibility, e.g. a test with a sensitivity of 0.75 is potentially useful if it is inexpensive and non-invasive compared to alternative diagnostic procedures, despite missing a large proportion of the population with a disease.

For Harmonic Shrinkage, the convergence rate, s, was set to 0.25 to balance early stopping and decision stability.

4.3 Baselines

The proposed stopping algorithms were compared against several baselines. Several of these approaches rely on hyperparameters which can be used to create conservative and aggressive versions which bias it towards later and earlier stopping, respectively. The conservative variants are Knee (-), Target (5), and SCAL (0.9), while Knee (2), Target (3), and SCAL (1.05) represent more aggressive variants.

70% Recall Baseline: Stop after at least 70% of relevant studies are retrieved. This represents a common heuristic used in some TAR applications, aiming to balance search effort with the goal of finding a substantial portion of relevant material. This baseline relies on perfect knowledge of the relevant documents in the collection, which is typically unknown during retrieval.

Knee: Monitor the recall-versus-cost (gain curve) during the review process and stop when the rate of finding new utility-labelled studies significantly decreases relative to the initial rate, identified algorithmically as the 'knee' point of diminishing returns on the curve [10]. Two variants are used with the ρ parameter set to 2 and "dynamic": "Knee (2)" and "Knee (-)".

Target: Randomly samples a set number of relevant studies, creating a target set. Then, a standard TAR process runs until all documents in this initial target set have been reviewed [10]. The target set value, $|T|$, is set to 3 and 5, indicated by "Target (3)" and "Target (5)".

SCAL (Scalability of Continuous Active Learning): Iteratively train rankers on a large random sample from the collection, then estimate the total number of relevant studies (R) via stratified sub-sampling (calibrated for the full collection) to determine a score threshold. This threshold, which achieves a target recall based on R, is applied to the final ranked list of the entire collection to select documents [11]. SCAL's target recall is set to 0.7 and the η values to 1.05 and 0.9: "SCAL (1.05)" and "SCAL (0.9)".

With the exception of 70% Recall, all baselines were computed using the reference implementations described by Li and Kanoulas [25].

4.4 Evaluation metrics

Decision Agreement (DA). Following Kusa et al. [22], a stopping algorithm's effectiveness was evaluated by calculating the proportion of outcomes for which the decision when the algorithm stops is the same as the decision would have been given all potential evidence.

The decision at index i is a binary value given by

$$\text{Decision}(i) = \begin{cases} 1, & \text{if } (\hat{p}_i > \theta \text{ for Sens.}) \ \vee \ (\hat{p}_i < \theta \text{ for FPR}) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Then the decision agreement for a set of outcomes, O, is calculated as the proportion of times the decisions at the stopping point ($Decision(I^*)$) and when all information is considered ($Decision(n)$) agree over all outcomes, i.e.,

$$\text{Decision Agreement} = \frac{1}{|O|} \sum_{o \in O} \mathbb{I}\big(\text{Decision}_o(I_o^*) = \text{Decision}_o(n_o)\big) \,. \tag{4}$$

where $\mathbb{I}$ is the indicator function and n_o denotes the index after all relevant studies for outcome o have been retrieved.

Savings (Sav.) The percentage of documents in the collection not reviewed before the stopping point, averaged over all outcomes.

Stops The percentage of outcomes where the stopping algorithm is activated, i.e. does not examine all documents in the ranking.

Appropriate Stops (AS) Stopping is not always appropriate, such as when the collection lacks sufficient information to address the information need. A stopping decision is deemed appropriate if either: (1) the algorithm stopped when sufficient information was available for a confident decision, or (2) the algorithm continued when insufficient information was available. Sufficient information is defined as a state where the decision threshold (θ) lies outside

Table 1: Performance of confidence-based and baseline stopping algorithms across 135 DTA review outcomes, evaluated on False Positive Rate and Sensitivity. Bold indicates best DA, Sav., and AS. Stops and Rec. are not highlighted as optimal, as they are task-dependent.

	False Positive Rate					*Sensitivity*				
	DA	**Sav.**	**Stops**	**AS**	**Rec.**	**DA**	**Sav.**	**Stops**	**AS**	**Rec.**
Baselines										
70% Recall	0.963	0.900	1.000	0.719	0.751	0.889	0.900	1.000	0.578	0.751
Knee (2)	0.993	0.747	0.993	0.726	0.909	0.981	0.747	0.993	0.585	0.909
Knee (-)	**1.000**	0.079	0.169	0.382	1.000	**1.000**	0.079	0.169	0.487	1.000
SCAL (1.05)	**1.000**	0.600	0.959	0.695	0.983	**1.000**	0.600	0.959	0.578	0.983
SCAL (0.9)	**1.000**	0.560	0.821	0.584	0.982	**1.000**	0.560	0.821	0.585	0.982
Target (3)	**1.000**	0.486	0.905	0.647	0.999	0.996	0.486	0.905	0.578	0.999
Target (5)	**1.000**	0.332	0.916	0.679	0.999	0.997	0.332	0.916	0.547	0.999
Confidence-based										
Harmonic Shrinkage	0.889	**0.956**	1.000	0.719	0.180	0.807	**0.933**	0.978	0.600	0.175
Confidence Boundary	0.970	0.853	0.896	**0.822**	0.309	0.926	0.741	0.778	**0.800**	0.415

the margin of error of the point estimate *after* all relevant studies have been retrieved.

Note that Appropriate Stops reflects alignment with the statistical sufficiency of the full evidence, not a cost-sensitive utility model. The evaluation, therefore, assesses *decision consistency under partial evidence*.

Let S_o be a binary variable indicating whether the algorithm stopped for outcome o (1 = stopped, 0 = did not stop) and I_o be a binary variable indicating whether information was sufficient for outcome o (1 = sufficient, 0 = insufficient). Appropriate stopping is then calculated as the percentage of outcomes where the stopping decision was appropriate:

$$\text{Appropriate Stops} = \frac{1}{|O|} \sum_{o \in O} \mathbb{I}(S_o = I_o). \tag{5}$$

Recall (Rec.) The percentage of *all relevant* documents in the collection reviewed before the stopping point, also averaged over all outcomes. Recall remains an informative measure, although, unlike in many other Information Retrieval scenarios, our goal here is not to maximise it.

5 Results

Table 1 presents the aggregated performance of the proposed confidence-based stopping algorithms compared against baseline methods across 135 DTA review outcomes. Results demonstrate that the proposed approaches, particularly Confidence Boundary, substantially reduce the number of documents that need to be examined while largely preserving decision agreement.

Savings for the majority of baselines (excluding 70% Recall which cannot be deployed in practice) are lower than those of the proposed approaches, demonstrating that baselines require more of the collection to be examined prior to stopping screening. Several of the baselines (Knee (-), Target (3) and Target (5)) require more than half of the collection to be examined while the proposed approaches reduce this to around a quarter, or less. These savings are reflected in the Recall scores which are noticeably lower for the proposed approaches than the baselines which typically retrieve almost all relevant documents before stopping, which the Decision Agreement scores demonstrate is not necessary.

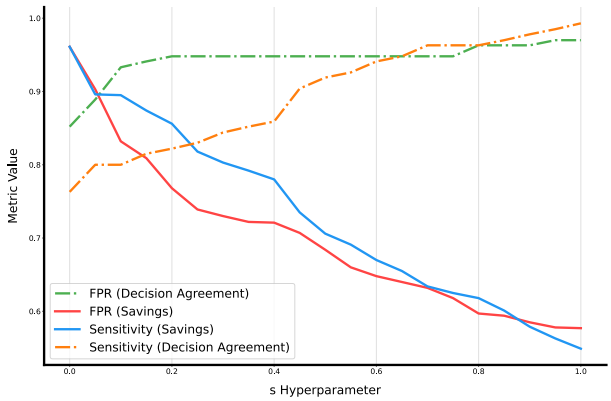

Figure 1: Effect of changing the s hyperparameter for Harmonic Shrinkage. Increasing s results in greater DA, at the cost of savings.

Efficiency gains were achieved while largely maintaining high effectiveness. Although some baselines achieved perfect DA, the top-performing of the proposed approaches also yielded high DA scores. Confidence Boundary consistently achieved high agreement (DA: 97% FPR, 92.6% Sens).

Confidence Boundary also proved able to make appropriate stopping decisions, outperforming all other methods on the AS metric. This highlights the core benefit of aligning the stopping decision with statistical sufficiency of the accumulated evidence rather than recall targets, which are agnostic to whether the review question can yet be answered reliably. In contrast, baseline methods exhibited lower AS scores ($\leq$72.6%), precisely because their stopping criteria (e.g., target recall) are independent of final information sufficiency.

Varying the value of hyperparameter s in the Harmonic Shrinkage approach had the expected effect: increasing s resulted in greater decision agreement but reduced savings, and vice versa, as shown in Figure 1.

The stopping methods proposed here may struggle with small sample sizes and skewed data distributions, which are common challenges in systematic reviews, particularly for those addressing

rare diseases or conditions. With few data points, estimates become unstable and sensitive to individual studies, potentially leading to premature or delayed stopping.

The approaches also assume that the sampling distribution of the point estimates is approximately normal. While the Central Limit Theorem suggests that this assumption holds for sufficiently large samples, it may not for small sample sizes.

6 Conclusion

This paper has introduced an alternative approach to TAR stopping that focuses on whether retrieved information is sufficient to support a downstream decision rather than on whether a predefined recall threshold has been reached. Experimental results from systematic DTA reviews demonstrate that the proposed approaches can greatly reduce relevant document screening requirements while preserving decision agreement. By monitoring evolving point estimates and uncertainty for key clinical metrics, these approaches stop when the accumulated evidence appears sufficient for the thresholded decision. Overall, the results suggest benefits in shifting from recall-oriented stopping methods to confidence-based approaches for stopping decisions.

The underlying principle, that stopping should be governed by evidence sufficiency for a thresholded decision rather than by recall targets, may transfer to other Information Retrieval settings, but the current study evaluates the methods exclusively on DTA systematic reviews. Extending the evaluation to intervention reviews, where outcome structure differs, and to non-medical TAR domains remains important future work.

Future work could extend this evidence-sufficiency framing into a more formal decision-theoretic approach, allowing practitioners to specify task-specific costs, benefits, and stopping criteria aligned with their information needs.

Acknowledgments

This work was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

[1] Leif Azzopardi. 2011. The economics in interactive information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 15–24. doi:10.1145/2009916.2009923

[2] Leif Azzopardi, Diane Kelly, and Kathy Brennan. 2013. How query cost affects search behavior. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 23–32. doi:10.1145/2484028.2484049

[3] Reem Bin-Hezam and Mark Stevenson. 2024. RLStop: A Reinforcement Learning Stopping Method for TAR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2604–2608. doi:10.1145/3626772.3657911 arXiv:2405.02525 [cs].

[4] Reem Bin-Hezam and Mark Stevenson. 2025. A Generalised and Adaptable Reinforcement Learning Stopping Method. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. ACM, 761–770. doi:10.1145/3726302.3729879

[5] David C. Blair and M. E. Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM* 28, 3 (March 1985), 289–299. doi:10.1145/3166.3197

[6] Michiel Bron, Peter GM van der Heijden, Ad Feelders, and Arno Siebes. 2025. Using Chao's estimator as a stopping criterion for technology-assisted review. *ACM Transactions on Information Systems* 43, 3 (2025), 1–51.

[7] Max W Callaghan and Finn Müller-Hansen. 2020. Statistical stopping criteria for automated screening in systematic reviews. *Syst. Rev.* 9, 1 (Nov. 2020), 273. doi:10.1186/s13643-020-01521-4

[8] William S. Cooper. 1973. On selecting a measure of retrieval effectiveness part II. Implementation of the philosophy. *Journal of the American Society for Information Science* 24, 6 (Nov. 1973), 413–424. doi:10.1002/asi.4630240603

[9] Gordon V. Cormack and Maura R. Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. arXiv:1504.06868 [cs.IR] https://arxiv.org/abs/1504.06868

[10] Gordon V. Cormack and Maura R. Grossman. 2016. Engineering Quality and Reliability in Technology-Assisted Review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 75–84. doi:10.1145/2911451.2911510

[11] Gordon V. Cormack and Maura R. Grossman. 2016. Scalability of Continuous Active Learning for Reliable High-Recall Text Classification. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 1039–1048. doi:10.1145/2983323.2983776

[12] Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. 2010. Overview of the TREC 2010 Legal Track. In *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010 (NIST Special Publication, Vol. 500-294)*. National Institute of Standards and Technology (NIST), 1–9.

[13] Giorgio Maria Di Nunzio. 2018. A Study of an Automatic Stopping Strategy for Technologically Assisted Medical Reviews. In *Advances in Information Retrieval*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer International Publishing, Cham, 672–677. doi:10.1007/978-3-319-76941-7_61

[14] Maureen Dostert and Diane Kelly. 2009. Users' stopping behaviors and estimates of recall. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 820–821. doi:10.1145/1571941.1572145

[15] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. 2016. TREC 2016 Total Recall Track Overview.. In *TREC*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA.

[16] Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. 2019. *Cochrane handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Chichester, UK. doi:10.1002/9781119536604

[17] Noah Hollmann and Carsten Eickhoff. 2017. Ranking and Feedback-based Stopping for Recall-Centric Document Retrieval. In *CLEF (working notes)*. CEUR-WS.org, Dublin, Ireland, 7–8.

[18] Fereshte Ilani, Mohsen Nowkarizi, and Sholeh Arastoopoor. 2024. Analysis of the factors affecting information search stopping behavior: A systematic review. *Journal of Librarianship and Information Science* 56, 3 (March 2024), 796–808. doi:10.1177/09610006231157091

[19] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview, In 18th Working Notes of CLEF Conference and Labs of the Evaluation Forum. *CEUR Workshop Proceedings* 1866, 1–29. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85034732447&partnerID=40&md5=a183b346edceb1918338abf473a69dcd

[20] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2018. CLEF 2018 technologically assisted reviews in empirical medicine overview, In Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. *CEUR Workshop Proceedings* 2125. https://strathprints.strath.ac.uk/66446/

[21] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview, In Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019. *CEUR Workshop Proceedings* 2380. https://strathprints.strath.ac.uk/71253/

[22] Wojciech Kusa, Guido Zuccon, Petr Knoth, and Allan Hanbury. 2023. Outcome-based Evaluation of Systematic Review Automation. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval* (Taipei, Taiwan) *(ICTIR '23)*. Association for Computing Machinery, New York, NY, USA, 125–133. doi:10.1145/3578337.3605135

[23] David D. Lewis, Eugene Yang, and Ophir Frieder. 2021. Certifying One-Phase Technology-Assisted Reviews. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 893–902. doi:10.1145/3459637.3482415

[24] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research* 5 (2004), 361–397. doi:10.5555/1005332.1005345

[25] Dan Li and Evangelos Kanoulas. 2020. When to Stop Reviewing in Technology-Assisted Reviews: Sampling from an Adaptive Distribution to Estimate Residual

Relevant Documents. *ACM Trans. Inf. Syst.* 38, 4, Article 41 (Sept. 2020), 36 pages. doi:10.1145/3411755

[26] Graham Mcdonald, Craig Macdonald, and Iadh Ounis. 2020. How the Accuracy and Confidence of Sensitivity Classification Affects Digital Sensitivity Review. *ACM Trans. Inf. Syst.* 39, 1, Article 4 (Oct. 2020), 34 pages. doi:10.1145/3417334

[27] Alessio Molinari and Andrea Esuli. 2024. SALτ: efficiently stopping TAR by improving priors estimates. *Data Mining and Knowledge Discovery* 38, 2 (2024), 535–568. doi:10.1007/s10618-023-00961-5

[28] Christopher Norman, Mariska Leeflang, and Aurélie Névéol. 2018. Data Extraction and Synthesis in Systematic Reviews of Diagnostic Test Accuracy: A Corpus for Automating and Evaluating the Process. *AMIA Annual Symposium proceedings.* 2018 (2018), 817–826.

[29] Christopher R. Norman, Mariska M. G. Leeflang, Raphaël Porcher, and Aurélie Névéol. 2019. Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *Systematic Reviews* 8, 1 (2019). doi:10.1186/s13643-019-1162-x

[30] Giorgio Maria Di Nunzio and Evangelos Kanoulas. 2023. Special issue on technology assisted review systems. *Intelligent Systems with Applications* 20 (2023), 200260. doi:10.1016/j.iswa.2023.200260

[31] Douglas W. Oard, Fabrizio Sebastiani, and Jyothi K. Vinjumur. 2018. Jointly Minimizing the Expected Costs of Review for Responsiveness and Privilege in E-Discovery. *ACM Trans. Inf. Syst.* 37, 1, Article 11 (Nov. 2018), 35 pages. doi:10.1145/3268928

[32] Robin R. Pennington and Andrea Seaton Kelton. 2016. How much is enough? An investigation of nonprofessional investors information search and stopping rule use. *International Journal of Accounting Information Systems* 21 (2016), 47–62. doi:10.1016/j.accinf.2016.04.003

[33] Chandra Prabha, Lynn Connaway, Lawrence Olszewski, and Lillie Jenkins. 2007. What is Enough? Satisficing Information Needs. *Journal of Documentation* 63 (01 2007), 74–89. doi:10.1108/00220410710723894

[34] Johannes B. Reitsma, Afina S. Glas, Anne W.S. Rutjes, Rob J.P.M. Scholten, Patrick M. Bossuyt, and Aeilko H. Zwinderman. 2005. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 58, 10 (2005), 982–990. doi:10.1016/j.jclinepi.2005.02.022

[35] Adam Roegiest, Gordon V Cormack, Charles LA Clarke, and Maura R Grossman. 2015. TREC 2015 Total Recall Track Overview.. In *TREC*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA.

[36] Tefko Saracevic. 2022. *The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?* Springer Nature, Springer, Cham. doi:10.1007/978-3-031-02302-6

[37] Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison. 1988. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science* 39, 3 (1988), 161–176. doi:10.1002/(SICI)1097-4571(198805)39:3<161::AID-ASI2>3.0.CO;2-0

[38] Ian Shemilt, Antonia Simon, Gareth J. Hollands, Theresa M. Marteau, David Ogilvie, Alison O'Mara-Eves, Michael P. Kelly, and James Thomas. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods* 5, 1 (2014), 31–49. doi:10.1002/jrsm.1093

[39] Mark Stevenson and Reem Bin-Hezam. 2023. Stopping Methods for Technology-assisted Reviews Based on Point Processes. *ACM Trans. Inf. Syst.* 42, 3, Article 73 (Dec. 2023), 37 pages. doi:10.1145/3631990

[40] Wan-Ching Wu and Diane Kelly. 2014. Online search stopping behaviors: An investigation of query abandonment and task stopping. *Proceedings of the American Society for Information Science and Technology* 51, 1 (2014), 1–10. doi:10.1002/meet.2014.14505101030

[41] Eugene Yang, David D. Lewis, and Ophir Frieder. 2021. On minimizing cost in legal document review workflows. In *Proceedings of the 21st ACM Symposium on Document Engineering* (Limerick, Ireland) *(DocEng '21)*. Association for Computing Machinery, New York, NY, USA, Article 30, 10 pages. doi:10.1145/3469096.3469872

[42] Lisl Zach. 2005. When is "enough" enough? Modeling the information-seeking and stopping behavior of senior arts administrators. *Journal of the American Society for Information Science and Technology* 56, 1 (2005), 23–35. doi:10.1002/asi.20092