Utility-Based Stopping Methods

Anonymous Author(s)

Abstract

Conducting efficient systematic reviews is crucial given the growing volume of scientific literature. Technology Assisted Review (TAR) offers tools to streamline this process, but traditional stopping approaches focus on achieving specific target recall levels, which may not always satisfy specific information needs. This paper introduces utility-based stopping algorithms, including Harmonic Shrinkage and Confidence Boundary, that prioritise the value of retrieved information in addressing a defined information need. These algorithms analyse statistical properties of point estimates, such as sensitivity and false positive rate, to determine when sufficient evidence has been gathered. Experiments using CLEF 2017-2019 datasets demonstrate that these moment-based approaches can significantly reduce the percentage of reviewed relevant documents (PRD) while maintaining high decision agreement. Notably, these methods achieve substantially reduced PRD while maintaining decision agreement when compared to traditional target-recall approaches. This research advances TAR methodologies by shifting from recall-oriented stopping rules to utility-driven approaches, aligning stopping decisions with users' information needs to enhance the efficiency and effectiveness of document review.

CCS Concepts

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

Keywords

Technology Assisted Review (TAR), Stopping Rules, Information Need, Utility, Statistical Moments, Efficiency, Effectiveness, Point Estimates, Systematic Reviews.

ACM Reference Format:
Anonymous Author(s). 2025. Utility-Based Stopping Methods. In *Proceedings of Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Re- trieval, (SIGIR '25).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Technology Assisted Review (TAR) aims to develop methods to support information identification when examining the entire collection is impractical. Key applications include the development of systematic reviews in areas such as medicine and software engineering [20] and eDiscovery in the legal domain [19, 34, 38, 42].

Within TAR, stopping rules help reviewers decide when to stop examining documents, thereby reducing workload by reviewing no more documents than necessary. A wide range of stopping methods have been proposed, all 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* [30, 46]. However, basing the stopping decision on target recall does not take account of whether a users' information need has been satisfied. Some information needs might be satisfied with information contained within a single document, even when many other relevant documents in the collection have not been discovered. On the other hand, information needs may not be satisfied at all in a given collection. Therefore, relying solely on recall as a stopping criterion may be inadequate for determining when a users' information need has been fulfilled. Recall relies on the concept of relevance, the "aboutness" of a document to query, which has been contrasted with utility, i.e. the value of a document to an information seeker [43, 44]. Deciding when to stop by considering the utility of the information discovered is more useful than considering the documents' relevance.

This paper introduces several novel utility-based stopping algorithms for TAR that prioritise fulfilling a users' information need over achieving specific recall targets. The proposed algorithms focus on the utility derived from the retrieved documents by analysing the evolving statistical properties of the retrieved information, using metrics that assess the point estimates generated through document retrieval. These stopping algorithms evaluate the utility of the recalled information so far and determine whether that utility is sufficient to justify stopping. In essence, these algorithms seek to determine when the marginal gain in utility from reviewing additional documents is outweighed by the cost and effort involved. This approach allows for a more dynamic and adaptive stopping criterion that aligns more closely with the users' actual needs.

This approach is compared against existing stopping rules (all of which base their decisions on target recall) and found to improve the efficiency of the review process significantly, achieving comparable results to target recall-based methods with substantially less relevant document review.

The contributions of this paper are the development of a range of TAR-stopping algorithms based on utility rather than recall and the evaluation of these algorithms using a real-world dataset of systematic reviews.

2 Background

Previous work on stopping methods has focused on two largely separate areas: stopping methods within TAR pipelines and methods to simulate user behaviour during search tasks.

2.1 Stopping in TAR

Previous work on stopping methods within TAR pipelines have aimed to reduce the number of documents that need to be examined while screening collections for relevance, e.g. [11, 28, 30, 46]. A wide range of approaches have been applied, including examination of

SIGIR '25, July 13–18, 2025, Padua, Italy

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

the rate at which relevant documents are observed [11, 46], estimating the number of remaining relevant documents by sampling or classification [7, 12, 30, 45] and analysis of ranking scores [15, 21].

A common feature of these methods is that they all use the concept of a *target recall*, namely a percentage of all of the relevant documents within the collection, and aim to identify a suitable stopping point after a sufficient number of documents have been examined for the target recall to have been achieved. Evaluation of these stopping methods generally compares the recall achieved at the proposed stopping point against the target recall. The reliability metric [11] measures how often the target recall is reached or exceeded, thereby treating it as a minimum recall for acceptability. An alternative approach measures the difference between the achieved and target recalls, e.g. [30, 46].

There are several advantages to evaluating stopping algorithms in terms of achieved recall. One is that recall is based on the widely accepted notion of relevance, making it straightforward to understand. Another is that a wide range of test collections are available in which documents have been labelled for relevance and these can be used to evaluate stopping algorithms, e.g. CLEF e-Health [23–25], TREC Total Recall [19], TREC Legal [13] and RCV1 [29]. Finally, recall-based evaluation is appropriate for some potential applications of stopping methods, such as within eDiscovery where the parties involved in litigation may agree on a target recall for documents relevant to a case to be disclosed (e.g. 80%) since the collections being reviewed are large enough to make the identification of all relevant documents impractical [49].

However, recall-based evaluation may not always be appropriate since recall alone provides little information about whether an information need has been satisfied. The information needed might be satisfied by a single document, or it might not be fully contained even within the entire set of relevant documents. In addition, recall-based evaluation generally assumes that all documents are of equal value, which is not always true.

This is the case in systematic review automation, a significant use case for TAR methods, where some documents (e.g. those describing high-quality randomised control trials with a large number of participants) contribute more towards the review's final outcome than others. In addition, systematic reviews aim to answer a specific question which may be possible to do without identifying all of the relevant documents within the collection, particularly those with only a minor contribution to its conclusion. Norman et al. [37] found that the number of documents screened could be reduced by more than half without affecting review conclusions.

2.2 Stopping Behaviour

Another strand of relevant work has focussed on analysis of user behaviour during search tasks, including when they decide to stop search sessions. This work suggests that users are motivated by a range of cognitive and environmental factors when deciding when to stop searching [22]. A key cognitive factor was found to be the individual's assessment of information sufficiency, where a feeling that the information being obtained so far is "good enough" motivates the decision to stop searching [8, 17, 40, 50]. Other cognitive factors include the users' experience and skills [39, 48]. Relevant environmental factors are the search task being undertaken, the Information Retrieval system being used, and the time available [3, 17, 48].

Analysis of user interactions within search systems led to the development of stopping rules. This work was largely carried out independently of the work on TAR stopping described above and generally aimed to create rules that could be applied within interactive search scenarios and which modelled observed user behaviour. Early examples of such rules were the *satisfaction rule*, which stopped when a fixed number of relevant documents have been encountered, and the *frustration rule*, where the search is stopped after a fixed number of irrelevant documents have been seen [8]. These rules have also been combined [26]. Stopping rules have also been developed based on assessments of the sufficiency of the information already encountered. Nickles [35] proposed four such rules: stop when the user does not believe they will learn anything new; stop when the user is satisfied with the cumulative amount of information has been acquired; stop when information has been obtained about a predefined list of criteria; stop when understanding of a topic has stabilised. Browne et al. [6] proposed a further rule where the user gathers information about a single criterion (the most important) and stops when sufficient information has been acquired. Experiments simulating search sessions revealed that variants of the frustration rule typically performed well both in terms of identifying suitable stopping points and modelling user behaviour, although the simple approach of stopping after a fixed number of documents also performed well [31–33].

Decision theory and economics methods have also been applied to propose stopping when the benefits of continuing the search process are outweighed by the cost of doing so [2, 9].

This paper aims to develop TAR-stopping rules that go beyond the current recall-based approaches and to develop stopping algorithms based on when sufficient information has been identified to satisfy an information need. The approaches are similar to those developed by Nickles [35] to model search behaviour, although these have not yet been applied to TAR problems.

3 Problem Formulation

Systematic reviews aim to analyse currently available evidence to answer a specific question, such as the effectiveness of a treatment for a given condition for a particular group of individuals. This work focuses on developing methods for identifying when enough information has been discovered to answer the question posed by the review. This contrasts existing stopping methods for systematic reviews, which aim to identify a fixed portion of the evidence, regardless of whether or not it provides enough information to answer the question. Following existing work [27], the effectiveness of these methods is assessed in terms of outcomes, i.e. whether or not the question was successfully answered.

Diagnostic Test Accuracy (DTA) reviews are an important type of systematic review that aims to summarise evidence regarding the accuracy of a medical test, such as a lateral flow test for COVID-19 [16]. DTA reviews are important in evidence-based medicine as the collated evidence synthesis they provide informs the appropriateness of a test for a particular circumstance and how their results should be interpreted. The majority of systematic reviews used in the CLEF e-Health task on "Technology Assisted Reviews in Empirical Medicine" were DTA reviews and were selected due to the challenge involved in identifying relevant studies [23–25].

DTA reviews are created by identifying research publications that report on a test's accuracy and combining their individual estimates together to provide a single overall estimate which can be regarded as the best possible estimate given the information available. They report test performance in terms of their *sensitivity* and *specificity*. Sensitivity measures the test's ability to identify individuals with the condition being tested for and represents the proportion successfully identified [1]. This metric is identical to precision and is computed similarly, i.e., the number of true positives divided by the sum of true positives and false negatives. A test's specificity measures its ability to correctly identify patients without a condition. It is computed as the number of true negatives divided by the sum of true negatives and false positives [18]. This test property may be reported as the *false positive rate* computed as 1 - specificity. Similar to the balance between precision and recall, diagnostic tests generally have a trade-off between sensitivity and specificity/false positive rate. This balance determines the appropriate circumstances for a test to be used. For example, large-scale screening would require a test with high sensitivity to avoid missing individuals with the condition being tested. In contrast, a test with a low false positive rate would be more appropriate before prescribing a high-risk intervention to avoid over-treatment.

Many types of systematic review aim to answer a binary question, such as whether a drug is superior to another in a particular set of conditions. The information need has been answered for these types of reviews when enough information has been gathered to answer the question. DTA reviews are somewhat different since they provide estimates of test performance, which can continue to be refined as more information is gathered, although the value of additional information generally diminishes. Although DTA reviews answer a clear information need about test accuracy, it is difficult to determine exactly at what point this has been met or what level of performance regarding the test results is required to do so since this largely depends on how an individual plans to make use of the systematic review. Consequently, we consider the information need for each DTA review to be a binary question: "Is the performance of this test > θ?", for some threshold θ and where performance is the value for some metric (e.g. sensitivity or false positive rate). This approach allows the information need to be stated clearly and simplifies the process of determining whether or not it has been met.

The best available estimates of the true values for test accuracy are those calculated by combining the results from all included studies in the review, and consequently, these are treated as gold standard values. Retrieving all relevant documents and combining their information would produce the same estimates and, therefore, the same answer to whether the test accuracy exceeds θ. In contrast, if a stopping algorithm only returns a subset of all relevant studies, only some of this information will be available, and the estimate of test accuracy may differ from the one that would be produced given all the information. Consequently, the answers to whether the test accuracy exceeds θ given all or a subset of the information may differ. The aim is to develop stopping methods that provide the correct answer to the binary question (i.e. the answer that would be given when all available information is considered) while minimising the number of documents that need to be examined. That is, stopping algorithms aim to minimise the number of documents that need to be reviewed to answer the question correctly.

4 Stopping Methods

Assume that documents in a collection have been ranked and examined in the ranked order. Let i be the ith relevant document encountered in the ranking. For each relevant document a point estimate, $\hat{p}_i$, of the metric of interest (e.g. sensitivity or false positive rate) can be extracted. The goal is to find the smallest index, I^*, such that sufficient information has been aquired for the binary question to be answered.

This paper introduces utility-based stopping algorithms that leverage the statistical properties of point estimates to determine when sufficient information has been gathered to satisfy the information need. The methods described in the following sections examine distributional properties of the cumulative $\hat{p}_i$. Each approach monitors one or more of the moments: mean (first moment), variance (second moment), skewness (third), and kurtosis (fourth) [5, 47]—as new documents are incorporated. These moments capture different indicators of convergence in the estimate's distribution. Moments are a framework in statistical analysis for characterising how data behave as they accumulate since each moment captures a different aspect of the distribution's shape. A stable mean and small variance signal that additional studies are unlikely to shift the central value significantly, whereas low skewness and kurtosis suggest that the distribution is well-balanced and not unduly influenced by outliers.

The rationale for this approach is that as relevant information accumulates, the statistical distribution of the performance estimates (like sensitivity and false positive rate) should exhibit predictable changes reflecting convergence and stability. Moments provide a quantifiable way to track these changes and infer when the marginal utility of reviewing further documents diminishes.

4.1 Harmonic Shrinkage

Early in retrieval, a $\hat{p}_i$ further from θ provides stronger evidence for stopping. This is because subsequent information accumulation is unlikely to significantly alter future $\hat{p}_i$, as decisions based on these estimates are also unlikely to change. Conversely, $\hat{p}_i$ within an upper and lower boundary from θ suggest additional information retrieval is required before making a decision, because the estimate is not yet stable enough for confident decision-making.

As more information is gathered, the confidence required for stopping can decrease. This decreasing confidence requirement is modelled through harmonic shrinkage of temporary decision thresholds. As more data is retrieved, these temporary thresholds gradually converge towards the final decision boundary, following Equation 1 and illustrated with a θ of 0.5 in Figure 1.

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

This approach shrinks upper and lower temporary decision boundaries until all previous point estimates are outside or on the $I'th$ decision boundaries. A shrinkage factor, s, controls the

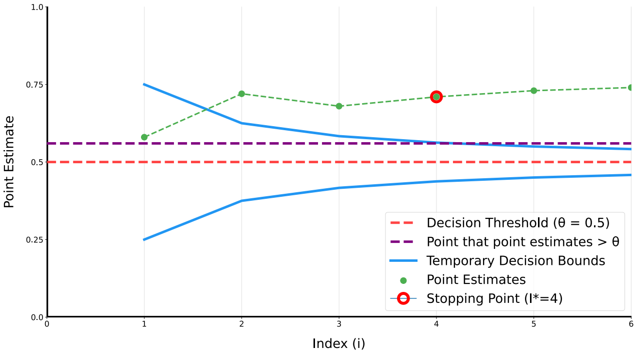

Figure 1: A simulated study run using the Harmonic shrinkage stopping rule. Note stopping at Index 4 because all previous point estimates are above the temporary decision bounds.

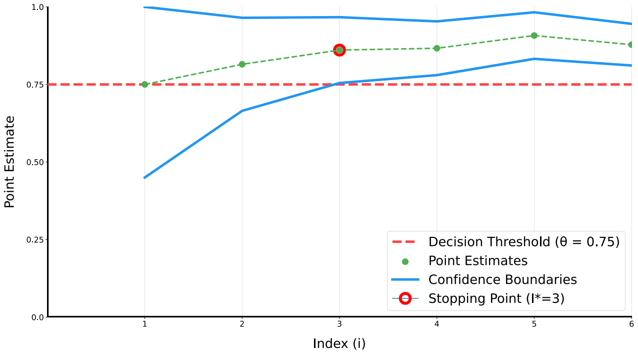

Figure 2: A simulated study run using the Confidence boundary stopping rule. Note stopping at index 3 due to a lower confidence boundary above the decision threshold.

convergence rate of the decision boundaries and represents the required decision confidence. When $s = 1$, the rule demands more extreme point estimates for stopping, reflecting a high confidence requirement. When s = 0, the temporary boundaries collapse to the decision threshold θ, indicating that any point estimate can trigger a decision, regardless of the distance from the final threshold. If $\hat{p}_i$ crosses θ from above or below at any point, the algorithm will never stop, as it requires all previous point estimates to be on the same side.

4.2 Confidence Boundary

An alternative approach is to construct confidence intervals around each $\hat{p}_i$ using estimated variance. This approach addresses one of the key limitations of the harmonic shrinkage method, namely its sole reliance on the first moment and its failure to consider for the variability of $\hat{p}_i$. A confidence interval is calculated from cumulative data so far, providing a range within which the true value of the mean is likely to lie with a certain level of confidence. The Confidence Boundary approach is outlined in Equation 2 and demonstrated using a θ of 0.75 in Figure 2.

$$I^* = \arg\min_i \left(\left(\hat{p}_i + \frac{z\alpha}{2} \cdot \frac{\sigma_i}{\sqrt{n_i}} < \theta \right) \vee \left(\hat{p}_i - \frac{z\alpha}{2} \cdot \frac{\sigma_i}{\sqrt{n_i}} > \theta \right) \right) \quad (2)$$

Where μ_i and σ_i are, respectively, the mean and standard deviation of the point estimate distribution.

In simpler terms, the stopping criterion is met when either the interval's lower confidence boundary is above θ, or the upper confidence boundary is below θ. This implies that, with the chosen confidence level (set here at 95%), the retrieved information provides sufficient statistical evidence to conclude that the true mean is either above or below the decision threshold.

4.3 Skewness

Skewness (a measure of asymmetry) can be used to inform stopping decisions as when more evidence is accumulated through document retrieval, the $\hat{p}_i$ distribution is expected to become less skewed, indicating convergence towards a stable value. This method assumes that a symmetric distribution (i.e., skewness approaching zero) suggests sufficient information has been gathered to yield a reliable $\hat{p}_i$.

Skewness is calculated according to Equation 3.

$$\text{Skewness}_i = \frac{\sum x = 1^{n_i} (\hat{p}x, i - \mu_i)^3}{n_i * \sigma_i^3} \quad (3)$$

I^* is determined when the absolute value of the calculated skewness is below a predefined threshold, γ, formally expressed in Equation 4.

$$I^* = \arg\min_i (|\text{Skewness}_i| < \gamma) \quad (4)$$

The hyperparameter choice, γ, can be considered a required confidence level within the point estimate needed before terminating information retrieval. Smaller γ values enforce stricter symmetry requirements, potentially leading to prolonged retrieval but more confidence in the point estimate. Larger γ values might result in premature stopping.

4.4 Kurtosis

Kurtosis measures the concentration of data around the peak and the heaviness of the tails. As more data is accumulated through document retrieval, the $\hat{p}_i$ distribution is anticipated to approach a stable shape, characterised by a specific kurtosis value. This method leverages changes in kurtosis to inform stopping decisions, suggesting that a stable distribution shape, as reflected in a relatively constant kurtosis, indicates that variability is consistent. This approach assumes that a kurtosis value approaching that of a normal distribution (i.e., excess kurtosis approaching zero) suggests sufficient information has been gathered to yield a reliable point estimate.

This stopping method is outlined in Equation 5:

$$\text{Kurtosis}_i = \frac{\sum x = 1^{n_i} (\hat{p}x, i - \mu_i)^4}{n_i \times \sigma_i^4} - 3 \quad (5)$$

The subtraction of 3 adjusts the kurtosis value to represent “excess kurtosis”, where a value of 0 corresponds to the kurtosis of a normal distribution [14]. Positive excess kurtosis (leptokurtic) indicates a distribution with heavier tails and a sharper peak than a normal distribution. In comparison, negative excess kurtosis (platykurtic) indicates lighter tails and a flatter peak [4].

As with skewness, I^* is determined when the absolute value of the calculated kurtosis is below γ, see Equation 6.

$$I^* = \arg\min_i \left(|\text{Kurtosis}_i| < \gamma \right) \tag{6}$$

4.5 Multi-moments

While a single-stopping criterion might lack sufficient discriminatory power or be overly sensitive to specific data characteristics, a collective assessment based on multiple moments can provide a more reliable and nuanced decision-making framework. By considering multiple facets of the evolving distribution, this method aims to capitalise on the early stopping potential offered by lower-moment methods like Harmonic Shrinkage while also assessing the stability and shape considerations provided by higher-moment methods like skewness and kurtosis-based stopping. The combined approach also benefits from the variance-aware nature of confidence boundary-based stopping.

The integration mechanism can be conceptualised as an "OR" operation, where the information retrieval process is terminated at the earliest index when any of the individual stopping criteria are met, as outlined in Equation 7.

$$I^* = \arg\min_i \left(\text{HS}(i) \lor \text{CB}(i) \lor \text{SK}(i) \lor \text{KU}(i) \right) \tag{7}$$

This ensures retrieval aligns with resource constraints. The D_{max} value is search-specific, reflecting domain priorities and resource availability.

5 Experiments

This section evaluates our proposed utility-based stopping algorithms against established recall-based baselines. DTA reviews from the CLEF datasets illustrate how utility-based stopping performs in a real-world context by identifying sufficient rather than complete evidence. These experiments aim to answer two key questions: First, do utility-based stopping methods maintain high decision agreement? Second, do they reduce the number of documents reviewed while maintaining high decision agreement?

5.1 Dataset Description

Norman et al. [36] developed the Limsi-Cochrane dataset to aid research in this domain. This dataset, derived from 63 DTA systematic reviews, captures crucial information needed for developing and evaluating automated stopping algorithms. The Limsi-Cochrane dataset was constructed using a combination of optical character recognition technology (Tesseract[1]), manual verification and post-editing.

This work obtained a low extraction error rate (0.06-0.3%) encompassing 5,848 test results from 1,354 diagnostic tests and 1,738 diagnostic studies. Each review within the dataset contains one or more outcomes (sets of tests), and each test is linked to a set of underlying studies. PubMed identifiers uniquely identify most studies.

5.2 Dataset Generation

The included studies were grouped by outcome from the Limsi-Cochrane dataset, each containing absolute values for true positives, true negatives, false positives and false negatives. The included studies were matched to the CLEF dataset via their Cochrane Database (CD) and PubMed identification numbers. If this lookup resulted in no matches, a document identifier, if present in the Limsi-Cochrane dataset, was converted to a PubMed identification number using Entrez[2] Python package. This PubMed identification number was then used to query the CLEF dataset for a matching document. Any included studies that could not be matched to the CLEF dataset were removed from the outcome study run. The studies were then ranked using AutoStop [30], an implementation of an active learning approach to document ranking in TAR [10] representing state-of-the-art performance on total recall tasks. This formed the CLEF 2017 and CLEF 2018 DTA datasets.

The Limsi-Cochrane dataset predates the CLEF 2019 dataset; and therefore, it did not contain those systematic reviews. For all the 8 CLEF 2019 DTA systematic reviews, Amazon's Textract optical character recognition technology[3] was used to read each outcome obtained on the Cochrane website. This data contained the absolute values for true positives, true negatives, false positives and false negatives. These results were manually linked to a PubMed identification number if one was reported on the Cochrane website, generating the CLEF 2019 DTA dataset.

Each dataset had an ordered list of included studies for each outcome within a systematic review. With each included research study, a position within the AutoTAR ranking and values for true positives, true negatives, false negatives and false positives were available. At each point of included research, sensitivity and false positive rate were calculated using the Reistma R Package[4], which uses a bivariate approach [41]. In total, 135 outcomes study runs were available after the dataset matching process that met the minimum number (5) of relevant studies threshold (90 from CLEF 2017, 29 from CLEF 2018 and 16 from CLEF 2019 DTA).

5.3 Hyperparameter selection

In diagnostic testing, appropriate decision thresholds (θ) are crucial for determining test outcomes. For this experiment, the decision thresholds were established based on values generally considered desirable within the broader domain of diagnostic testing, particularly for screening purposes, given the context-agnostic nature of this experimental setup.

Sensitivity's θ was set to 0.75. A sensitivity of 0.75 may be acceptable in screening scenarios where the primary objective is to identify a substantial proportion of individuals potentially affected by a disease and warrant further confirmatory testing. This is especially relevant in cases where the screening test is relatively inexpensive and non-invasive compared to subsequent diagnostic procedures. A higher sensitivity is often preferred in early-stage

[1] https://tesseract-ocr.github.io/

[2] https://pypi.org/project/entrezpy/

[3] https://aws.amazon.com/textract/

[4] https://www.rdocumentation.org/packages/mada/versions/0.5.11/topics/reitsma

screening to avoid missing cases. False positive rate's θ was set to 0.1. Lower false positive rate values are generally preferred as they minimise the risk of misclassifying healthy individuals as having the condition. The acceptability of this rate is contingent upon the potential ramifications of false positives, such as undue anxiety, unnecessary follow-up tests, or unwarranted treatments. A false positive rate of 0.1 might be considered acceptable if the follow-up confirmatory tests are relatively accurate and the consequences of a false positive are not severe.

While these thresholds are reasonable starting points for this context, it is important to acknowledge that selecting appropriate decision thresholds (θ) for each outcome (sensitivity, false positive rate) is ideally informed by the specific diagnostic testing context, including the disease's nature, the condition's prevalence, and the relative consequences of false positive and false negative classifications. A universal threshold is unlikely to be optimal across all scenarios, and further research could explore the impact of varying these thresholds within specific contexts.

s was set to 0.25, γ was set to 0.5, as these values showed promise in preliminary empirical testing as reasonable starting points balancing early stopping and decision stability. Both parameters can be interpreted as considered confidence levels, with higher values of s and lower values of γ corresponding to higher confidence levels.

5.4 Baselines

The proposed utility-based stopping algorithms were compared against several baselines. All of these are *idealised* in that they rely on perfect knowledge of the relevant documents in the collection, which is not available in a real-world review scenario.

First Relevant Stop: A simple baseline where information retrieval ceases after encountering the first relevant document. This represents an extreme version of minimal effort.

70% Recall Stop: Information retrieval is terminated once at least 70% of relevant documents are retrieved. This represents a common heuristic used in some TAR applications, aiming to balance search effort with the desire to find a substantial portion of relevant material.

Last Relevant Stop: This is an idealised baseline method where retrieval stops immediately after the last relevant document has been found. It is not realistically achievable but is an upper bound on efficiency.

All stopping algorithms (utility-based and baselines) were applied to all outcomes.

5.5 Evaluation metrics

Following Kusa et al. [27], a stopping algorithm's effectiveness was evaluated by calculating the proportion of time which the decision when the algorithm stops is the same as the decision would have been given all potential evidence, i.e. comparing the decision at index I^* against that at n. The decision at index i is a binary value given by

$$Decision(i) = \begin{cases} 1, & \text{if } \hat{p}_i > \theta \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Then the decision agreement (**DA**) for a set of outcomes, O, is calculated as the proportion of times the decisions agree over all outcomes, i.e.

$$\text{Decision Agreement} = \frac{\sum_O \mathbb{I}(Decision(I^*), Decision(n))}{|O|} \tag{9}$$

where O is a set of outcomes and $\mathbb{I}$ an indicator function that returns 1 when the values are identical and 0 otherwise.

As this research aims to determine if utility-based approaches can make high-quality decisions but with less information, efficiency was additionally measured by reporting the percentage of relevant documents (**PRD**) looked at (i.e., out of the total relevant documents within an outcome, how many were recalled before arriving at the stopping decision). In line with existing research, a stopping algorithm's retrieval efficiency is measured by calculating document retrieval savings (**Savings**). The calculation subtracts the number of documents reviewed before the stopping point from the total documents in the document collection. Lower savings values indicate less efficient retrieval.

A key aspect of utility-based stopping is not just efficiency but also *appropriateness*, i.e., stopping when enough information is available to make a sound decision, but not before. For each outcome, the appropriateness of the stopping algorithm's decision is assessed, quantified as (**App. Stops**). 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 the margin of error of the point estimate *after* all relevant documents have been retrieved (see Figure 3).

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, as shown in Equation 10.

$$\text{Appropriate Stops} = \frac{\sum_O \mathbb{I}(S_i = I_i)}{|O|} \tag{10}$$

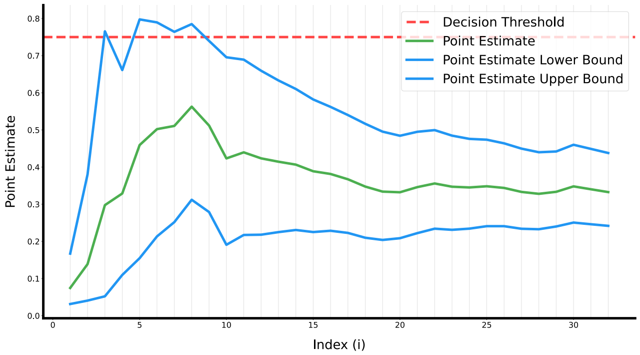

Figure 3: Sufficency of information demonstration on run data as at the last index, the decision threshold is outside of the margin of error.

The frequency of a stopping algorithm's stopping frequency was also reported as a percentage (**Stops**).

Due to the paired nature of the data and the small sample sizes in some of the comparisons, a non-parametric Wilcoxon signed-rank test was chosen to evaluate the performance differences between the moment-based approaches and the 70% Recall baseline on DA and PRD, with a Bonferroni correction for multiple comparisons.

6 Results

The performance of the utility-based stopping algorithms is compared against several baseline methods presented in Table 1 for false positive rate and sensitivity.

Decision agreement compares algorithm decisions to those made with complete data. As expected, the baseline Last Relevant Stop method achieved perfect decision agreement, while the First Relevant Stop method showed a wide decision agreement range (0.688-1.000), highlighting the risks of premature stopping. The 70% Recall baseline consistently achieved strong decision agreement (0.931-0.989).

Crucially, none of the moment-based methods showed a statistically significant decrease in decision agreement compared to the 70% Recall baseline, indicating they maintain comparable decision accuracy. Harmonic Shrinkage's decision agreement ranges from 0.793 to 1.000. Confidence Boundary achieves decision agreement ranging between 0.828 and 1.000. While the Multi-moments method typically yields lower decision agreements than individual moment approaches, with its range of decision agreement values matching that of First Relevant Stop (0.688-1.000), it improves upon or matches the First Relevant Stop decision agreement 12/12 times. It potentially suggests an over-aggressiveness in the Multi-moments stopping point but still makes it a superior choice to First Relevant Stop.

The PRD metric indicates the proportion of relevant documents reviewed before a stopping decision is made. As expected, the Last Relevant Stop baseline method reviewed all relevant documents, while the First Relevant Stop method reviewed the least (0.056-0.094). The 70% Recall baseline has a narrow range of PRD (0.654-0.677)

The lower moment-based methods, Harmonic Shrinkage and Confidence Boundary, on average, reviewed a smaller proportion of relevant documents (0.142-0.584 and 0.146-0.365, respectively) than the 70% Recall baseline, which, in combination with their high decision agreements, suggests lower moments are more efficient in identifying when a sufficient subset of relevant documents have been retrieved to make accurate stopping decisions. This trend was statistically significant within the 2017 and 2018 datasets. However, in the 2019 dataset, statistical significance was only observed for Harmonic Shrinkage on the sensitivity point estimate. The lack of significance for other comparisons in the 2019 dataset could be attributed to the small sample size (n=16), which reduces statistical power. Given this small sample size, the absence of statistical significance does not strongly indicate a lack of effect. Conversely, the statistically significant findings in the 2017 and 2018 datasets provide stronger evidence for the effectiveness of the Harmonic Shrinkage and Confidence Boundary methods. Skewness and Kurtosis required more relevant documents before stopping (0.406-0.782). The Multi-moments method exhibits a very low PRD (0.062-0.150), which, combined with similar decision agreement levels with First Relevant Stop, makes it a superior stopping algorithm choice if minimising the review of relevant documents is a priority, even at the risk of poorer decision agreement.

Savings, the proportion of documents not reviewed, highlight the reduction in screening efforts. The Last Relevant Stop achieves substantial savings, ranging from 0.919-0.973, benefiting from perfect knowledge of the last relevant document, and highlights the difficulty of achieving savings while preserving a robust decision agreement. The First Relevant Stop method achieves near-maximum savings (0.994-0.999). The 70% Recall baseline achieves relatively consistent savings (0.961-0.988). All baseline methods savings highly benefit from the effectiveness of screening prioritisation and always having a stopping point.

In cases where a moment-based approach did not stop, it incurred a penalty of the entire document pool (relevant and irrelevant), affecting the lower boundaries of these ranges. Harmonic Shrinkage and Confidence Boundary have a wide range of savings (0.510-0.948). Skewness and Kurtosis generally result in lower savings (0.385-0.975) than the lower moments, which is expected given that higher moments typically require more point estimates to arrive at a stopping decision. The Multi-moments method achieves high savings (0.988-0.998) across all datasets and point estimates.

An analysis of outcomes using lower-moment methods (e.g., Harmonic Shrinkage and Confidence Boundary) revealed a limitation of using overall savings as an evaluation metric. While these methods often reviewed a smaller proportion of relevant documents (lower PRD) than the 70% recall baseline, their overall savings were sometimes lower. This discrepancy arises because a failure of the moment-based algorithms to identify a stopping point resulted in a significant penalty – the review of the entire document set. Figure 5, which illustrates the Confidence Boundary approach on the CLEF 2018 dataset for sensitivity, demonstrates this phenomenon. Although savings were achieved in 68.97% of outcomes, the few instances where the algorithm failed to stop, and thus reviewed all documents, drastically reduced the average savings.

This suggests that not stopping was appropriate for these outcomes, which is supported by the Appropriate Stops metric for the Confidence Boundary approach. The Confidence Boundary approach exceeded the Appropriate Stopping rate of all baseline methods, in all datasets. Notably, all baseline approaches, and the Multi-moments share the same values as they always stop, regardless of whether sufficient information has been obtained. In contrast to target-recall-based approaches, which force a decision even with potentially insufficient information, the Confidence Boundary approach yielded a 10.3% continuation rate, demonstrating the approaches ability to identify cases where further evidence was needed. An example of such a case is given for 70% recall from the CLEF 2019 dataset for the sensitivity point estimate is given in Figure 4. Regardless of *any* specified target recall, stopping is not particularly appropriate, which can be identified with utlity-based approaches.

In conclusion, these results indicate that the confidence boundary approach maintains good decision agreement, utilises less relevant documents and stops appropriately more than the 70% baseline.

Table 1: Performance Comparison of Stopping Algorithms on CLEF Datasets. Note * indicates statistical significance detected between the moment-based approach when compared with a 70% baseline metric using a Wilcoxon signed rank test with a Bonferroni correction performed on DA and PRD ($\alpha = 0.05$).

		False Positive Rate					Sensitivity				
Dataset	**Algorithm**	**DA**	**PRD**	**Savings**	**Stops**	**App. Stops**	**DA**	**PRD**	**Savings**	**Stops**	**App. Stops**
2017	*Baselines*										
	Last Relevant Stop	1.000	1.000	0.973	1.000	0.789	1.000	1.000	0.973	1.000	0.567
	First Relevant Stop	0.767	0.094	0.999	1.000	0.789	0.767	0.094	0.999	1.000	0.567
	70% Recall	0.989	0.745	0.988	1.000	0.789	0.944	0.745	0.988	1.000	0.567
	Moment-based										
	Harmonic Shrinkage	0.889	0.284*	0.827	0.811	0.667	0.833	0.279*	0.799	0.878	0.578
	Confidence Boundary	0.978	0.298*	0.948	0.911	0.878	0.967	0.365*	0.873	0.811	0.756
	Skewness	1.000	0.685	0.486	0.489	0.500	0.944	0.641	0.560	0.567	0.422
	Kurtosis	0.989	0.593	0.567	0.600	0.633	0.956	0.637	0.555	0.611	0.533
	Multi-moments	0.867	0.142*	0.998	1.000	0.789	0.811	0.150*	0.998	1.000	0.567
2018	*Baselines*										
	Last Relevant Stop	1.000	1.000	0.936	1.000	0.655	1.000	1.000	0.936	1.000	0.552
	First Relevant Stop	0.724	0.065	0.994	1.000	0.655	0.724	0.065	0.994	1.000	0.552
	70% Recall	0.931	0.735	0.961	1.000	0.655	0.897	0.735	0.961	1.000	0.552
	Moment-based										
	Harmonic Shrinkage	0.793	0.182*	0.788	0.897	0.621	0.862	0.584*	0.835	0.759	0.651
	Confidence Boundary	0.828	0.146*	0.831	0.897	0.759	0.862	0.220*	0.829	0.897	0.759
	Skewness	0.931	0.452	0.436	0.517	0.379	0.897	0.782	0.385	0.483	0.379
	Kurtosis	0.966	0.479	0.799	0.724	0.793	0.931	0.476	0.761	0.620	0.793
	Multi-moments	0.759	0.081*	0.992	1.000	0.655	0.793	0.126*	0.988	1.000	0.552
2019	*Baselines*										
	Last Relevant Stop	1.000	1.000	0.919	1.000	0.438	1.000	1.000	0.919	1.000	0.689
	First Relevant Stop	0.688	0.056	0.998	1.000	0.438	1.000	0.056	0.998	1.000	0.689
	70% Recall	0.938	0.726	0.968	1.000	0.438	0.875	0.726	0.968	1.000	0.689
	Moment-based										
	Harmonic Shrinkage	0.875	0.354	0.556	0.688	0.750	1.000	0.142*	0.945	0.938	0.750
	Confidence Boundary	0.938	0.361	0.510	0.688	0.750	1.000	0.271	0.887	0.750	0.938
	Skewness	0.875	0.517	0.499	0.562	0.375	0.938	0.410	0.671	0.750	0.688
	Kurtosis	0.875	0.406	0.975	0.812	0.375	1.000	0.472	0.912	0.812	0.625
	Multi-moments	0.812	0.108*	0.995	1.000	0.438	1.000	0.094*	0.997	1.000	0.689

7 Discussion

The experimental results demonstrate the potential and challenges of transitioning from recall-centric to utility-based adaptive stopping algorithms in TAR. By grounding the stopping decision in the *value* rather than fixed recall thresholds, the moment-based stopping algorithms achieved notable gains in PRD while maintaining a robust alignment in decision agreement with outcomes based on full document sets.

Moment-based methods, particularly Harmonic Shrinkage and Confidence Boundary, address this limitation by adapting to the specific characteristics of the data. These methods dynamically assess the evolving statistical properties of key metrics (e.g., sensitivity, false positive rate) as more documents are reviewed. Instead of aiming for a fixed recall target, they monitor the convergence of these metrics, using their moments (mean, variance, skewness, kurtosis) to infer when additional information is unlikely to alter the overall conclusions significantly. The results show that Harmonic Shrinkage and Confidence Boundary often achieve comparable or superior decision agreement to the 70% Recall baseline while reviewing a smaller proportion of relevant documents, underscoring their adaptive nature.

Some moment-based methods, especially those based on higher moments (Skewness and Kurtosis), tend to delay stopping, potentially reviewing a large proportion of the document collection. This can negate the efficiency gains of early stopping. The experiments also revealed trade-offs associated with using higher moments (Skewness and Kurtosis) for stopping decisions. While these methods generally resulted in higher decision agreement compared to lower moments (harmonic shrinkage or confidence boundary), they achieved this at the cost of reduced savings. This is likely because higher moments typically require more data points

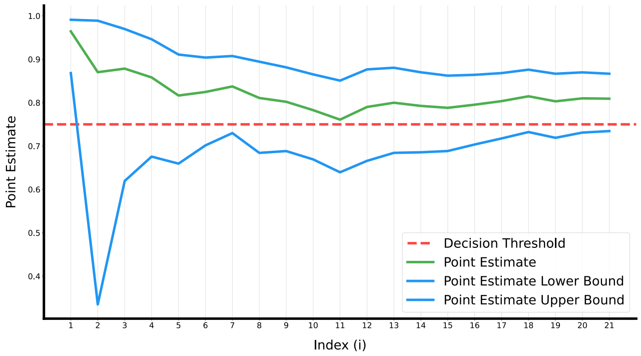

Figure 4: Example from CLEF 2019 dataset, demonstrating cases where specifying any target recall level would result in an inappropriate decision - further information is needed.

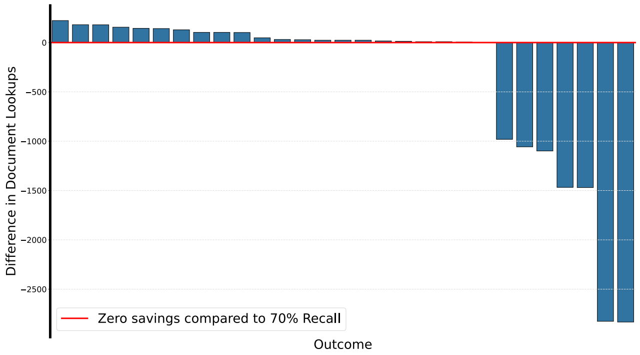

Figure 5: Relative savings of the Confidence Boundary approach on CLEF 2018 dataset for sensitivity. Note that the savings metric includes relevant and non-relevant documents, with the black horizontal line indicating savings made with 70% recall method.

(i.e., reviewed documents) to stabilise and provide reliable estimates of distribution shape.

This finding suggests that higher moments may not be ideal for datasets with few relevant documents. Such datasets are common in systematic reviews of niche topics or rare diseases. In such scenarios, the instability of higher moment estimates due to limited data can lead to delayed or even absent stopping signals, undermining the efficiency of the review process.

It is crucial to acknowledge the limitations of moment-based stopping methods, especially in small sample sizes and skewed data distributions, which are common challenges in systematic reviews.

Small sample sizes are common without systematic reviews, particularly for rare diseases or conditions. With few data points, moment estimates become highly unstable and sensitive to individual studies. A single study with an unusually high or low point estimate (e.g., of sensitivity or false positive rate) can disproportionately influence the calculated moments. This can lead to erratic fluctuations in the stopping methods, causing premature or delayed stopping.

Several stopping methods, such as the Confidence Boundary method, 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 often fails with small sample sizes. When normality is violated, the calculated confidence intervals can be inaccurate, leading to unreliable stopping decisions. The algorithm might stop too early or too late based on a flawed understanding of the uncertainty surrounding the point estimate. This could be mitigated through implementing a minimum number of studies to be reviewed before stopping.

Even a single outlier study can have a disproportionately large impact on moment estimates in small samples. This can further exacerbate the instability of the stopping criteria and lead to erroneous stopping decisions. An outlier might artificially inflate or deflate the mean, skew the distribution, or distort the kurtosis, misleading the stopping algorithm.

While small sample sizes can pose challenges for moment-based stopping criteria, it is important to recognise that the systematic review process protects against this issue. A core component of systematic reviews is the rigorous assessment of evidence quality, which involves critically appraising each included study for potential biases and limitations. This quality assessment often leads to the down-weighting or exclusion of studies deemed to be of low quality. Consequently, the remaining studies contributing to the synthesis tend to be more homogeneous and of higher quality. This effective filtering mechanism reduces the likelihood of highly divergent or unreliable data points skewing the overall synthesis, even with a limited number of included studies. In essence, the quality control inherent in systematic reviews can result in a narrower, more reliable distribution of evidence, making moment-based stopping methods more robust even with limited sample sizes.

8 Conclusion

This paper has presented a series of utility-based stopping methods for technology-assisted review. These methods focus on the utility of retrieved information rather than pursuing predefined recall thresholds. Experimental results with systematic reviews of DTA reviews suggest that lower moment-based stopping approach can reduce relevant document screening requirements significantly while preserving decision agreement. By modelling statistical moments such as the mean, variance, skewness, and kurtosis of key clinical metrics, these approaches adapt to the usefulness of the evidence, supporting more efficient screening processes. Overall, the results emphasise the benefits of shifting from recall-oriented stopping methods to user-centric, utility-based approaches for stopping decisions, with potential applications in systematic reviews and beyond. Future research should focus on refining stopping criteria by incorporating information redundancy, such as analysing how new documents alter point estimates. Adaptive decision thresholds, adjusted dynamically based on accumulating evidence, including quality and prior knowledge, should be explored.

References

[1] D. G. Altman and J. M. Bland. 1994. Diagnostic tests. 1: Sensitivity and specificity. *BMJ (Clinical research ed.)* 308, 6943 (1994), 1552. https://doi.org/10.1136/bmj.308.6943.1552

[2] 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. https://doi.org/10.1145/2009916.2009923

[3] 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. https://doi.org/10.1145/2484028.2484049

[4] Kevin P. Balanda and H. L. MacGillivray. 1988. Kurtosis: A Critical Review. *The American Statistician* 42, 2 (1988), 111–119. http://www.jstor.org/stable/2684482

[5] María J. Blanca, Jaume Arnau, Dolores López-Montiel, Roser Bono, and Rebecca Bendayan. 2013. Skewness and Kurtosis in Real Data Samples. *Methodology* 9, 2 (2013), 78–84. https://doi.org/10.1027/1614-2241/a000057

[6] G.J. Browne, M.G. Pitts, and J.C. Wetherbe. 2005. Stopping Rule Use During Web-Based Search. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. 271b–271b. https://doi.org/10.1109/HICSS.2005.556

[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. https://doi.org/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 (1973), 413–424. https://doi.org/10.1002/asi.4630240603

[9] William S. Cooper. 1976. The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness. *Information Processing and Management* 12, 6 (1976), 367–375. https://doi.org/10.1016/0306-4573(76)90034-0

[10] Gordon Cormack and Maura Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. *arXiv preprint arXiv:1504.06868* (apr 2015). arXiv:1504.06868

[11] 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. https://doi.org/10.1145/2911451.2911510

[12] 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. https://doi.org/10.1145/2983323.2983776

[13] 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), Gaithersburg, MD, USA, 1–9.

[14] Lawrence T. DeCarlo. 1997. On the meaning and use of kurtosis. *Psychological Methods* 2, 3 (1997), 292–307. https://doi.org/10.1037/1082-989X.2.3.292

[15] 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. https://doi.org/10.1007/978-3-319-76941-7_61

[16] Jacqueline Dinnes, Pawana Sharma, Sarah Berhane, Susanna S Van Wyk, Nicholas Nyaaba, Julie Domen, Melissa Taylor, Jane Cunningham, Clare Davenport, Sabine Dittrich, Devy Emperador, Lotty Hooft, Mariska Mg Leeflang, Matthew Df McInnes, René Spijker, Jan Y Verbakel, Yemisi Takwoingi, Sian Taylor-Phillips, Ann Van Den Bruel, Jonathan J Deeks, and Cochrane COVID-19 Diagnostic Test Accuracy Group. 2022. Rapid, point-of-care antigen tests for diagnosis of SARS-CoV-2 infection. *Cochrane Database of Systematic Reviews* 2022, 7 (2022). https://doi.org/10.1002/14651858.CD013705.pub3

[17] 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. https://doi.org/10.1145/1571941.1572145

[18] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[19] 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.

[20] 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.

[21] Noah Hollmann and Carsten Eickhoff. 2017. Ranking and Feedback-based Stopping for Recall-Centric Document Retrieval. In *CLEF (working notes)*. 7–8.

[22] 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 (2024), 796–808. https://doi.org/10.1177/09610006231157091

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

[24] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2018. CLEF 2018 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings* 2125. https://strathprints.strath.ac.uk/66446/

[25] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings* 2380. https://strathprints.strath.ac.uk/71253/

[26] Donald H Kraft and T Lee. 1979. Stopping rules and their effect on expected search length. *Information Processing and Management* 15, 1 (1979), 47–58. https://doi.org/10.1016/0306-4573(79)90007-4

[27] 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. https://doi.org/10.1145/3578337.3605135

[28] 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. https://doi.org/10.1145/3459637.3482415

[29] 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. https://doi.org/10.5555/1005332.1005345

[30] 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. https://doi.org/10.1145/3411755

[31] David Maxwell. 2021. *Modelling search and stopping in interactive information retrieval*. Ph. D. Dissertation. New York, NY, USA. https://doi.org/10.1145/3458537.3458543

[32] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. An Initial Investigation into Fixed and Adaptive Stopping Strategies. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 903–906. https://doi.org/10.1145/2766462.2767802

[33] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 313–322. https://doi.org/10.1145/2806416.2806476

[34] 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. https://doi.org/10.1145/3417334

[35] Kathryn Ritgerod Nickles. 1995. *Judgment-based and reasoning-based stopping rules in decision-making under uncertainty*. University of Minnesota.

[36] 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. AMIA Symposium* 2018 (2018), 817–826.

[37] Christopher R Norman, Mariska MG Leeflang, Raphaël Porcher, and Aurelie Neveol. 2019. Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *Systematic reviews* 8 (2019), 1–18.

[38] 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. https://doi.org/10.1145/3268928

[39] 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. https://doi.org/10.1016/j.accinf.2016.04.003

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

[41] 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. https://doi.org/10.1016/j.jclinepi.2005.02.022

[42] 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.

[43] Tefko Saracevic. 2022. *The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?* Springer Nature.

[44] 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. https://doi.org/10.1002/(SICI)1097-4571(198805)39:3<161::AID-ASI2>3.0.CO;2-0

[45] 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. https://doi.org/10.1002/jrsm.1093

[46] 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. https://doi.org/10.1145/3631990

[47] Dennis D. Wackerly, William Mendenhall III, and Richard L. Scheaffer. 2002. *Mathematical Statistics with Applications* (sixth edition ed.). Duxbury Advanced Series.

[48] 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. https://doi.org/10.1002/meet.2014.14505101030

[49] 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. https://doi.org/10.1145/3469096.3469872

[50] 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. https://doi.org/10.1002/asi.20092