Null hypothesis:

METHODOLOGY

Statistical Stopping Criteria for Active Learning in Systematic Review Screening

Max W Callaghan[??,??*] and Finn Müller-Hansen[??]

[*]Correspondence:
callaghan@mcc-berlin.net
[??]Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Active learning for systematic review screening promises to reduce the human effort required to identify relevant documents for a systematic review. Machines and humans work together, with the human providing training data, and the machine optimising the documents a human screens, so that they can identify all relevant documents after viewing only a fraction of the total documents. However, current approaches lack robust stopping criteria, so that reviewers do not know when they have seen all or a certain proportion of relevant documents. This means that such systems are hard to implement in live reviews. This paper introduces a workflow for working with robust and flexible statistical stopping criteria, that offer real work reductions on the basis of a given confidence level of reaching a given recall. The stopping criteria is shown on test datasets to achieve a reliable level of recall, while still providing consistent work reductions, while other methods proposed are shown to provide inconsistent recall and work reductions across datasets.

Keywords: Systematic Review; Machine Learning; Active Learning; Stopping Criteria

Background

Machine learning for evidence synthesis is a growing field, where machine learning enabled interventions are introduced at various points in the workflow, in order to reduce the human effort required to produce systematic reviews and other forms of evidence synthesis. A major strand of the literature works on screening - the identification of relevant documents in a set of documents whose relevance is uncertain [**?**]. For this task, several papers have developed or evaluated active learning.

Active learning is an iterative process where documents screened by humans are used to train a machine learning model to predict the relevance of unseen papers [**?**]. The most relevant studies are passed back to the human and rated, generating more labels to feed back to the machine. By prioritising those studies most likely to be relevant, a human reviewer most often identifies all relevant studies - or a given proportion of relevant studies (recall) - before having seen all the documents in the corpus. The proportion of documents not yet seen by the human when they reach the given recall threshold are referred to as the work saved.

In live review settings, however, the recall remains unknown until all documents have been screened. In order for work to really be saved, reviewers have to stop screening while uncertain about the recall. This is particularly problematic in systematic reviews because low recall increases the risk of publication bias The lack

of appropriate stopping criteria has therefore been identified as a research gap [**?**], although some approaches have been suggested. These fall into the following categories

- **Sampling criteria:** Reviewers estimate the number of relevant documents by taking a random sample at the start of the process. They stop when this number, or a given proportion of it, has been reached [**?**]
- **Heuristics:** Reviewers stop when a given number of irrelevant articles are seen in a row [**?**].
- **Pragmatic criteria:** Reviewers stop when they run out of time [**?**].

We show in this paper how the first two criteria are inadequate. We argue that the inadequacy lies in the unreliability - particularly across different domains, or datasets with different properties - both of the work saved and the recall achieved. Not achieving the desired level of recall is a particular challenge for systematic review screening. In systematic reviews, the larger the proportion of the relevant literature is not considered, the greater the risk of publication bias . Without the reliable or reportable achievement of a desired level of recall, AL systems in live reviews remain challenging.

ref on cross-domain reliability

f cochrane handbook

This study proposes a system for estimating the recall based on random sampling of remaining documents. We draw on the literature on binomial distributions [**?**], to illustrate how AL users can predefine a threshold in terms of uncertainty and recall, and use this to transparently save work with machine learning, while making a statement like "There is a ¡5% chance that we achieve a recall under 95%".

The information retrieval literature discusses similar stopping criteria for ranking algorithms like BM25 and variants [**?**, **?**]. However, the estimators they use to determine the recall rely on the specific ranking functions and depend on their search input. Therefore, the quality of the estimation depends on the adequacy of the model. Our approach, on the contrary, is independent of model choice or model performance.

We evaluate this stopping criteria on real-world systematic review datasets on which active learning systems have previously been tested.

Methods

Existing Stopping Criteria for Active Learning

We start by explaining the sampling and heuristic based stopping criteria before showing with toy data how the criteria fall short. Then we introduce our own suggested stopping criteria, and show its benefits with toy data, before testing all criteria on real world datasets.

Sampling Based Stopping Criteria

The stopping criterion suggested by [**?**] involves establishing the Baseline Inclusion Rate (BIR), by taking a random sample at the beginning of screening. This is used to estimate the number of relevant documents in the whole dataset. Reviewers continue to screen until this number, or a proportion of it corresponding to the desired level of recall, is reached.

However, the estimation of the BIR fails to take into account the sampling uncertainty. Figure 1.a shows for 50 random samples of 2,000 documents from 20,000

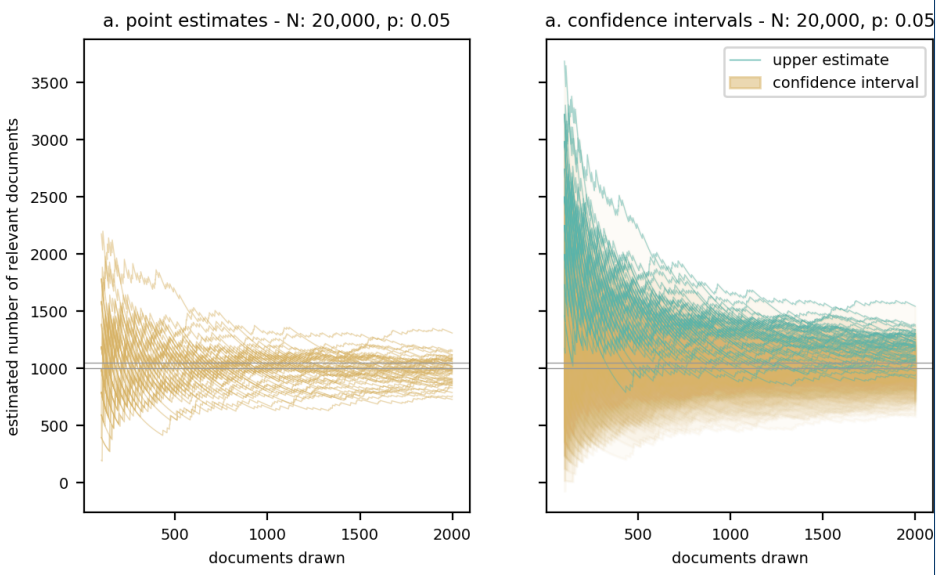

Figure 1 Uncertainty in estimating the baseline inclusion rate. a. shows the estimated BIR after drawing each document, for 50 random samples of from 20,000 documents, after each document is drawn. b. Shows the 95% Confidence interval (brown) and the upper bound (green) for each of the same samples. The two lines correspond to 100% of relevant documents, and $\frac{100}{0.95}$% of relevant documents.

documents (where 5% of documents are relevant), the predicted number of documents after each document drawn. Depending on the luck of the draw, one might estimate the true number of documents as being

- much higher than the true value, meaning that the stopping criterion would never be reached - and no work could be saved
- much lower than the true value, meaning that the stopping criterion would be reached before the desired level of recall was actually achieved

We can estimate the uncertainty of a binomial outcome using an Agresti-Coull confidence interval (which has been shown to be conservative at levels near 0 [**?**]). Figure 1.b shows the 95% confidence interval (brown) and its upper bound (green) in the same set of random draws. Where the green line is above the above the lower (upper) grey line, identification of 100% (95%) of the estimated number of studies would be impossible, and no work would be saved.

this is less relevant here, go back to literature and see if another is better

Figure 2 explores different scenarios of true relevance and probability thresholds, showing that the higher the true proportion of relevant documents, the fewer documents one would have to sample to make the identification of a given fraction (the true number of relevant results divided by the upper estimate) possible. Even with a relatively high proportion of relevant documents, and a sample of 500, all relevant documents would have to be identified to have a 95% likelihood of achieving 95% recall. With low proportions of relevant documents, only an 80% recall can be estimated with confidence, even after a sample of 2000, and subsequently identifying all relevant documents. N.B, these results ignore the random element of sample order, which explains the variation on figure 1. They show upper confidence bounds for accurate estimates of relevance.

Why does it behave like this, is there a theoretical maximum for each? something to do with absolute values? a Bug? Is uncertainty a function of being near to 0?

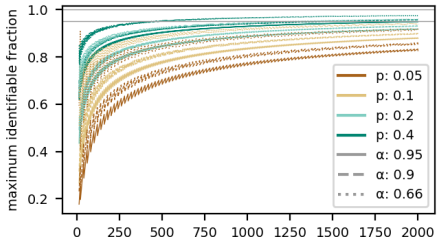

Figure 2 Maximum indentifiable fraction of relevant results under different levels of true relevance p and confidence thresholds α. Maximum identifiable fraction is the true number of relevant documents divided by the upper estimate.

Heuristic Based Stopping Criteria

Some studies give the example of heuristic based stopping criteria based on drawing a given number of irrelevant articles in a row [**?**] [+REFS]. We take this as a proxy for estimating that the proportion of documents remaining in the unseen documents is low. We find this a promising intuition, but argue that 1) it ignores uncertainty, as discussed in relation to the previous method; and 2) it misunderstands the significance of a low proportion of relevant documents in estimating the recall.

Figure 3 illustrates this second point. We show two scenarios with identical low proportions of relevant documents (note that estimating a proportion lower than 0.02 with 95% confidence requires 160 consecutive irrelevant documents). Recall is not simply a function of the relevance of unseen documents, but also of the number of unseen documents. This also means that where machine learning has performed well (as in the top figure), low proportions of irrelevant documents in those that remain are indicative of lower recall than where ML has performed less well.

monstrate?

Other stopping criteria

[**?**] Develop a "simple, operational stopping criterion": stopping after half the documents have been screened. Although the criterion worked in their experiment, it is unclear how this could be generalised, and its development depended on knowledge of the true relevance values. [**?**] note that "the reviewer can elect to end the process of classifying documents at any point, recognizing that stopping before reviewing all documents involves a trade-off of lower recall for reduced workload", although clearly the reviewer lacks information about probable recall. [**?**] adopt a more complicated stopping criterion which allows the user to target a specific level of recall. However, reviewers are not given the opportunity to specify a confidence level, and for two of the four datasets in which they tested their criteria, the median achieved recall at a stopping criteria targeting 95% recall was below 95%. [**?**] also present an innovative stopping criteria, but it does not take into account uncertainty, and produces results *near* a target recall threshold. These last examples are promising developments, but [are not ready for live reviews because hard to understand/communicate?? do not give uncertainty about reaching a threshold]

ow does it actually
ork??

A Statistical Stopping Criterion for Active Learning

Random Sampling

Our aim is to formulate a stopping criterion such that a recall of τ has been reached with confidence level α. To achieve this, we formulate a hypothesis test: We first calculate the number $\hat{n}$ of relevant documents that can maximally remain in the unseen documents to achieve a given recall target. The recall is defined as the ratio of seen relevant documents over all relevant documents

$$\tau = \frac{\hat{\rho}}{\rho + \tilde{n}}. \tag{1}$$

Reorganization of this equation gives

$$\tilde{n} = \frac{\hat{\rho}}{\tau} - \rho. \tag{2}$$

todo: argue why to take the ceil

$$\hat{n} = \lceil \frac{\rho}{\tau} - \hat{\rho} \rceil \tag{3}$$

The null-hypothesis to reject is then: There are $\hat{n}$ or more relevant documents remaining in the set of unseen documents.

In theory, we draw a random sample of N documents, evaluate how many of them are relevant, and denote this number with k. We can estimate the probability that we observe k or fewer relevant documents in a random draw of N documents without replacement from a pool of M documents, of which n are relevant using the cumulative distribution function of the hypergeometric distribution. This function is defined as the sum over the probability mass function of the hypergeometric distribution $p(x, M, n, N)$:

$$cdf(k, M, n, N) = \sum_{x \leq k} p(x, M, n, N) = \sum_{x \leq k} \frac{\binom{n}{x}\binom{M-n}{N-x}}{\binom{M}{N}}, \tag{4}$$

where $\binom{a}{b}$ denote binomial coefficients. In our case, M is the number of unseen documents before we start random sampling.

We can reject Hypothesis 1 if the p-value $p = cdf(k, M, \hat{n}, N) < 1 - \alpha$ [references to hypergeometric testing?, needs further argumentation?]. Because we tested for the lower tail, this implies that we can be confident with a level α that the number of unseen relevant documents is lover than $\hat{n}$.

In practice, we can draw random unseen documents repeatedly. After each draw, we calculate the p-value using the number of relevant documents k and the number randomly sampled of documents N. We can stop the screening once a p-value $p < 1 - \alpha$ is reached.

Pseudo-random sampling

In order to decide when to begin a random sample, we employ pseudo-random sampling, where we treat previously screened documents as a random sample. The distribution of relevant documents among previously screened documents is clearly

Parameter	Description	Estimation
P	Total number of studies	*Observed*
S	Number of studies coded by humans	*Observed*
U	Number of studies not yet coded by humans	*Observed* ($P - S$)
α	Acceptable uncertainty level	*Given*
τ	Target recall threshold	*Given*
k	Number of relevant documents drawn	
N	Number of randomly drawn documents	*Observed*
M	Number of remaining unseen documents before sampling	
n	Number of relevant documents in sample	
$\hat{n}$	Minimum number of relevant documents in sample that would mean the recall threshold had been reached	

Table 1 Parameters

not random, as documents predicted to be relevant are prioritised. It is reasonable to assume, though, that the density of relevant documents is greater among previously screened documents than among remaining unseen documents. This would make the following estimates conservative.

After reviewing each document, S documents are screened, and U documents are yet to be seen. We treat $i = 1 \dots S$ of the previously screened documents as a random sample, and calculate p as before for each sample, taking the minimum across all samples p_{min}. If p_{min} is less than $1 - \frac{\alpha}{2}$, we switch to random sampling. We also calculate p_{min} for the remaining values as if we had not switched to random sampling and present these in the results below.

Old description

Figure 4 shows a workflow for the approach proposed in this paper. The random sampling occurs at the end of the process, and is used to estimate the number of relevant documents remaining and the total number of relevant documents. The rationale is to limit the uncertainty to a subsection of the dataset: that which has not yet been screened. As reviewers continue to draw random documents, the uncertainty range decreases, and the proportion of the data about which one is uncertain also decreases.

Table 1 shows the parameters, known, estimated, and given, available during the random sampling process and required to estimate recall. Expanding on what was stated before, recall - at a given confidence threshold is a function of 1) the upper estimate of the relevance of remaining documents, 2) The estimated relevance of all documents in the dataset, 3) The proportion of documents not yet seen.

Figure 5 shows the minimum estimated recall for a set of confidence intervals, along with the actual recall (in grey) for a case where 800 out of 2,000 documents have been reviewed, 2% of remaining documents are relevant, and 20% of all documents are relevant. We see that after 95% recall has actually been achieved, but before 100% of documents have been been seen, we can be confident at each of the given confidence levels that 95% of relevant documents have been identified. All estimates of recall only reach 100% after all documents have been seen, as it is not possible to exclude the possibility, at any given confidence level, that the proportion of relevant documents is greater than 0 . Figure 6 shows the various trajectories of a 95% minimum recall (blue) and actual recall (grey) given the same paramaters for 200 random samples.

ctually maybe this is
ssible if we work with
hole numbers? How
es that change things?

That the blue lines consistently meet a 95% threshold to the right of the grey lines indicates that reviewers have to see more than 95% of relevant documents in order to be confident that they have achieved their threshold. We call the number of documents it is necessary to review to establish with confidence that reviewers have achieved 95% recall after they have already achieved 95% recall the *additional burden*.

In figure 7, we investigate how the total number of documents and the proportion of relevant documents affect the additional burden . We demonstrate the additional burden decreases with increasing total numbers of documents. With decreasing relevance of the remaining documents (holding the recall constant at 95%, this means that fewer documents remain unseen), the additional burden increases. This implies that where the machine learning has more successfully identified documents and reduced burden, the additional burden required to establish the level of recall is higher. This is because the proportion of documents not seen is greater, so uncertainty around the relevance of that proportion has a greater impact on the estimation of recall.

Use the number of documents seen instead of p?

how? Non-linearly?

When to move to random sampling remains an open question. Our current approach is to use the predicted relevance of the machine learning model as a guide. We calculate the [minimum/maximum] estimated recall based on observing the predicted relevance level in [x] random documents. If the estimated recall would be below the threshold, then we begin drawing random documents.

The stopping criteria explained here is

Evaluation

We evaluate each of the three criteria discussed on real world test data. With each method, we run 100 runs and record

- **Actual Recall**: The recall when the stopping criteria is met
- **WS-SC**: Work saved when the stopping criteria is met [1]
- **Additional Burden**: The proportion of the literature that need to be screened in order to *be confident* that 95% recall has been achieved, after 95% of documents have actually been identified.
- **Missed Recall**: The amount by which, if any, the recall target was missed - in percentage points

We operationalise the heuristic stopping criteria with 50, 100, and 200 irrelevant records in a row. For simplicity, we use a basic SVM model [REFS], with 1-2 word n-grams in the document abstracts used as input data. We start with random samples of [100, 200, and 500] documents.

As the intention is to evaluate stopping criteria, rather than machine learning approaches, we focus on additional burden and missed recall. These parameters correspond to the tradeoff inherent in stopping criteria: stopping too early means missing the recall target, stopping too late means undergoing additional burden than was actually required to meet the target.

We use the collection of systematic reviews used to develop machine learning applications for document screening by Aaron Cohen and co-authors in 2006 [**?**], along with PB COPD, and FASTREAD . We feel that testing on datasets with different

fill in gaps

[1] we do not use work saved over sampling, is it is not axiomatic that seeing 95% of a random sample would achieve 95% recall

properties and from different domains is key to establishing criteria appropriate for general use. In this way we can show how well the criteria perform even when the model performs badly.

Results

Discussion

Conclusion

We argue that the stopping criteria proposed here has several important advantages.

- The accuracy of results is independent of modelling decisions. When a model fits badly, you have to work more, but you don't miss results without knowing it.
- A minimum threshold with uncertainty is understandable and communicable in a live review. It is based solely on arithmetic and simple statistics.
- A minimum threshold better takes into account the undesirability of missed recall.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
Max Callaghan is supported by a PhD scholarship from the Heinrich Böll Foundation

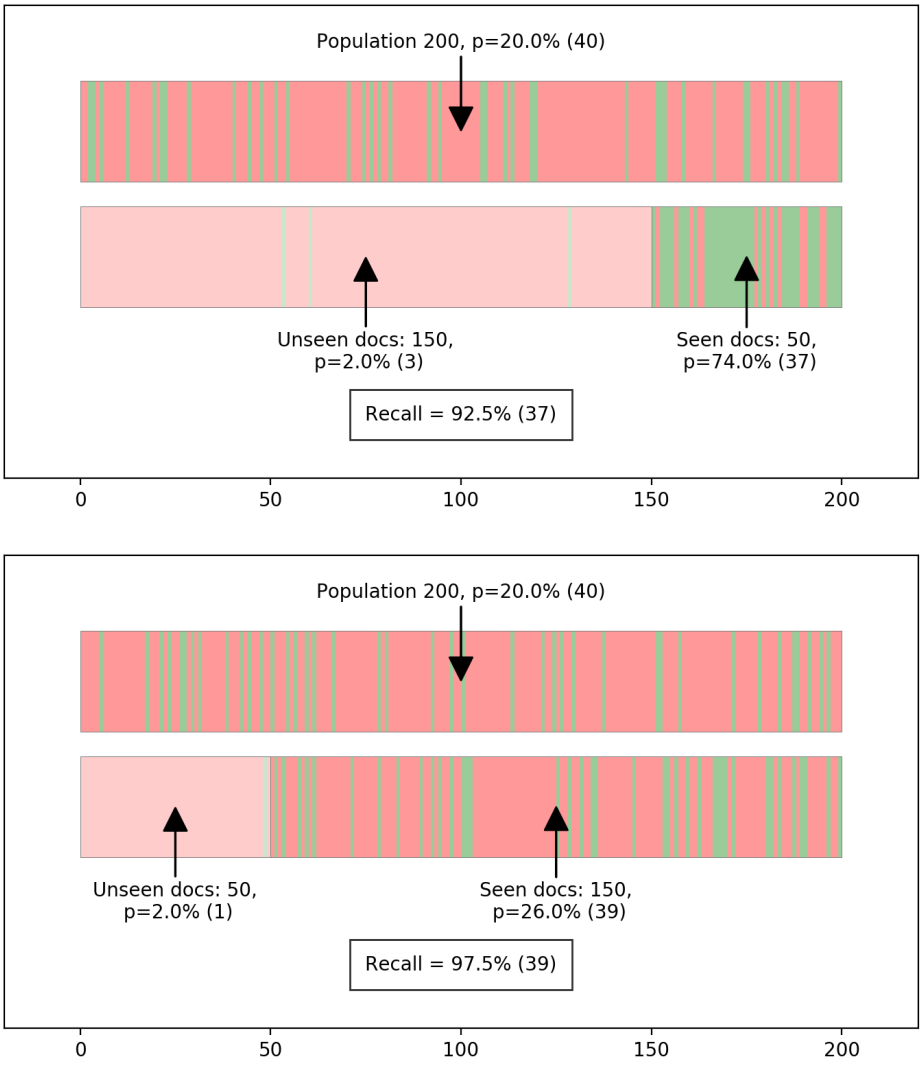

Figure 3 Similar low proportions of relevant documents in unseen documents with different consequences for recall. The top bar shows a random distribution of relevant documents (green) and irrelevant documents (red) at a given proportion of relevance. The bottom bar shows distributions of relevant and irrelevant documents in hypothetical sets of seen (right) and unseen (left - transparent) documents.

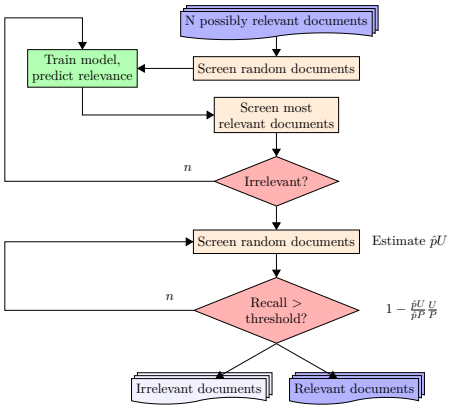

Figure 4 A workflow for active learning in screening with a statistical stopping criterion

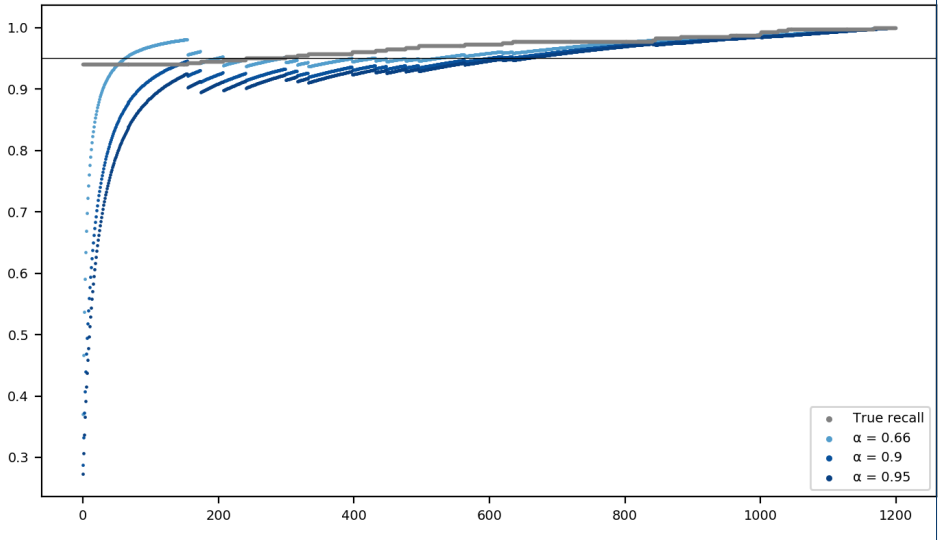

Figure 5 Sample recall

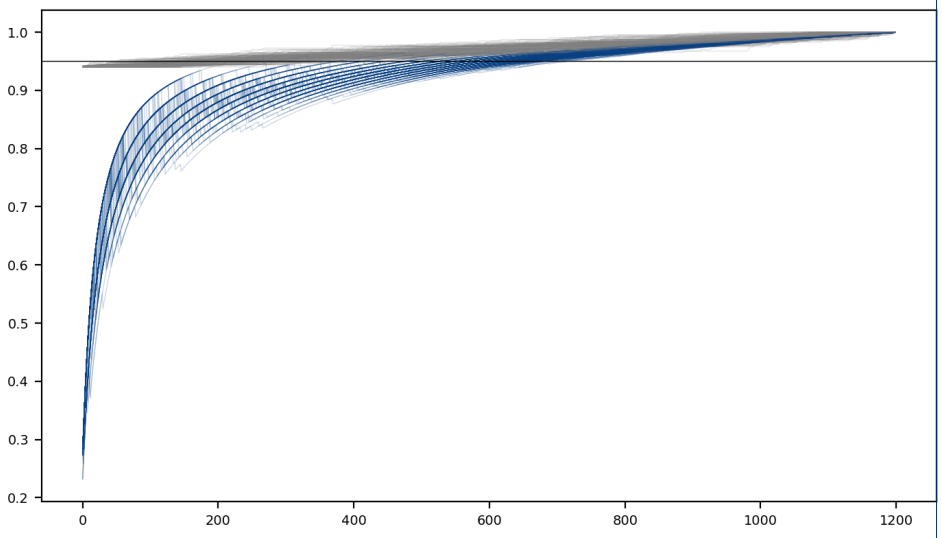

Figure 6 Sample recall variation

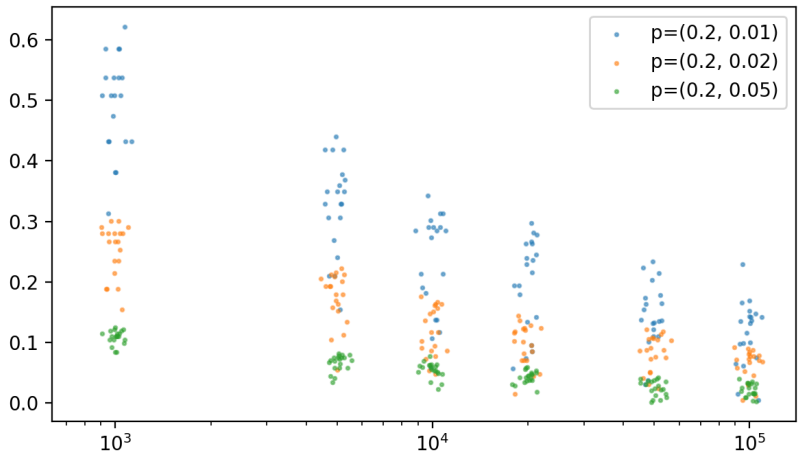

Figure 7 Additional burden for 20 random samples where the 95% threshold has been reached, and the remaining documents are 1%, 2% or 5% relevant, for different total numbers of documents.

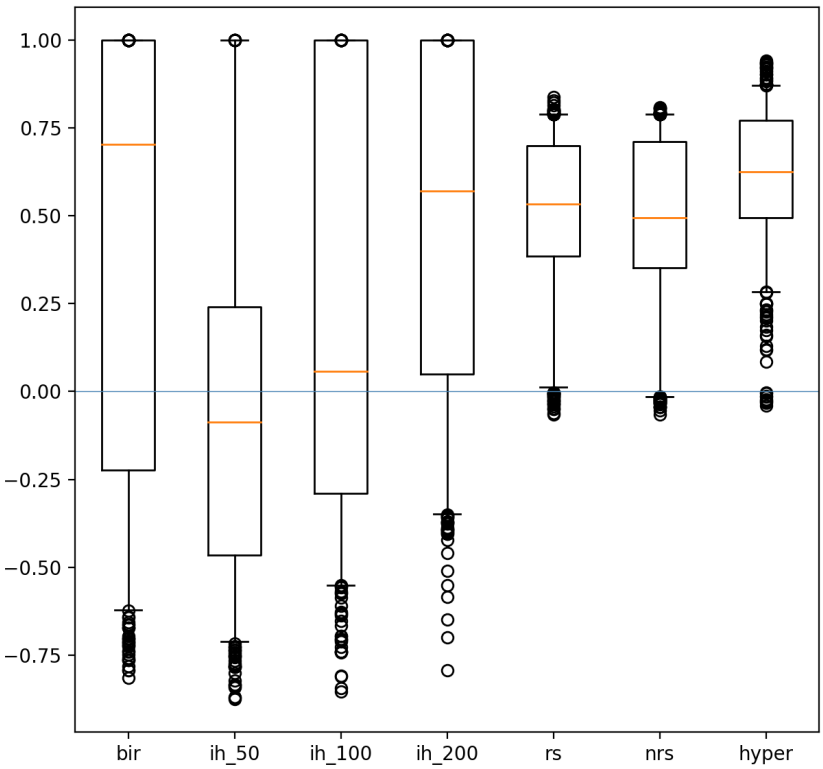

Figure 8 Additional burden ($y > 0$) and missed recall ($y < 0$) for the range of stopping criteria tested.

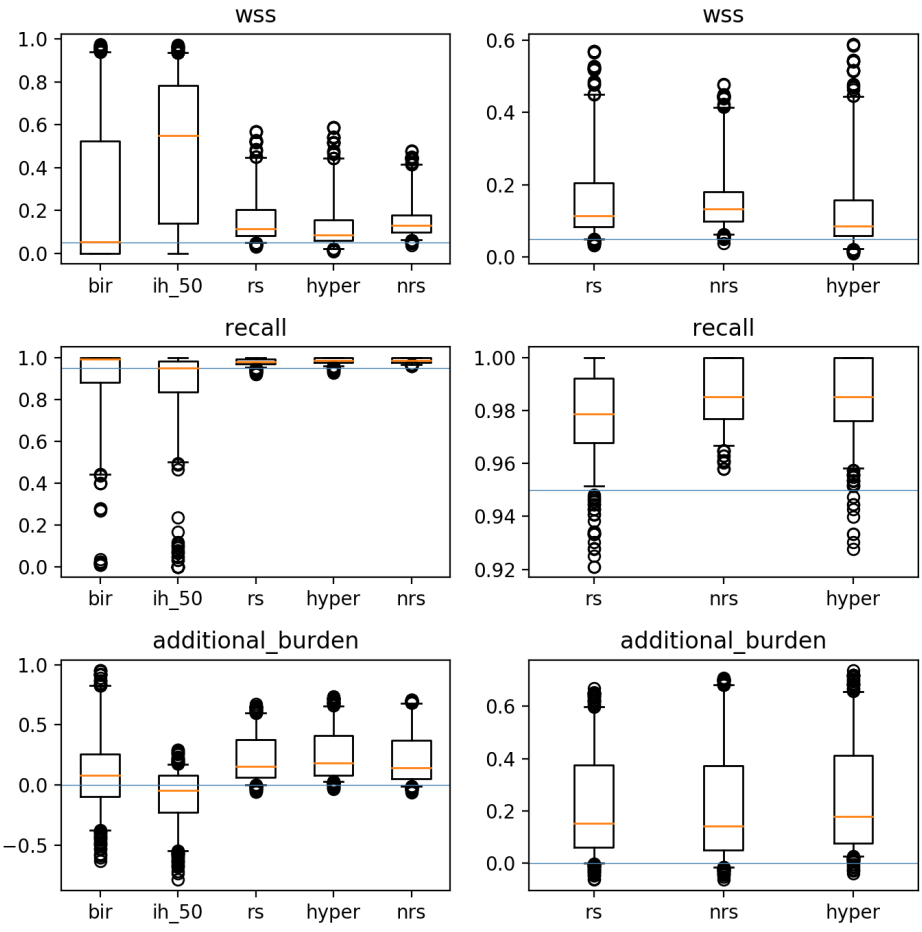

Figure 9 Work saved, recall and additional burden across a variety of stopping criteria

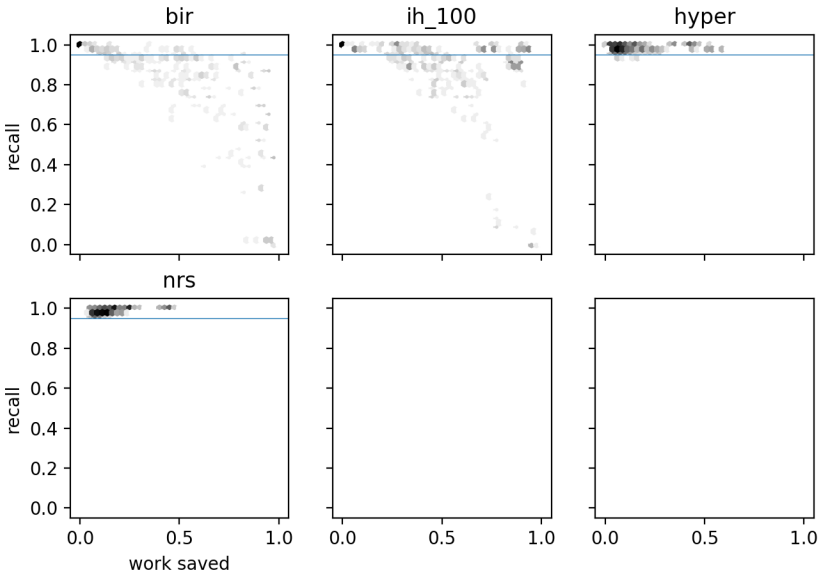

Figure 10 Work saved, recall across a variety of stopping criteria

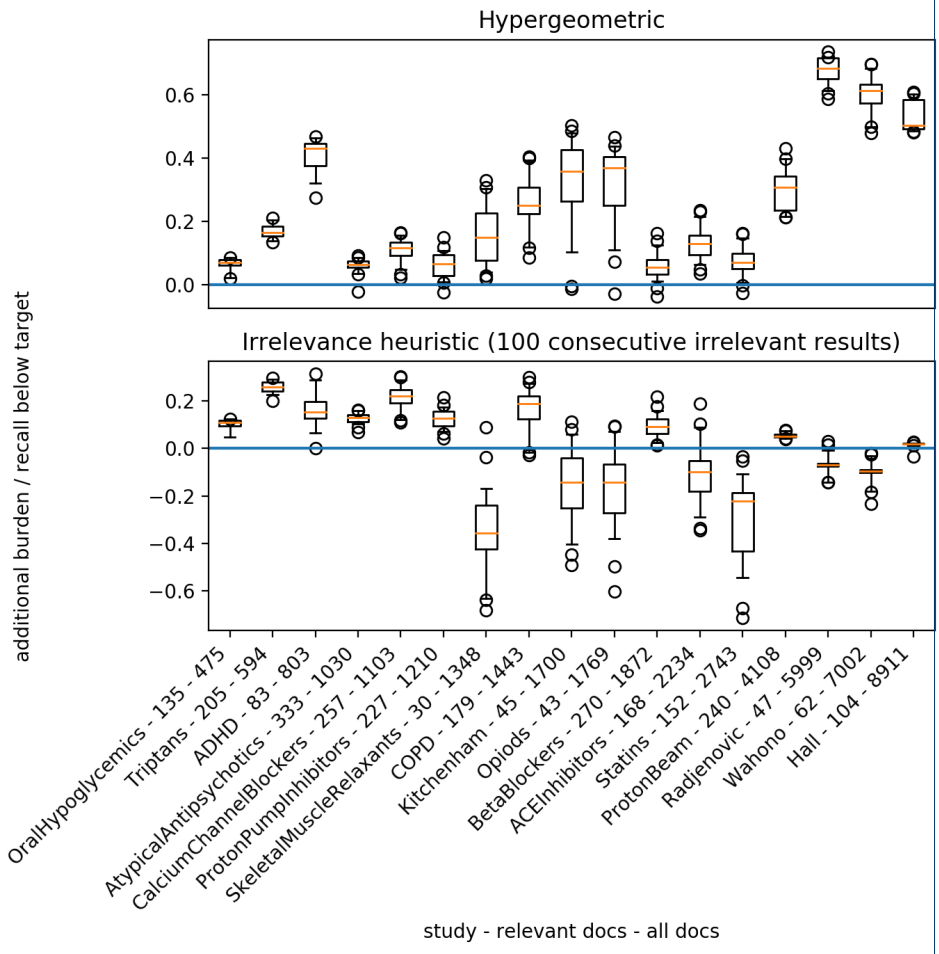

Figure 11 Work saved, recall across a variety of stopping criteria