METHODOLOGY

Statistical Stopping Criteria for Active Learning in Systematic Review Screening

Max W Callaghan[1,2*] and Finn Müller-Hansen[1]

*Correspondence: callaghan@mcc-berlin.net
[1]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 humans providing training data, and the machine optimising the documents that the humans screen. This enables the identification of 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

Evidence synthesis technology is a rapidly emerging field that promises to change the practice of evidence synthesis work [1]. Interventions have been proposed at various points 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 [2]. This is a time consuming and repetitive task, and in a research environment with constrained resources and increasing amounts of literature, this may limit the scope of the evidence synthesis projects undertaken. Several papers have developed Active Learning (AL) approaches [3, 4, 5, 6, 7] to reduce the time required to screen documents. This paper sets out how current approaches are unsuitable in practice, and outlines and evaluates a small modification that would make AL systems ready for live reviews.

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 [8]. The algorithm chooses which studies will next be screened by humans, often those which are likely to be relevant or about which the model is uncertain, in order to generate 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 is referred to as the work saved, as this represents the proportion of documents that they do not have to screen, which they would have had to without machine learning.

Machine learning applications are often evaluated using sets of documents from already completed systematic reviews for which inclusion or exclusion labels already exist. As all human labels are known *a priori*, it is possible to simulate the screening process, recording when a given recall target has been achieved. In live review settings, however, 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 recall. This is particularly problematic in systematic reviews because low recall increases the risk of bias [9]. The lack of appropriate stopping criteria has therefore been identified as a research gap [10], 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 [11]
- **Heuristics:** Reviewers stop when a given number of irrelevant articles are seen in a row [6, 7].
- **Pragmatic criteria:** Reviewers stop when they run out of time [3].

We show in this paper how existing criteria are inadequate. We demonstrate with several previously used datasets that their inadequacy lies in the unreliability - particularly across different domains, or datasets with different properties [2] - both of the work saved and the recall achieved. Without the reliable or reportable achievement of a desired level of recall, deployment of AL systems in live reviews remains challenging.

This study proposes a system for estimating the recall based on random sampling of remaining documents. We use a simple statistical method to iteratively test a null hypothesis that the recall achieved is less than a given target recall. This allows AL users to predefine a target in terms of uncertainty and recall, so that they can make transparent, easily communicable statements 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 [12, 13]. 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 our stopping criteria, along with heuristic and sampling based criteria on real-world systematic review datasets on which AL systems have previously been tested [14, 13, 15, 16].

Methods

We start by explaining the sampling and heuristic based stopping criteria and showing their methodological weaknesses. Then we introduce our own suggested stopping criteria before testing all criteria on real world datasets.

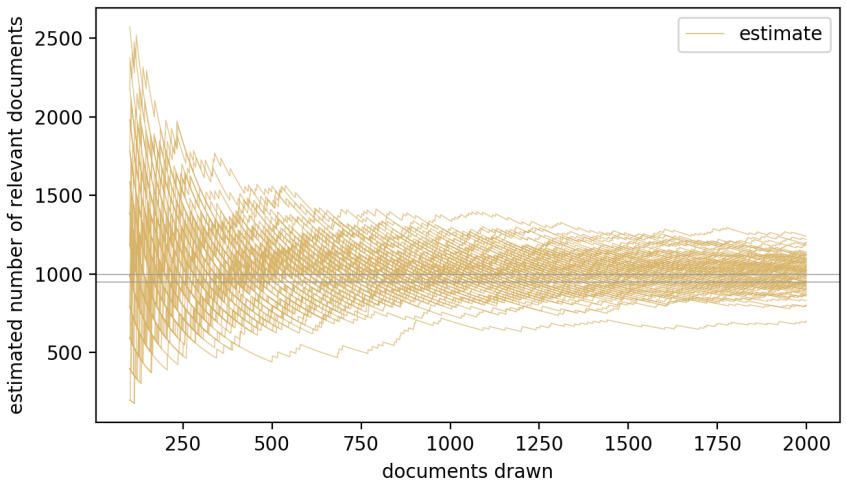


Figure 1: Uncertainty in estimating the baseline inclusion rate. For 50 random samples of 2,000 documents from 20,000 documents, we show the estimated number of relevant documents, after each document is drawn. The two horizontal grey lines correspond to the actual number of relevant documents, and to $\frac{100}{0.95}\%$ of relevant documents. This shows how even after taking a large sample, baseline estimation inevitably offers poor reliability, both in terms of recall and in work saved.

Existing Stopping Criteria for Active Learning

Sampling Based Stopping Criteria

The stopping criterion suggested by [11] 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 sampling uncertainty. Figure 1 shows the predicted number of documents after each document drawn for 50 random samples of 2,000 documents from 20,000 documents (where 5% of documents are relevant). Depending on the luck of the draw, one might estimate the true number of documents as being

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

Where the yellow line is above the above the upper grey line, identification of 95% of the estimated number of studies would be impossible, and no work would be saved. Where it is below, true recall, where recall is estimated to be 95% would be less than 95%.

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 [6, 7]. 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 2 illustrates this second point. We show two scenarios with identical low proportions of relevant documents observed in the unseen documents. In the top figure, machine learning (ML) has performed well, and 74% of documents seen were relevant. In the bottom figure, ML has performed less well, and only 26% of documents seen were relevant. In both cases, only 2% of unseen documents are relevant, but 2% of a larger number means more relevant documents are missed. 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 ML has performed well (as in the top figure), a low proportion of relevant documents in those not yet checked is indicative of lower recall than where ML has performed less well. Likewise, where the proportion of relevant documents in the whole corpus is low, a similarly low proportion of relevant documents is likely to be observed, even when true recall is low.

Other stopping criteria

[4] 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. [6] 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. [13] 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%. [12] also present an innovative stopping criteria, but it does not take into account uncertainty, and produces results *near* a target recall threshold, rather than above it in a reliable proportion of cases. These last examples are promising developments, but fail to take into account the needs of live systematic reviews, where the reliability of and ease of communication about recall are paramount.

A Statistical Stopping Criterion for Active Learning

Random Sampling

After using machine learning to select which documents are screened by humans as described above, we begin drawing a random sample from the remaining documents (when this happens is described below).

We use random sampling to estimate the probability that a target recall τ has been achieved. Because we are sampling a binary outcome without replacement, we

can use the hypergeometric distribution to formulate a statistical test. The hypergeometric distribution tells us the statistical significance of observing k successes (relevant documents) in n draws from a finite population of N documents with K successes.

$$k \sim Hypergeometric(N, K, n) \tag{1}$$

In our case, we know k, n and N after each draw, but K is unknown. We therefore substitute a hypothetical value $\hat{K}$: the minimum number of relevant documents in the sample had the recall target been missed.

Recall R is given by the number of relevant documents that have been seen $\dot{\rho}$ over the number of relevant documents in the whole dataset ρ

$$R = \frac{\dot{\rho}}{\rho} \tag{2}$$

The number of relevant documents in the whole dataset is composed of $\hat{\rho}$ relevant documents seen before random sampling began and $\tilde{K}$ relevant documents unseen at the start of random sampling. We can therefore express R as

$$R = \frac{\dot{\rho}}{\hat{\rho} + \tilde{K}}, \tag{3}$$

Substituting the target recall τ for R, reorganising and rounding up to the next integer, we can, after each draw, calculate $\hat{K}$, which is the minimum number of relevant documents that could have been remaining when random sampling started, if recall were lower than the target.

$$\hat{K} = \lceil \frac{\rho}{\tau} - \hat{\rho} \rceil \tag{4}$$

We use the cumulative distribution function to estimate the probability p of having observed k or fewer relevant documents in the sample given $\hat{K}$, or, in other words, the probability of observing as many relevant documents in our random sample as we did, if our recall target had not been achieved. If this is below our confidence level $1-\alpha$, we can reject the null hypothesis that the recall target was not achieved.

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 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.

	dataset	data_source	N	r_docs	p
0	UrinaryIncontinence	cohen	284	68	0.24
1	Antihistamines	cohen	287	90	0.31
2	Estrogens	cohen	349	79	0.23
3	NSAIDS	cohen	358	83	0.23
4	OralHypoglycemics	cohen	475	135	0.28
5	Triptans	cohen	594	205	0.35
6	ADHD	cohen	803	83	0.10
7	AtypicalAntipsychotics	cohen	1030	333	0.32
8	CalciumChannelBlockers	cohen	1103	257	0.23
9	ProtonPumpInhibitors	cohen	1210	227	0.19
10	SkeletalMuscleRelaxants	cohen	1348	30	0.02
11	COPD	copd_pb	1443	179	0.12
12	Kitchenham	fastread	1700	45	0.03
13	Opiods	cohen	1769	43	0.02
14	BetaBlockers	cohen	1872	270	0.14
15	ACEInhibitors	cohen	2234	168	0.08
16	Statins	cohen	2743	152	0.06
17	ProtonBeam	copd_pb	4108	240	0.06
18	Radjenovic	fastread	5999	47	0.01
19	Wahono	fastread	7002	62	0.01
20	Hall	fastread	8911	104	0.01

Table 1: Dataset properties

After reviewing each document, S documents have been 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, using the method above, 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 documents as if we had not switched to random sampling and record the recall and work saved when $p_{min} < 1 - \alpha$. We present these in the results below as the psuedo-random sampling criterion.

Analysis

Evaluation

We evaluate each of the criteria discussed, operationalising the heuristic stopping criteria with 50, 100, and 200 consecutive irrelevant records, on real world test data. We run 100 iterations on each dataset and record

- **Actual Recall**: The recall when the stopping criteria was met
- **WS-SC**: Work saved when the stopping criteria was met [1]
- **Additional Burden**: the work saved when the criterion was triggered subtracted from the work saved when the recall target was actually achieved.

For simplicity, we use a basic SVM model [17, 18], with 1-2 word n-grams taken from the document abstracts used as input data. We start with random samples of 200 documents. Subsequently, we "screen", that is, we reveal the labels of, batches of the 20 documents with the highest predicted relevance scores, retraining the model after each batch. Each criterion is evaluated after each document is "screened". For our criteria, we set the target recall value to 95% and the confidence level to 95%.

The systematic review datasets used for testing are described in table 1. We use the seminal collection of systematic reviews used to develop machine learning

[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

applications for document screening by Aaron Cohen and co-authors in 2006 [14], along with the widely used Proton Beam [15] and COPD [16] datasets, and computer science datasets used to test FASTREAD [13]. Testing on datasets with different properties and from different domains is key to establishing criteria appropriate for general use. Choosing as broad as possible data also prevents us from being able to "tune" our machine learning approach in ways that may work well for specific datasets but not generalise well. Work savings, even maximum work savings are therefore below the state of the art recorded for each of these datasets. In this way we can show how well the criteria perform even when the model performs badly.

Results

Figure 4 shows the actual recall and work savings achieved when each stopping criteria has been satisfied. For comparison, we also include the results that would have been achieved with *a priori* knowledge of the data, that is, the work saved when the 95% recall target was actually reached. In a live systematic review, reviewers would never know when this had been reached, but these are the work savings most often reported in machine learning for systematic review screening studies.

The baseline sampling criteria (Figure 4a) misses the 95% recall target in 39.67% of cases, while the most common work saving is 0%. This is in line with our expectations that, due to random sampling error, the expected number of documents will often be over-estimated or under-estimated, resulting in zero work savings or poor recall.

The Heuristic based stopping criteria, both for 50 consecutive irrelevant results (Figure 4b - IH50), and for 200 irrelevant results (Figure 4e) also perform unreliably. Although the mean work saved for IH50 is 41%, the target is missed in 39% of cases. The cases below the horizontal grey line indicate instances where work has been saved at the expense of achieving the recall target.

Both the random sampling and the pseudorandom sampling criteria achieve the target threshold of 95% in more than 95% of cases. In fact, the pseudorandom sampling criterion outperforms the random sampling criterion with respect to both recall and work savings. In theory, the pseudorandom sampling criteria is conservative if the assumption holds that documents chosen by machine learning are not less likely to be relevant than those chosen at random. Based on our experiments, this assumption seems reasonable, and accounts for the higher recall. Because the pseudorandom sampling criterion can flexibly choose its sample, whereas the random criterion has to wait for a random sample to be triggered, the criterion is also triggered earlier, as it can make use of more data. This accounts for the higher work savings.

In figure 5 we rescale the x axis, calling it additional burden, which is simply the work saved when the criterion is triggered subtracted from the work saved when the recall target was actually achieved. This measure indicates whether the stopping criterion was triggered too early (positive values), or too late (negative values). Here we directly highlight the tradeoffs involved in deciding when to stop screening.

To help explain the different work savings that were observed in our experiments, we show the distribution of work savings from our pseudorandom criterion for each

dataset in figure 6. In general, higher work savings are possible when the total number of documents is larger. However, in datasets with a low proportion of relevant documents, smaller work savings are possible.

Figure 7 shows the recall and the probability of the null hypothesis for the best performing iteration of four datasets. Although the 95% recall target is achieved very quickly in the Radjenovic dataset, the null hypothesis cannot be excluded until much later. This is because the dataset has only 47 relevant documents out of a population of 5,999. After the 95% recall target was achieved, 45 out of 47 relevant documents had been seen and 5,029 documents remained. The null hypothesis was therefore that 3 or more of these 5,029 documents were relevant, which requires a lot of evidence to disprove. The burden of proof was smaller in the case of the Proton Beam dataset, where the null hypothesis at the point that the 95% recall threshold was reached was that 13 out of 3,369 remaining documents were relevant.

The Statins and Triptans datasets show how the criterion performs when the machine learning model has performed poorly in predicting relevant results. In each case, 95% recall is achieved with close to 20% of documents remaining. With fewer documents remaining, it takes less time to rule out the possibility that the number of relevant documents left is greater than that which would be compatible with the achievement of the recall target.

Discussion

Our results show that it is possible to use machine learning to achieve a given level of recall with a given level of uncertainty. The tradeoff for achieving recall reliably is that the work saving achieved is less than the maximum possible work saving. However, for large datasets with a significant proportion of relevant documents, the additional effort required to satisfy the criterion will be smaller. This makes the approach well suited to broad topics with lots of literature. In other words, it is precisely where machine learning will be most useful that the additional effort will be smaller.

Different use cases for machine learning enhanced screening may also carry different requirements for recall, or different tolerances for uncertainty. These can be flexibly accommodated within our stopping criterion. Importantly, the ability to make probabilistic statements about the chance of achieving a given recall target makes it possible to clearly communicate the implications of using machine learning enhanced screening to readers and reviewers who are not machine learning specialists. This is extremely important in live systematic reviews.

Our criteria has the further advantage that it is independent of the modelling choice or performance. If a machine learning model performs badly at discerning relevant from irrelevant results, the only consequence will be that the work saved will be low. With other criteria this may result in poor recall. When using machine learning for screening, poor recall can result in biased results, while low work savings represent no loss to the reviewer as compared to not using machine learning.

So far, systematic review standards have no way of accommodating screening with machine learning. We hope that the reliability and clarity of reporting offered by our stopping criteria make them suitable for incorporation into standards, so that machine learning for systematic review screening can fulfil its promise of reducing workload and making more ambitious reviews tractable.

Conclusion

This paper demonstrates the unsuitability of existing stopping criteria for machine learning approaches to document screening, and proposes a simple method that delivers reliable recall, independent of machine learning approach or model performance. Our robust statistical stopping criteria allow users to easily communicate the implications of their use of machine learning, making machine learning enhanced screening ready for live reviews.

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

Author details
[1]Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany.
[2]Priestley International Centre for Climate, University of Leeds, Leeds , LS2 9JT Leeds, United Kingdom.

References

1. Westgate MJ, Haddaway NR, Cheng SH, McIntosh EJ, Marshall C, Lindenmayer DB. Software support for environmental evidence synthesis. Nature Ecology & Evolution. 2018;2:588–590.
2. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: A systematic review of current approaches. Systematic Reviews. 2015;4(1):1–22.
3. Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. Journal of Biomedical Informatics. 2014;51:242–253. Available from: http://dx.doi.org/10.1016/j.jbi.2014.06.005.
4. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010;11.
5. Wallace BC, Small K, Brodley CE, Trikalinos TA. Active learning for biomedical citation screening. 2010;(July):173.
6. Jonnalagadda S, Petitti D. A new iterative method to reduce workload in systematic review process. International Journal of Computational Biology and Drug Design. 2013;6(1/2):5.
7. Przybyła P, Brockmeier AJ, Kontonatsios G, Le Pogam MA, McNaught J, von Elm E, et al. Prioritising references for systematic reviews with RobotAnalyst: A user study. Research Synthesis Methods. 2018;9(3):470–488.
8. Settles B. Active Learning Literature Survey. University of Wisonsin-Madison; 2009.
9. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Cochrane Handbook for Systematic Reviews of Interventions. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. version 5. ed. The Cochrane Collaboration; 2011. Available from: www.handbook.cochrane.org.
10. Bannach-Brown A, Przybyła P, Thomas J, Rice ASC, Ananiadou S, Liao J, et al. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Systematic Reviews. 2019;8(1):1–12.
11. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, et al. Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews. Research Synthesis Methods. 2014;5(1):31–49.
12. Di Nunzio GM. A study of an automatic stopping strategy for technologically assisted medical reviews. In: Pasi G, Piwowarski B, Azzopardi L, Hanbury A, editors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 10772 LNCS. Cham: Springer International Publishing; 2018. p. 672–677.
13. Yu Z, Menzies T. FAST 2 : An intelligent assistant for finding relevant papers. Expert Systems with Applications. 2019;120:57–71. Available from: https://doi.org/10.1016/j.eswa.2018.11.021.
14. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association. 2006;13(2):206–219.
15. Terasawa T, Dvorak T, Ip S, Raman G, Lau J, Trikalinos T. Review Annals of Internal Medicine Systematic Review : Charged-Particle Radiation Therapy for Cancer. Annals of Internal Medicine. 2009;(5).
16. Castaldi PJ, Cho MH, Cohn M, Langerman F, Moran S, Tarragona N, et al. The COPD genetic association compendium: A comprehensive online database of COPD genetic associations. Human Molecular Genetics. 2009;19(3):526–534.
17. Cortes C, Vapnik V. Support-Vector Networks. In: Machine Learning; 1995. p. 273–297.
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python Fabian. Journal of Machine Learning Research. 2011;12:2825–2830.

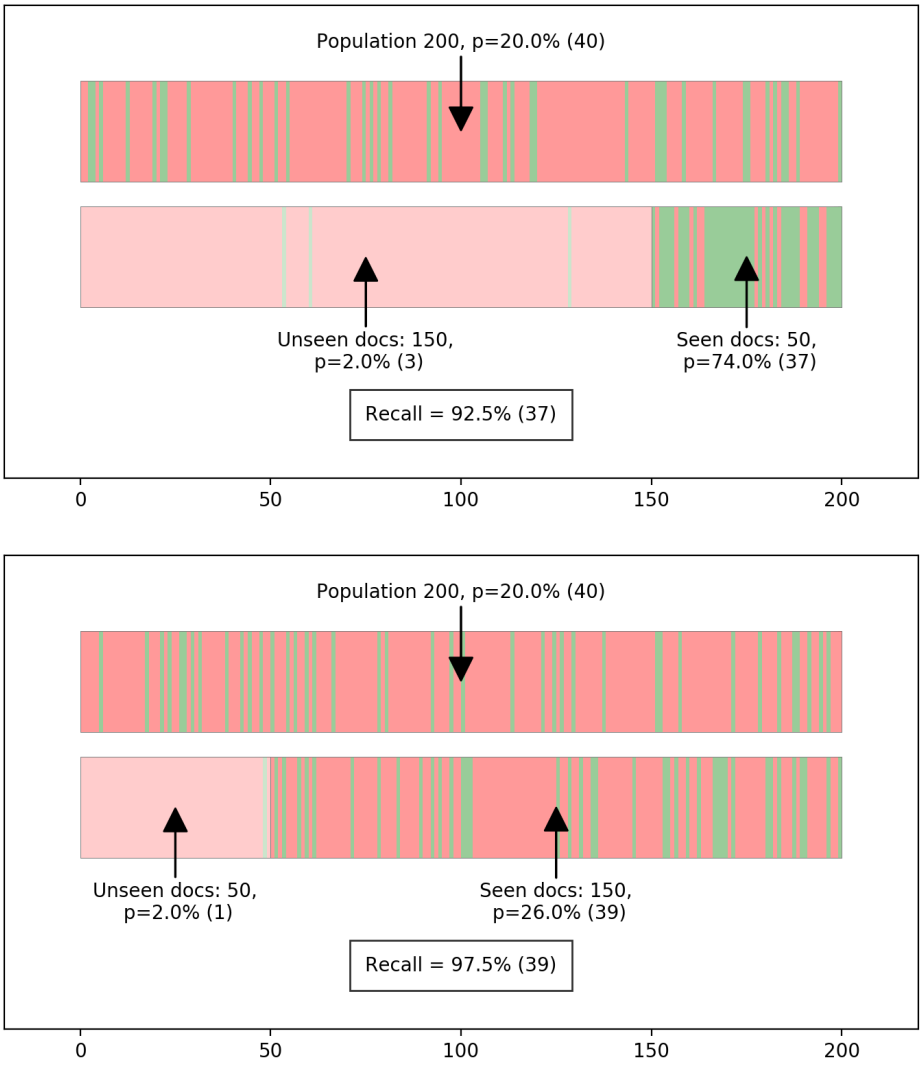


Figure 2: Similar low proportions of relevant documents in unseen documents with different consequences for recall. The top bar shows a random distribution of relevant documents (green) and irrelevant documents (red) at a given proportion of relevance. The bottom bar shows distributions of relevant and irrelevant documents in hypothetical sets of seen (right) and unseen (left - transparent) documents.

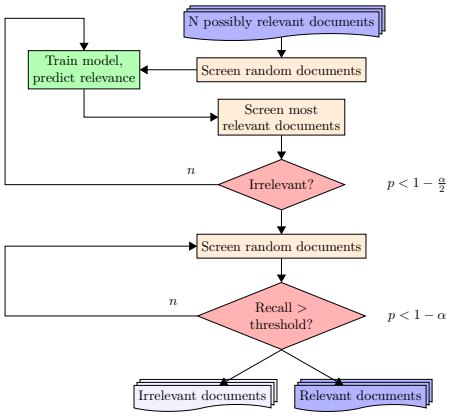
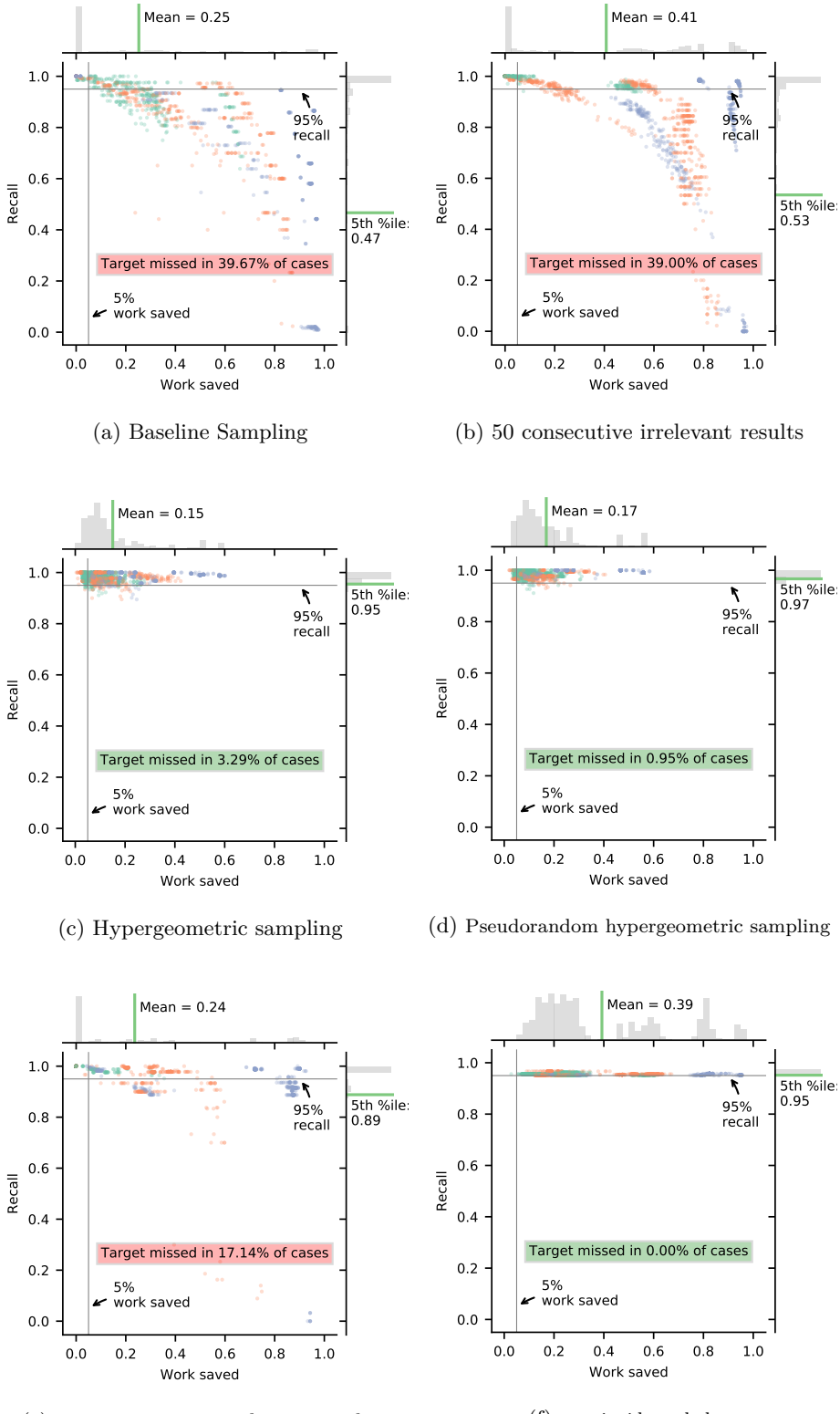


Figure 3: A workflow for active learning in screening with a statistical stopping criterion



(a) Baseline Sampling

(b) 50 consecutive irrelevant results

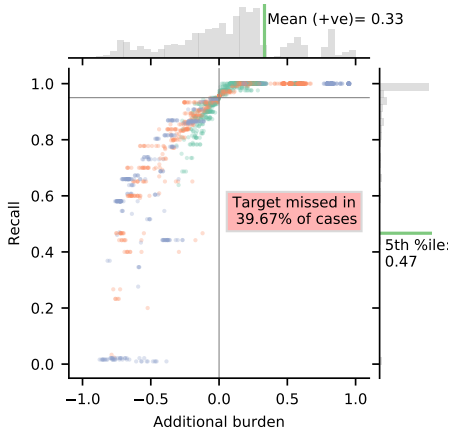
(c) Hypergeometric sampling

(d) Pseudorandom hypergeometric sampling

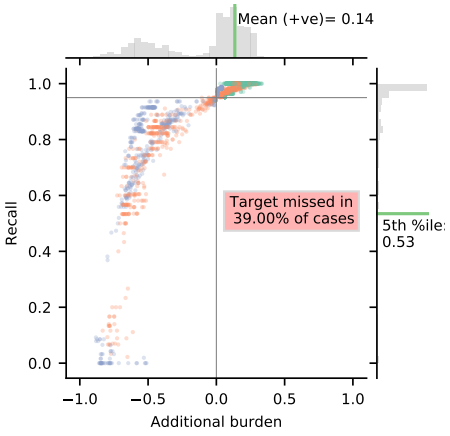
(e) 200 consecutive irrelevant results

(f) *a priori* knowledge

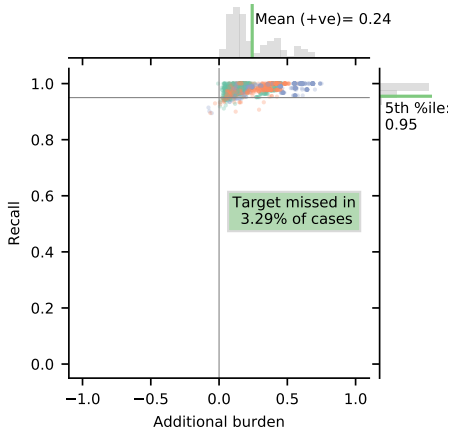
Figure 4: Distribution of recall and work saved after each stopping criteria. Green dots show results for datasets with less than 1,000 documents, orange dots show datasets with 1,000 - 2,000 documents, and blue dots show datasets with more than 2,000 documents



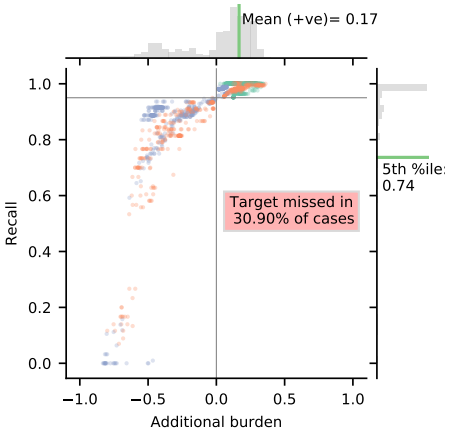
(a) Baseline Sampling



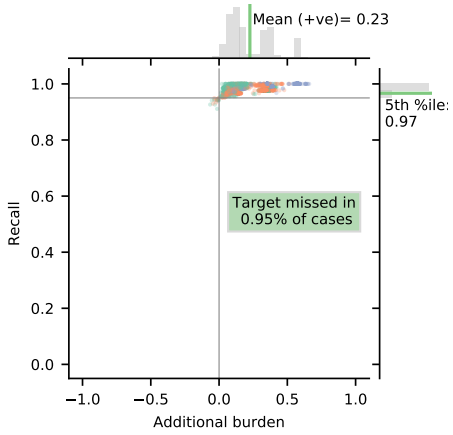
(b) 50 consecutive irrelevant results



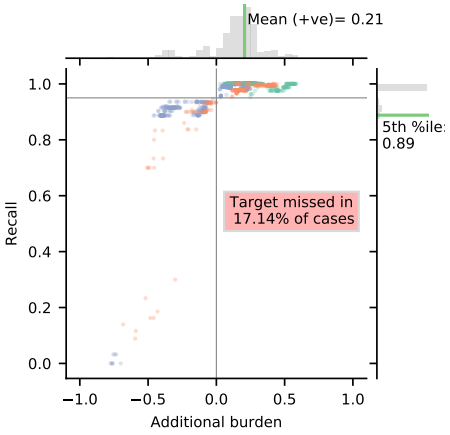
(c) Hypergeometric sampling



(d) 100 consecutive irrelevant results



(e) Pseudorandom hypergeometric sampling



(f) 200 consecutive irrelevant results

Figure 5: Distribution of recall and additional burden after each stopping criteria. Additional burden is the work saved when the criteria was trigerred subtracted from the work saved when the target was reached.

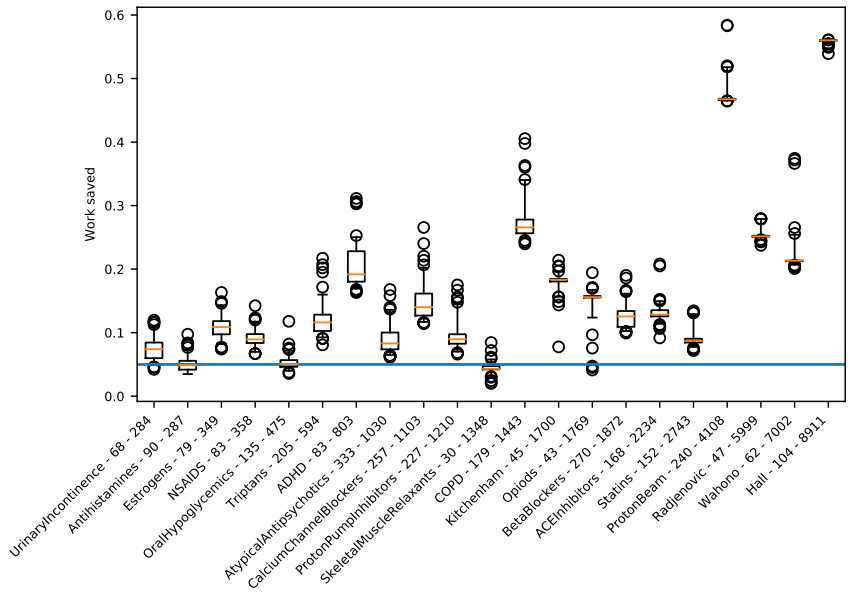
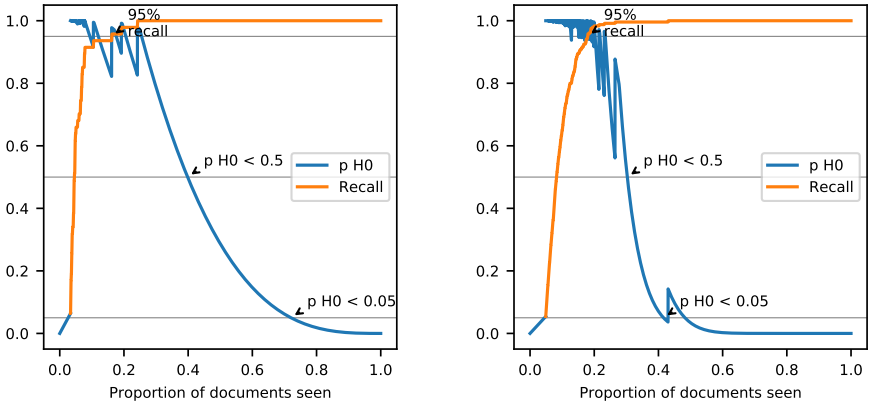
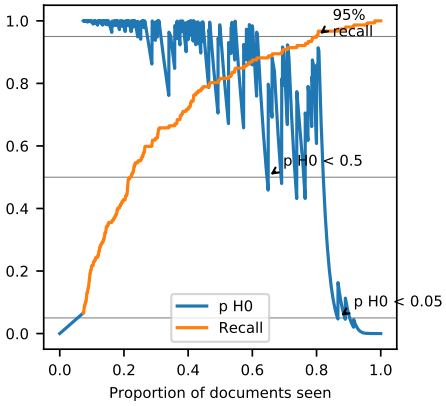


Figure 6: Work saved for the pseudorandom sampling method in each dataset. Labels show the number of relevant documents and the total number of documents. The datasets are presented in order of the number of documents. The whiskers represent the 5th and 95th percentiles.

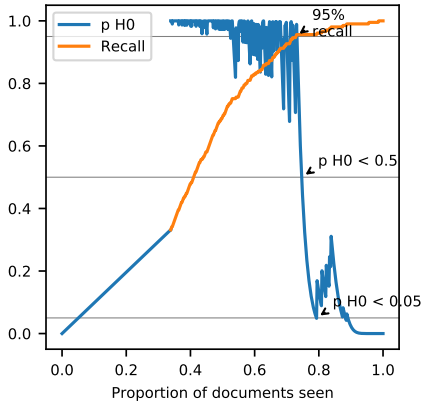


(a) Radjenovic

(b) ProtonBeam



(c) Statins



(d) Triptans

Figure 7: The path of recall (yellow) and the p-value of H0 for four different datasets