METHODOLOGY

Robust Statistical Stopping Criteria for Automated Screening in Systematic Reviews

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 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 with robust and flexible statistical stopping criteria, which offer real work reductions on the basis of a given confidence level of reaching a given recall. The stopping criteria are shown on test datasets to achieve a reliable level of recall, while still providing work reductions of on average 17%. Other methods proposed previously 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 [?]. 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 [?]. 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 [? ? ? ? ?] to reduce the time required to screen documents. This paper sets out how current approaches are [..[1]]unreliable in practice, and outlines and evaluates [..[2]]modifications 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 [?]. 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

[1] removed: unsuitable
[2] removed: a small modification

likely to be relevant, a human reviewer most often identifies all relevant studies [..[3]]– or a given proportion of relevant studies ([..[4]]described by recall: the number of relevant studies identified divided by the total number of relevant studies) – 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. 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 [**?**]. The lack of appropriate stopping criteria has therefore been identified as a research gap [**? ?**], although some approaches have been suggested. These [..[5]]have most commonly fallen 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 [**?**].
- **Novel automatic stopping criteria:** Recent papers have proposed more complicated novel systems for automatically deciding when to stop screening [**? ? ?**]

We review the first three classes of these methods in the following section and discuss their theoretical limitations. They are then tested on several previous systematic review datasets. We demonstrate theoretically and with our experimental results, that these three classes of methods can not deliver consistent levels of work savings or recall - particularly across different domains, or datasets with different properties [**?**]. We also discuss the limitations of novel automatic stopping criteria, which have all demonstrated promising results, but do not achieve a given level of recall in a reliable or reportable way. 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. If the hypothesis can be rejected, we conclude that the recall target has been achieved with a given confidence level and screening can be stopped. This allows AL users to predefine a target in terms of uncertainty and recall, so that they can make transparent, easily communicable statements like “A recall of more than 95% was achieved with a confidence of more than 95%”.

[3] removed: -
[4] removed: recall) -
[5] removed: fall

[..[6]]

In the remainder of the paper, we first discuss in detail the shortcomings of existing stopping criteria. Then, we introduce our new [..[7]]criteria based on a hypergeometric test. We evaluate our stopping criteria, and compare their performance with heuristic and sampling based criteria on real-world systematic review datasets on which AL systems have previously been tested [? ? ? ?].

Methods Review

We start by explaining the sampling and heuristic based stopping criteria and discussing their methodological limitations.

Sampling Based Stopping Criteria

The stopping criterion suggested by Shemilt et al. [?] involves establishing the Baseline Inclusion Rate (BIR), by taking a random sample at the beginning of screening. The BIR 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 correctly take into account sampling uncertainty [8]. This uncertainty is crucial, as errors can have severe consequences. Let us assume that users will stop screening when they have identified 95% of the relevant number of documents. If the estimated number of relevant documents is [..[9]]more than the true number of relevant documents divided by 0.95, then the users will never see 95% of the estimated number. This means that they will keep screening until they have seen all documents, and no work savings will be achieved. [..[10]]Conversely, if the number of relevant documents is underestimated by even a single unit, then the recall achieved will be [..[11]]lower than the target.

The number of relevant documents drawn without replacement from a finite sample of documents follows the hypergeometric distribution. Figure 1a shows the distribution of the predicted number of documents after drawing 1,000 documents from a total of 20,000 documents, where 500 documents (2.5%) are relevant. The left shaded portion of the graph shows all the cases where the recall will be less than 95%. This occurs [..[12]]48% of the time. The right shaded portion of the graph shows the cases where the number of relevant documents is overestimated [..[13]]so much that no work savings could be made to achieve a target recall of 95%.

[6] removed: 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.

[7] removed: criterion

[8] Although Shemilt et al. [?] employ a method to choose a sample size based on uncertainty, they fail to acknowledge the potential implications for recall of their choice. Their margin of error of 0.0025 and observed proportion of relevant studies of 0.0005 translate to estimates of 400 ± 451 relevant results. To reduce the margin of error to ±5% of estimated relevant studies, they would have had to screen 638,323 out of 804,919 results. See the notebook `https://github.com/mcallaghan/rapid-screening/blob/master/analysis/bir_theory.ipynb` that accompanies this paper for a detailed discussion.

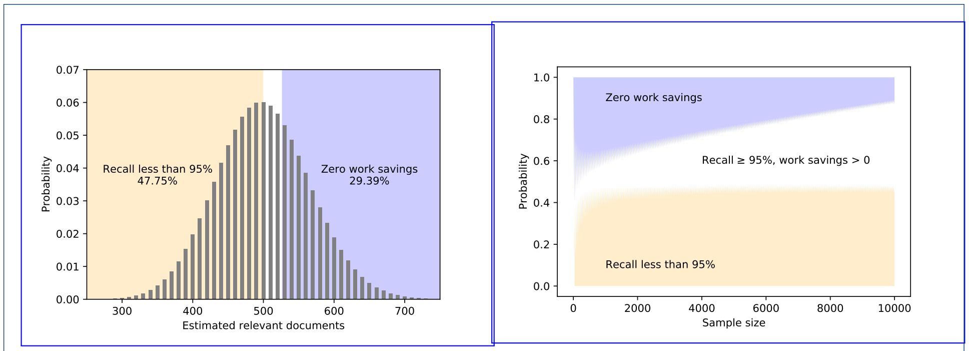
[9] removed: even one unit above the true value, then

[10] removed: If

[11] removed: less than 100%

[12] removed: 35.91

[13] removed: and



(a) Error types after 1,000 documents

(b) Error types across sample sizes

Figure 1: Distribution of under- or over-estimation errors using the BIR sampling method in a dataset of 20,000 documents of which 500 are relevant. Panel (a) shows the probability distribution of the estimated number of relevant documents after a sample of 1,000 documents. Panel (b) shows the probability of each type of error according to the sample size.

This occurs [..[14]]29% of the time. In only [..[15]]23% of cases can work savings be achieved while still achieving a recall of at least 95%.

Figure 1b shows the probability distribution of these errors according to the sample size. Even in very large samples both types of error remain frequent[..[16]]. This shows how baseline estimation inevitably offers poor reliability, [..[17]]either in terms of recall [..[18]]or in work saved.

Heuristic Stopping Criteria

Some studies give the example of heuristic stopping criteria based on drawing a given number of irrelevant articles in a row [? ?]. We take this as a proxy for estimating that the proportion of documents remaining in the unseen documents is low, as the probability of observing 0 relevant documents in a given sample (analogous to a set of consecutive irrelevant results) is a decreasing function of the number of relevant documents in the population. We find this a promising intuition, but argue that 1) it ignores uncertainty, as discussed in relation to the previous method; [..[19]]2) it lacks a formal description that would help to find a suitable threshold for the criterion; and 3) it misunderstands the significance of a low proportion of relevant documents in estimating the recall.

Figure 2 illustrates this [..[20]]third 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 the screened documents were relevant. In the bottom figure, ML has performed less well, and only 26% of the screened documents were relevant. In both cases, only 2% of unseen documents are relevant, but 2% of a larger number means more relevant documents

[14] removed: 46.24
[15] removed: 17.85
[16] removed:)
[17] removed: both
[18] removed: and
[19] removed: and
[20] removed: second

are missed. Recall is not simply a function of the [..[21]]proportion of unseen relevant 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. This shows us that even a perfect estimator of the proportion of unseen relevant documents is insufficient on its own to provide sufficient information about when to stop screening. To estimate recall reliably, it is necessary to take into account the total number of unseen relevant documents (or their proportion times the number of unseen documents).

[..[22]]Pragmatic stopping criteria

Wallace et al. [?] 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. Jonnalagadda and Petitti [?] 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.

Novel automatic stopping criteria

Two examples come from the information retrieval literature. Di Nunzio [?] presents a novel automatic stopping criterion based on BM25, although recall reported is "often between 0.92 and 0.94 and consistently over 0.7". Yu and Menzies [?] [..[23]]also present a stopping criterion based on BM25 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%. [..[24]]In each case, the reliability of the estimate is dependent on the performance of the model.

Finally, Howard et al. [?] present a method to estimate recall based on the number of irrelevant documents D observed in a list of documents since the δth previous relevant document. They reason that this should follow the negative binomial distribution based on the proportion of remaining relevant documents p, and use this information to estimate $\hat{p}$, and with this, the total number of relevant articles and the estimated recall.

However, their method does not quantify uncertainty, but can only claim that the method "*[..[25]][..[26]]*tends to result in a conservative estimate of recall" (emphasis

[21] removed: relevance of unseen

[22] removed: Other

[23] removed: adopt a more complicated stopping criterion

[24] removed: Di Nunzio [?] also present an innovative stopping criteria, but it does not take into account uncertainty, and produces results

[25] removed: near

[26] removed: a target recallthreshold, rather than above it in a reliable proportion of cases.

ours). This is not guaranteed by the criterion itself but rather a finding of the simulation with example datasets. Further the authors do not give sufficient information to reproduce their results, providing neither code (they describe their own proprietary software), nor an equation for $\hat{p}$. Additionally, the criterion requires a tuning parameter δ, which users may have insufficient information to set optimally. Lastly, because screening is a form of sampling without replacement, the negative hypergeometric distribution should be preferred to the negative binomial, even though the latter can be a good approximation for cases with large numbers of documents.

These last examples are promising developments, but they all fail to take into account the needs of live systematic reviews, where the reliability of and ease of communication about recall are paramount, and the results are independent of model performance. In the following, we explain our own method, which provides clearly communicable estimates of recall, which manage uncertainty in a way robust to model performance.

Methods

A Statistical Stopping Criterion for Active Learning

[..[27]]

In our screening setup, we start off with N_{tot} documents that are potentially relevant. ρ_{tot} of these documents are actually relevant, but we don't know this value *a priori*. As we screen relevant documents we include them, so ρ_{seen} represents the number of relevant documents screened, and recall τ is given by

[..[28]]

$$\tau = \frac{\rho_{\mathrm{seen}}}{\rho_{\mathrm{tot}}} \tag{1}$$

We set a target recall τ_{tar} and a confidence level α. We want to keep screening until $\tau \geq \tau_{tar}$, and devise a hypothesis test to estimate whether this is the case with a given level of confidence. We do this based on interrupting the active-learning process and drawing a random sample from the remaining [..[29]]unseen documents. We first describe this test, before showing how a variation on the test can be used to decide when to begin drawing a random sample.

Random Sampling

At the start of the sample, N_{AL} is the number of documents seen during the active learning process, and N is the number of documents remaining, so that

$$\mathrm{N} = \mathrm{N}_{\mathrm{tot}} - \mathrm{N}_{\mathrm{AL}} \tag{2}$$

We refer to the number of relevant documents seen during active learning as ρ_{AL}, and the number of remaining relevant documents as K. We do not know the value

[27] removed: Random Sampling

[28] removed: After using machine learning to select which documents are screened by humans as described above, we begin

[29] removed: documents (when this happens is described below)

of K but know that it is given by the total number of relevant documents minus the number of relevant documents seen during active learning.

[..[30]]

$$\mathsf{K} = \rho_{\mathsf{tot}} - \rho_{\mathsf{AL}} \tag{3}$$

We now take random draws from the remaining N documents, and denote the number of documents drawn with n and the number of relevant documents drawn with k. The number of relevant documents seen is updated by adding the number of relevant documents seen since sampling began to the number of relevant documents seen during active learning.

$$\rho_{\mathsf{seen}} = \rho_{\mathsf{AL}} + \mathsf{k} \tag{4}$$

We proceed to form a null hypothesis that the true value of recall is less than our target recall:

$$\mathsf{H}_0 : \tau < \tau_{\mathsf{tar}} \tag{5}$$

Because we are sampling [..[31]]without replacement, we can use the hypergeometric distribution to [..[32]]find out the probability of observing k relevant documents in a sample of n [..[33]]documents from a population of N documents [..[34]]of which K [..[35]]are relevant. We know that k is distributed hypergeometrically:

$$k \sim Hypergeometric(N, K, n) \tag{6}$$

[..[36]]

We introduce a hypothetical value [..[37]]for K, which we call K_{tar}. This represents the minimum number of relevant documents [..[38]]remaining at the start of sampling compatible with our null hypothesis that recall is below our target.

[..[39]]

[..[40]]

[30] removed: We use random sampling to estimate the probability that a target recall τ has been achieved.

[31] removed: a binary outcome

[32] removed: formulate a statistical test. The hypergeometric distribution tells us

[33] removed: draws from a finite

[34] removed: with

[35] removed: relevant documents.

[36] removed: In our case, we know k, n and N after each draw, but K is unknown. We therefore substitute

[37] removed: $\hat{K}$:

[38] removed: in the sample had the recall target been missed

[39] removed: Recall R is given by the number of relevant documents that have been seen ρ_s over the number of relevant documents in the whole dataset ρ_{tot}

$$K_{tar} = \lfloor \frac{\rho_{seen}}{\tau_{tar}} - \rho_{AL} + 1 \rfloor \tag{7}$$

[..[41]]This equation is derived by combining Eqs. 1 and 4. Because k can only take integer values, K_{tar} is the smallest integer that satisfies the inequality in Eq. 5. With K_{tar}, we can reformulate our null hypothesis: the true number of relevant documents in the [..[42]]sample is greater than our hypothetical value.

[..[43]] $$H_0 : K > K_{tar} \tag{8}$$

[..[44]]We test this by calculating the probability of observing k or fewer relevant documents from the hypergeometric distribution given by K_{tar}, using the cumulative probability mass function.

$$p = \text{[..}^{[45]}\text{]} P(X \leq k), \text{ where } X \sim Hypergeometric(N, K_{tar}, n) \tag{9}$$

[..[46]]Because the cumulative probability mass function $P(X \leq k)$ is decreasing with increasing K, this gives the maximum probability of observing k [..[47]]for all values of K compatible with our null hypothesis. Similar arguments have been made to derive confidence intervals for estimating the parameter K in the hypergeometric distribution function [? ?] and the derivation of an equivalent criterion could use the upper limit of such a confidence interval of an estimated K from the observation of k.

We can reject our null hypothesis and stop screening if the maximum probability of obtaining our observed results given our null hypothesis p is below $1 - \alpha$. This means, we can [..[48]]report the likelihood that we achieve a recall above our target as being more than α [49].

[..[50]]Ranked quasi-sampling

[..[51]]We now proceed to describe a special case of the method described above which we (1) use as a heuristic in order to decide when to begin [..[52]]random sampling; and

[41] removed: The

[42] removed: whole dataset is the sum 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

[44] removed: 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

[46] removed: We use the cumulative distribution function of the hypergeometric distribution to estimate the probability p of having observed

[47] removed: or fewer relevant documents in the sample given $\hat{K}$. This function gives us an upper bound on the probability of observing no more than the number of relevant documents in our random sample that we did, if our recall target had not been achieved. If this is below our confidence level $1 - \alpha$

[48] removed: reject the null hypothesis that the recall target was not achieved

[49] The notebook, `https://github.com/mcallaghan/rapid-screening/blob/master/analysis/hyper_criteria_theory.ipynb`, in the github repository accompanying this paper contains a step by step explanation of this method with code and examples

[50] removed: Pseudo-random sampling

[51] removed: In

[52] removed: a random sample, we employ pseudo-random sampling, where we treat previously screened documents as a random sample. The distribution of relevant documents among

(2) test as an independent stopping criterion. The method works by treating batches of previously screened documents [..[53]]as if they were random samples.

[..[54]]We calculate p as above for subsets of the already screened documents. Concretely, we use subsets of documents A_i by looking back to the last i documents, $A_i = \{d_{N_{seen}-1}, ..., d_{N_{seen}-i}\}$, where the documents d are indexed in the order in which they have been screened. For a specific i, this corresponds to random sampling beginning after seeing i documents in the section above. Thus, we set N_{AL} to i, n to $N_{seen} - i$, ρ_{AL} to the number of relevant documents seen when i documents had been seen, and k to the number of relevant documents seen since i documents had been seen, and calculate p [..[55]]according to Eq. 9. We compute p for all sets A_i with $i \in N_{seen} - 1 \dots 1$. This gives us a vector $\boldsymbol{p}$, representing the values of p which would have been estimated at each point at which we could have stopped active learning and began random sampling. The lowest probability of our null hypothesis being true that we would have thereby obtained is given by p_{min}. With the vectorized implementation included in our accompanying code, these calculations are completed in less than the time it would take a human to code the next document.

First, we use this method as a useful heuristic for deciding when to stop active learning, and switch to random sampling. [..[56]]For this, we choose a higher threshold for the likelihood, $p_{min} < 1 - \frac{\alpha}{2}$. Second, we use the same ranked quasi-sampling as an independent stopping criterion, by continuing screening with active learning until $p_{min} < 1 - \alpha$. We present [..[57]]the [..[58]]results of this second procedure separately below.

Given that the documents seen during active learning are ranked according to predicted relevance, they do not in fact represent a random sample. This means that the test is unlikely to be accurate. It would be reasonable to assume that the proportion of relevant documents in each ranked quasi-sample is as high if not higher than the proportion of relevant documents in the unseen documents. This assumption would make this estimator conservative. As such it works in a similar way to the criterion proposed by Howard et al. [?], although it makes use of more information and provides hypothesis testing rather than just a point estimate of recall.

Evaluation

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

- **Actual Recall**: The recall when the stopping criteria was met

[53] removed: 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 screeneddocuments than among remaining unseen documents. This would make the following estimates conservative.

[54] removed: 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

[55] removed: , 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

[56] removed: 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

[57] removed: these in the results below as

[58] removed: psuedo-random sampling criterion

	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

- **WS-SC**: Work saved when the stopping criteria was met
- **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 [? ?], with 1-2 word n-grams taken from the document abstracts used as input data. We start with random samples of 200 documents (we do not employ Shemilt et al's methods for identifying the "optimal" sample size, as we showed these in the methods section to be unhelpful). 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. Theoretically, using smaller batch sizes could mean that the recall target is achieved more quickly, but this is a trade-off between computational time spent training, and the speed at which the algorithm can "learn". However this is a modelling choice which may affect work saved, but not recall. 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 applications for document screening by Aaron Cohen and co-authors in 2006 [?], along with the widely used Proton Beam [?] and COPD [?] datasets, and computer science datasets used to test FASTREAD [?]. 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.

All computational steps required to reproduce this analysis are documented online at `https://github.com/mcallaghan/rapid-screening`.

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.

Both the random sampling and the [..[59]]ranked sampling criteria achieve the target threshold of 95% in more than 95% of cases. [..[60]]That this is greater than 95% is accounted for by the fact that random sampling sometimes begins after the target has been achieved, in which case the null hypothesis would be a priori impossible. The ranked quasi-sampling criterion outperforms the random sampling criterion with respect to both recall and work savings, saving a mean of 17% of the work compared to 15%, and missing the target in only 0.95% compared to 3.29% of cases. In theory, the [..[61]]ranked 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 [..[62]]ranked quasi-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.

The baseline sampling criteria (Figure 4c) 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 stopping criteria, both for 50 consecutive irrelevant results (Figure 4d - 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.

In figure 5 we rescale the x axis, calling it additional burden, which is simply the work saved when the criterion is triggered minus the work saved when the recall target was actually achieved. This measure indicates whether the stopping criterion was triggered too early (negative values), or too late (positive values). The figure directly highlights the tradeoffs involved in deciding when to stop screening: For our criteria, there is mostly a small additional burden which comes with the necessity to make sure the desired recall target has been reached and reject the null hypothesis that this has not been the case. For the other criteria, there are many cases in which additional burden is negative, i.e. the criterion has been triggered too early. In these cases, however, the desired recall is hardly ever reached.

To help explain the different work savings that were observed in our experiments, we show the distribution of work savings from our [..[64]]ranked quasi-sampling

[59] removed: pseudorandom
[60] removed: In fact , the pseudorandom sampling
[61] removed: pseudorandom
[62] removed: pseudorandom sampling
[64] removed: 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, many documents need to be screened to achieve a high confidence that there are only few relevant documents remaining in the unseen ones. Therefore, 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: at the point that the 95% recall threshold was reached, the null hypothesis to disprove was that a minimum of 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 fewer screening decisions to rule out the possibility that the number of relevant documents left is incompatible 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 confidence. 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 small compared to the work saved by using machine learning. 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 small.

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 have the further advantage that they are independent of the choice or performance of the machine learning model. If a 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.

One caveat in the derivation of our criteria is that we did not address the potential problem of multiple testing formally. Such a derivation is mathematically challenging and beyond the scope of this paper. However, the performance of the criteria shows that this is of limited practical concern. Formally describing screening procedures with iterative testing should be a next step towards even more rigorous stopping criteria and should be fully worked out in future research.

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 [..[65]]drawbacks of existing stopping criteria for machine learning approaches to document screening, [..[66]]particularly with regard to reliability. We propose 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.

Ethics approval
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
All computational steps required to reproduce this analysis are documented online at `https://github.com/mcallaghan/rapid-screening`.

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

Author's contributions
MC designed the research and conducted the experiments. FMH contributed to the development of the statistical basis for the stopping criterion. Both authors wrote and edited the manuscript.

Acknowledgements
Max Callaghan is supported by a PhD scholarship from the Heinrich Böll Foundation. Finn Müller-Hansen acknowledges funding from the German Federal Ministry of Research and Education within the Strategic Scenario Analysis (START) project (grant reference: 03EK3046B).

[65] removed: unsuitability

[66] removed: and proposes

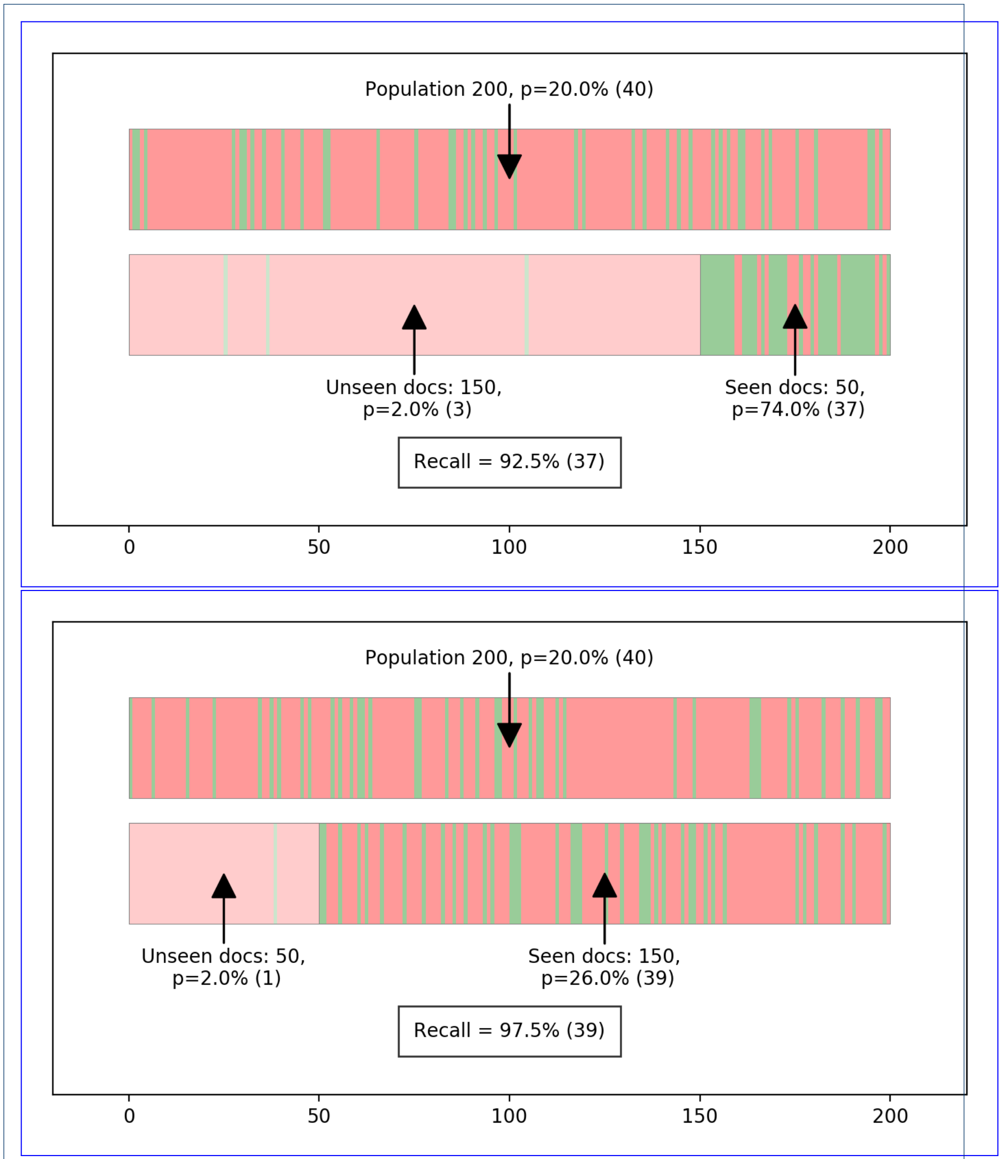


Figure 2: Similar low proportions of relevant documents in unseen documents with different consequences for recall. The top bar shows a random distribution of relevant documents (green) and irrelevant documents (red) at a given proportion of relevance. The bottom bar shows distributions of relevant and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

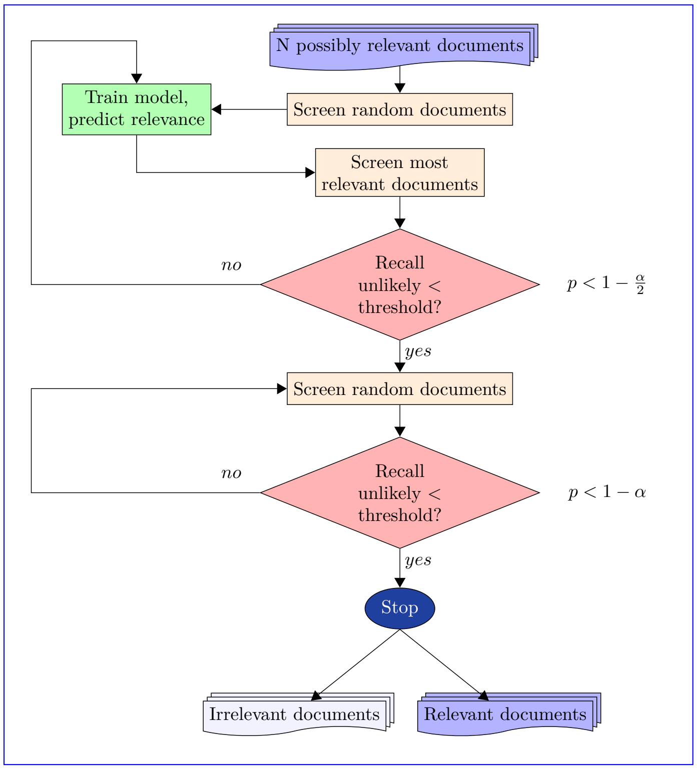
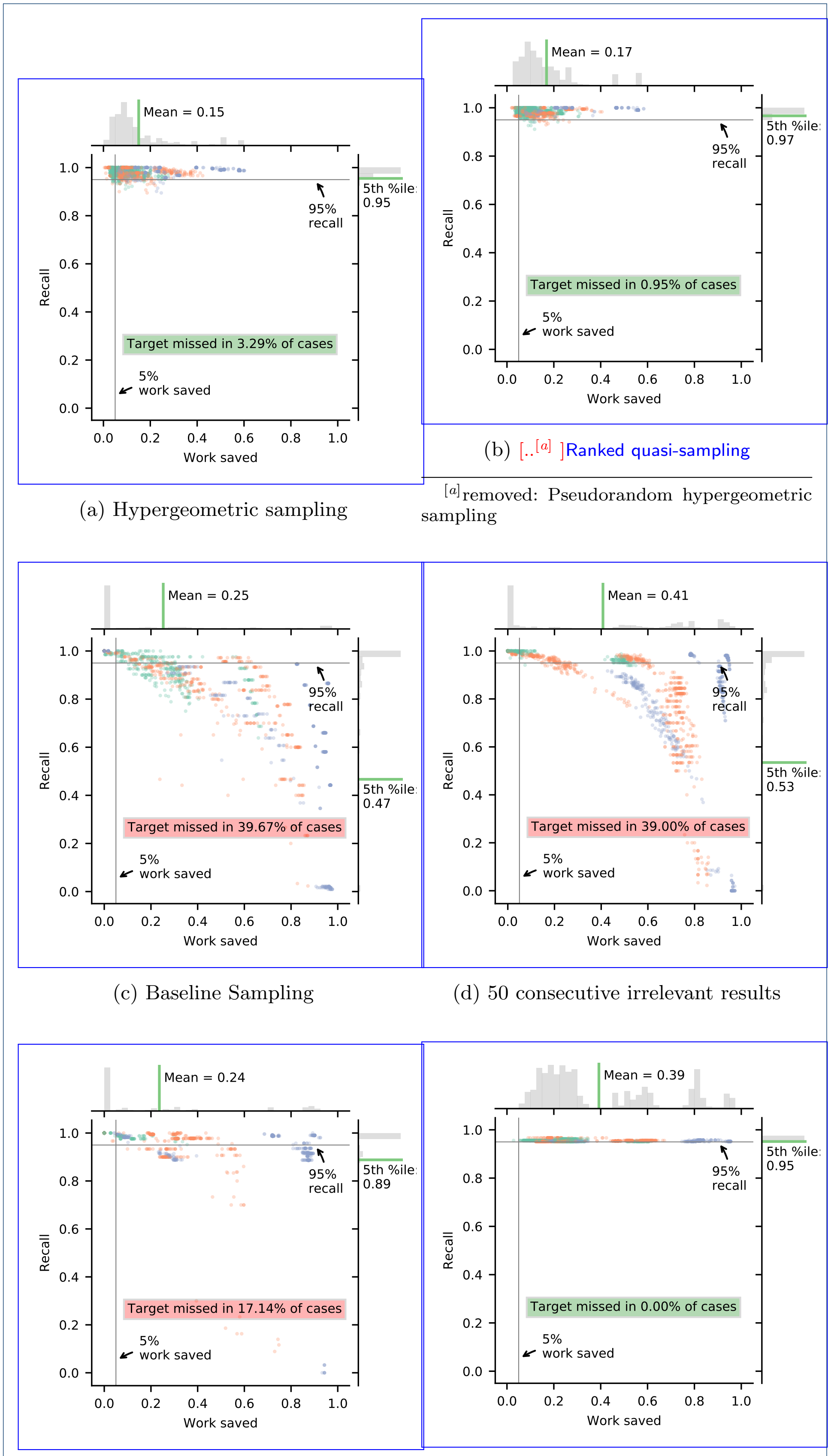

N possibly relevant documents
Screen random documents
Train model, predict relevance
Screen most relevant documents
Recall unlikely < threshold?
no
yes
$p < 1 - \frac{\alpha}{2}$
Screen random documents
Recall unlikely < threshold?
no
yes
$p < 1 - \alpha$
Stop
Irrelevant documents
Relevant documents

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



(a) Hypergeometric sampling

(b) [..[a]]Ranked quasi-sampling

[a]removed: Pseudorandom hypergeometric sampling

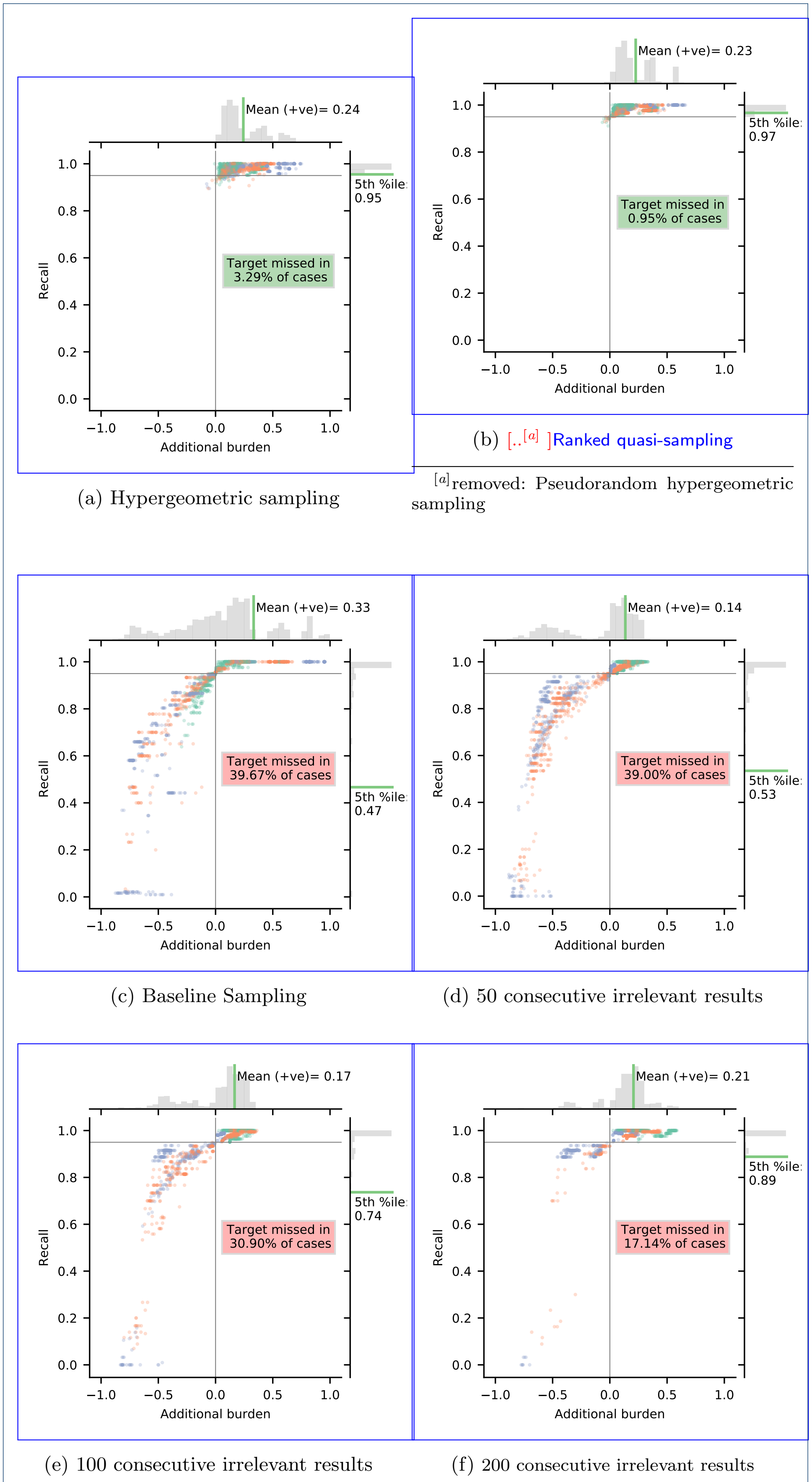
(c) Baseline Sampling

(d) 50 consecutive irrelevant results

(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) Hypergeometric sampling

(b) [..[a]]Ranked quasi-sampling

[a]removed: Pseudorandom hypergeometric sampling

(c) Baseline Sampling

(d) 50 consecutive irrelevant results

(e) 100 consecutive irrelevant results

(f) 200 consecutive irrelevant results

Figure 5: Distribution of recall and additional burden after each stopping criterion. Additional burden is the work saved when the criterion was triggered minus the work saved when the target was reached. Coloring of data points as in Fig. 4.

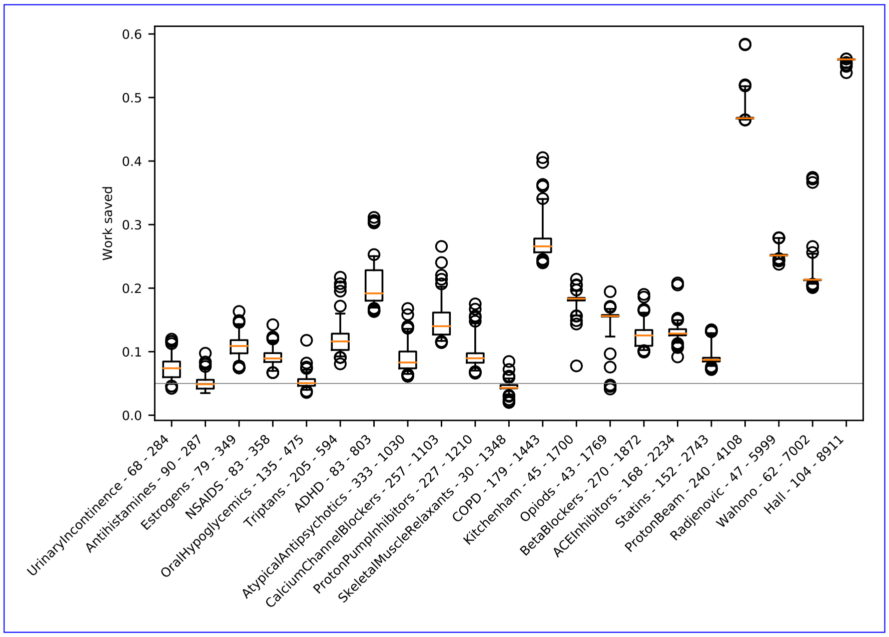
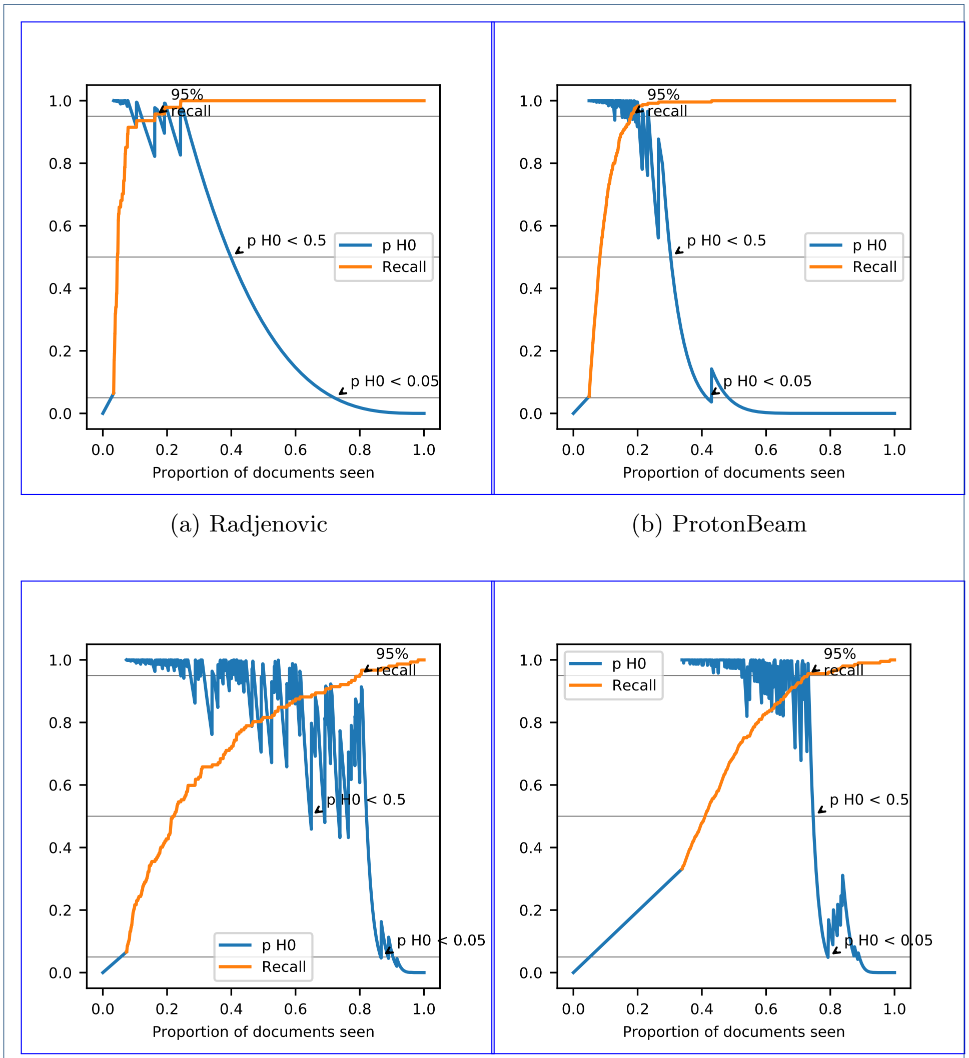


Figure 6: Work saved for the [..[63]]ranked quasi-sampling method in each dataset. Labels show the number of relevant documents and the total number of documents. The datasets are presented in order of the number of documents. The whiskers represent the 5th and 95th percentiles. The grey line shows work savings of 5%.



(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