Statistical Stopping Criteria for Automated Screening in Systematic Reviews

Max Callaghan, Finn Müller-Hansen

September 26, 2019

Context - the promise of work savings through machine learning

- When doing a systematic review, or other evidence synthesis project, you often need to screen a lot of documents.
- Large literature on training machine learning algorithms to recognise relevant documents and give these to us first O'Mara-Eves et al. (2015).
- The promise of this field is that we can stop before screening all documents, saving humans work.

But,

- How do we know what a good time to stop is?
- How can we report this?

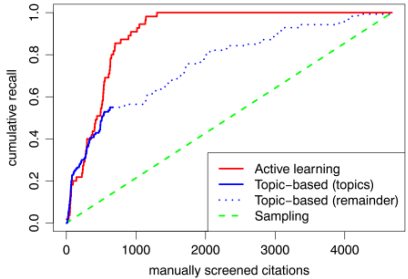
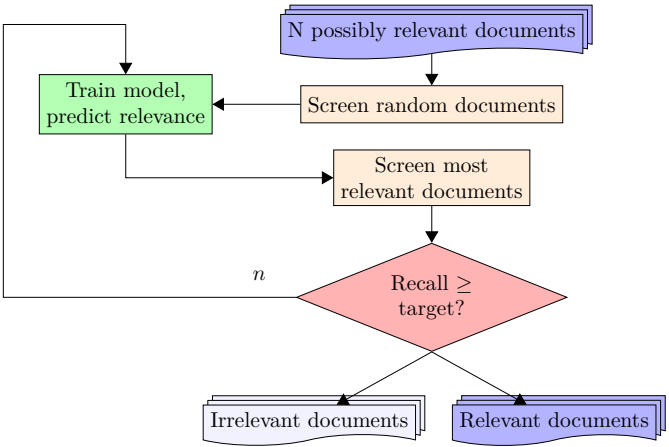


FIGURE 3 Cumulative recall curves for the *Tuberculosis* collection when using active learning versus topic-based screening at the beginning followed by random sampling for the remainder [Colour figure can be viewed at wileyonlinelibrary.com]

Source: Przybyła et al. (2018)

Workflows for reducing human effort in systematic review screening



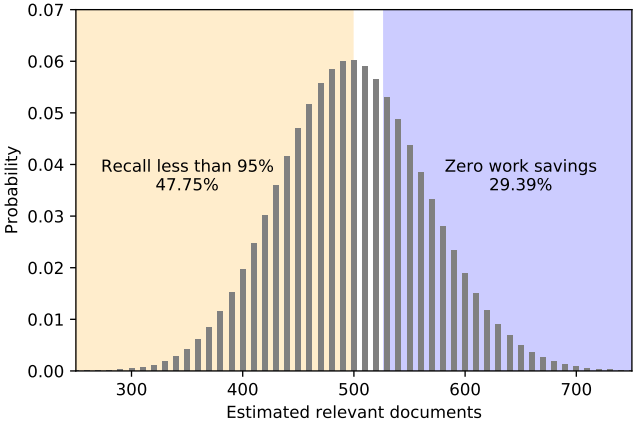
Existing approaches to deciding when to stop screening are unsatisfactory Bannach-Brown et al. (2019); Marshall and Wallace (2019), but fall into the following categories

- **BIR Sampling:** Estimate the number of relevant documents via sampling, stop when you have seen that number Shemilt et al. (2014)
- **Heuristics:** Stop after [x] consecutive irrelevant documents Jonnalagadda and Petitti (2013); Przybyła et al. (2018)
- **Novel automatic stopping criteria:** More sophisticated systems for automatically deciding when to stop screening Yu and Menzies (2019); Di Nunzio (2018); Howard et al. (2020)

Baseline Inclusion Rate (BIR) Sampling based criteria

1. Sample a fraction of large set of documents
2. Estimate the number of relevant documents based on the number seen in the sample
3. Screen until this number of relevant documents (or a proportion corresponding to target recall) has been seen.

Sampling error is not accounted for and can have serious consequences

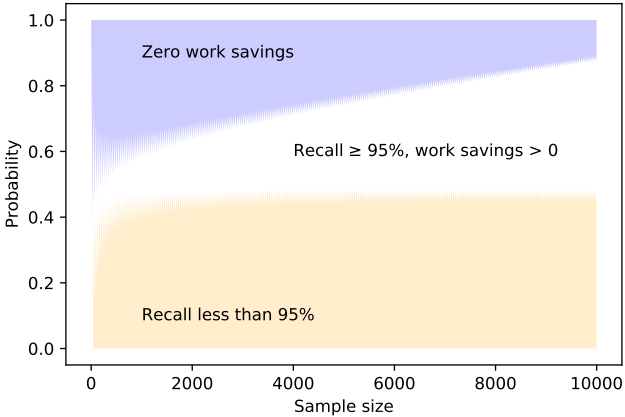


Distribution of errors after a sample of 1,000 using the BIR sampling method in a dataset of 20,000 documents of which 500 are relevant.

Baseline Inclusion Rate (BIR) Sampling based criteria

1. Sample a fraction of large set of documents
2. Estimate the number of relevant documents based on the number seen in the sample
3. Screen until this number of relevant documents (or a proportion corresponding to target recall) has been seen.

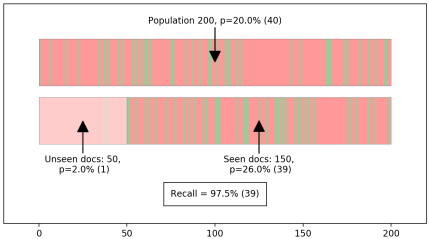
Sampling error is not accounted for and can have serious consequences



Distribution of errors across sample sizes using the BIR sampling method in a dataset of 20,000 documents of which 500 are relevant.

Heuristic criteria

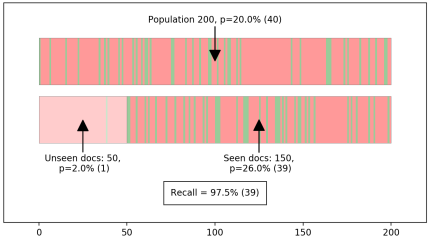
- If we see several consecutive irrelevant documents, we can surmise that the proportion of relevant documents remaining is low

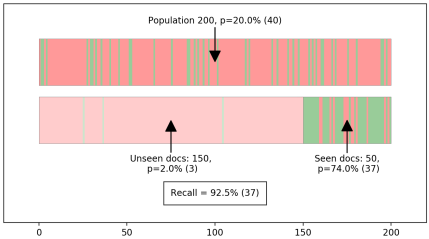


Heuristic criteria

- If we see several consecutive irrelevant documents, we can surmise that the proportion of relevant documents remaining is low

This intuition is helpful, but misses the fact that the proportion of relevant documents in the set of unseen documents can mean very different things





Novel automatic stopping criteria

Several recent papers Yu and Menzies (2019); Di Nunzio (2018); Howard et al. (2020) have suggested more sophisticated stopping criteria.

None of these have the same *fundamental* problems as those discussed, but

- None account properly for uncertainty
- None deliver *reliable* meeting of recall targets
- None offer users the ability to communicate when they stopped

If we want to use machine learning in live systematic reviews, we need criteria that can communicate the probable outcome of stopping early in a way that is clear and independent of machine learning approach and performance

A statistical stopping criteria for Active Learning

We test a null hypothesis that a given recall target has not been achieved, starting a random sample at an arbitrary point.

Definitions

- N_{tot} is the total number of documents
- ρ_{tot} is the total number of relevant documents
- ρ_{seen} is the number of relevant documents seen by a screener
- τ, or recall is $\frac{\rho_{seen}}{\rho_{tot}}$
- τ_{tar} is our recall target
- N_{AL} is the number of documents seen after active learning has finished (at start of random sample)
- N is the number of documents in the sample ($N_{tot} - N_{AL}$)
- K is the number of relevant documents in the sample ($\rho_{tot} - \rho_{AL}$)

After each draw from the sample:

- n is the number of documents drawn
- k is the number of relevant drawn

A statistical stopping criteria for Active Learning - II

We form a null hypothesis that the target recall has not been achieved

$$H_0 : \tau < \tau_{tar} \tag{1}$$

Accordingly, our alternative hypothesis is that recall is at least as large as our target:

$$H_1 : \tau \geq \tau_{tar} \tag{2}$$

Because we are sampling without replacement, we know that k is distributed hypergeometrically:

$$k \sim Hypergeometric(N, K, n) \tag{3}$$

A statistical stopping criteria for Active Learning - III

We introduce a hypothetical value for K, which we call K_{tar}. This represents the lowest value for K compatible with H_0

$$K_{tar} = \lfloor \frac{\rho_{seen}}{\tau_{tar}} - \rho_{AL} + 1 \rfloor \tag{4}$$

An testable analogue of H_0 is therefore

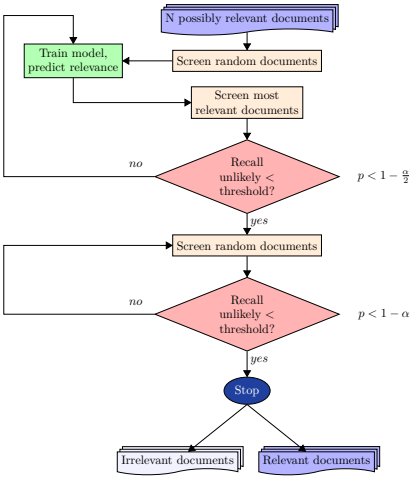
$$H_0 : K \geq K_{tar} \tag{5}$$

The cumulative distribution function gives us the probability of observing what we observed, if our null hypothesis were true

$$p = P(X \leq k), \text{ where } X \sim Hypergeometric(N, K_{tar}, n) \tag{6}$$

When $p < 1 - \alpha$, we can stop screening, and report, for example, that we reject the null hypothesis that we achieve a recall below 95% at the 5% significance level

A statistical stopping criteria for active learning - When to start a random sample?



In the approach described, we need to stop machine learning at some point and switch to random sampling

- If we stop too early, it will take us a long time to get to our recall target
- If we stop too late, we will save less work

We define a subcriterion, "ranked quasi-sampling", which treats previously screened documents as random samples.

- Ranked quasi-sampling is always conservative, as long as the proportion of relevant documents given to the screener by the machine learning algorithm $\geq$ the proportion in the remaining documents
- We test this as an independent criteria, but also use it to decide when to stop screening

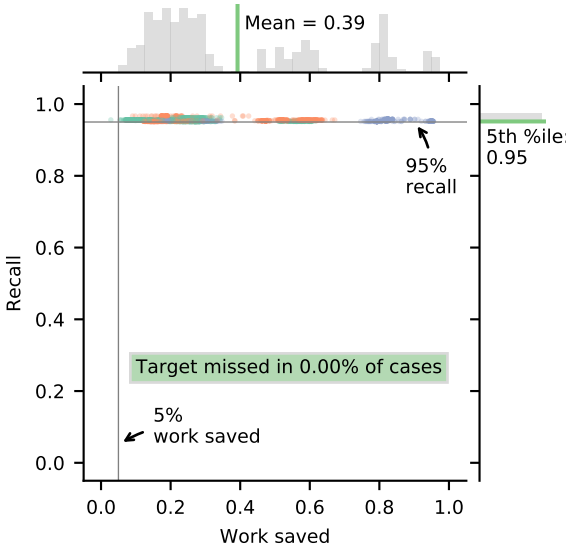
Evaluation

	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

Dataset properties

- We assemble a dataset of systematic reviews which other systems have been tested on
- We simulate 100 reviews on each of these
- In each review, a sample is drawn to begin with, then every 10 documents:
  - a simple SVM is trained on documents already seen
  - the relevance of unseen documents is predicted
  - the real relevance values of the next 10 documents are revealed
- We record when each criteria would have been achieved

Criteria performance - A priori knowledge

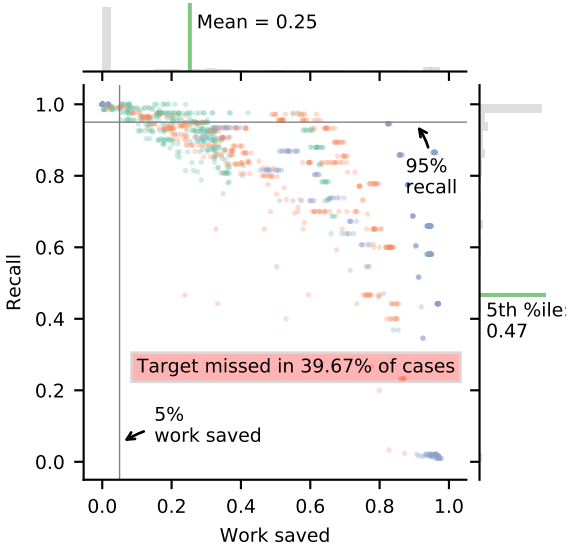


This is what would happen if we already knew how many relevant documents there were.

The target is never missed, and we can achieve some very large work savings.

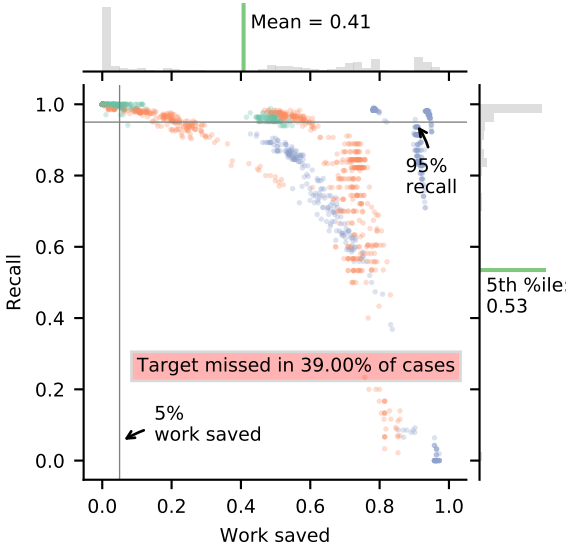
These are the results most often reported!

Criteria performance - Baseline estimation



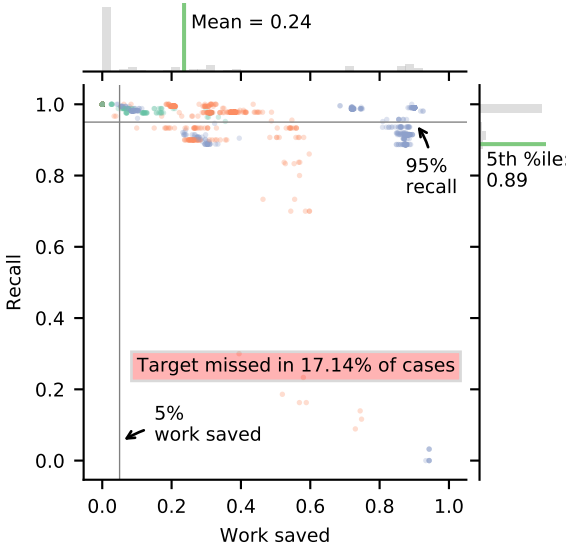
Estimating the number of relevant results based on a sample means we often (drastically!) miss our target, and we often save no work at all

Criteria performance - Heuristics



Stopping after 50 consecutive irrelevant results sometimes works, but often results in awful recall or no work savings

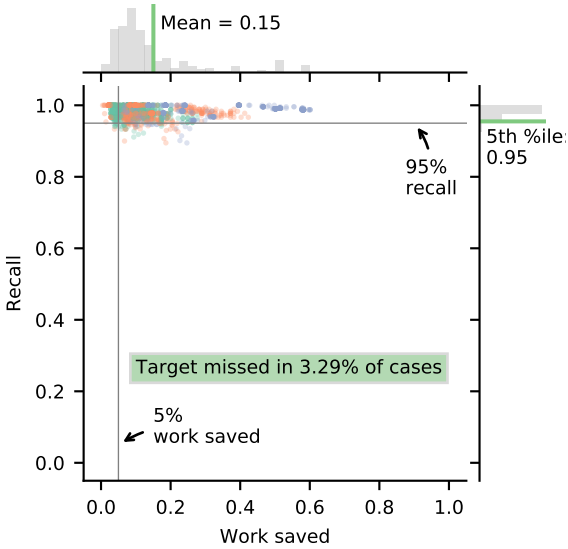
Criteria performance - Heuristics



Stopping after 50 consecutive irrelevant results sometimes works, but often results in awful recall or no work savings

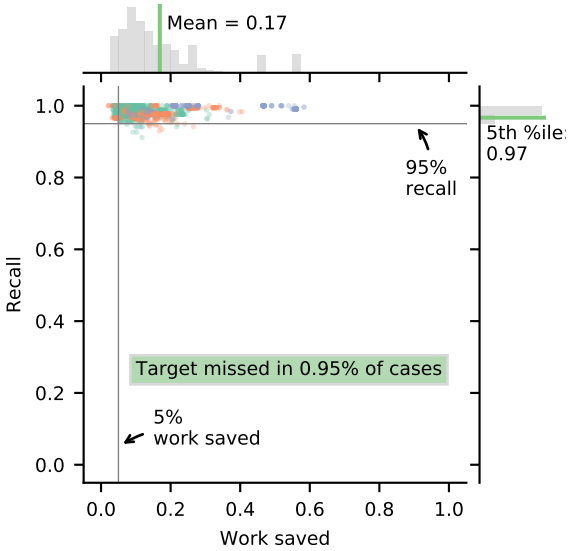
A stricter heuristic (200 consecutive irrelevant results) less often leads to poor recall, but also means less work is saved)

Criteria performance - Statistical stopping criteria (random sampling)



Our basic criterion makes modest work savings possible with a reliable achievement of the recall target

Criteria performance - Statistical stopping criteria (ranked quasi-random sampling)



Our criterion using ranked quasi-sampling turns out to be more conservative, and to allow slightly larger work savings, but this is not robust to catastrophic machine learning failure!

Discussion

- We provide reliable stopping criteria that realise *some* of the work savings in systematic review screening promised by machine learning
- Previously the outcome of a badly performing machine learning model could have been low recall levels (as low as a few percent). Now only work savings are at risk.
- We can use these in live reviews, because we can report the implications of our decision to stop early for recall

Work savings are modest but

- We used very general machine learning models and did not tune the parameters to perform better on individual datasets
- Larger savings are possible in larger datasets (where work savings are most helpful!)
- Forthcoming work will investigate how work savings depend on data features -> savings calculator for new projects

---

Thanks!

Working paper: `https://doi.org/10.21203/rs.2.18218/v2`
Contact: `callaghan@mcc-berlin.net`, Twitter: @MaxCallaghan5
Code: `https://github.com/mcallaghan/rapid-screening`

References

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S. C., Ananiadou, S., Liao, J., and Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, 8(1):1–12.

Di Nunzio, G. M. (2018). A study of an automatic stopping strategy for technologically assisted medical reviews. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A., editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10772 LNCS, pages 672–677, Cham. Springer International Publishing.

Howard, B. E., Phillips, J., Tandon, A., Maharana, A., Elmore, R., Mav, D., Sedykh, A., Thayer, K., Merrick, B. A., Walker, V., Rooney, A., Shah, R. R., Llc, S., Durham, D. D., Toxicology, N., Ntp, P., Sciences, H., and Rtp, T. W. A. D. (2020). SWIFT-Active Screener : Accelerated document screening through active learning and integrated recall estimation. *Environment International*, 138(April 2019):105623.

Jonnalagadda, S. and Petitti, D. (2013). A new iterative method to reduce workload in systematic review process. *International Journal of Computational Biology and Drug Design*, 6(1/2):5.

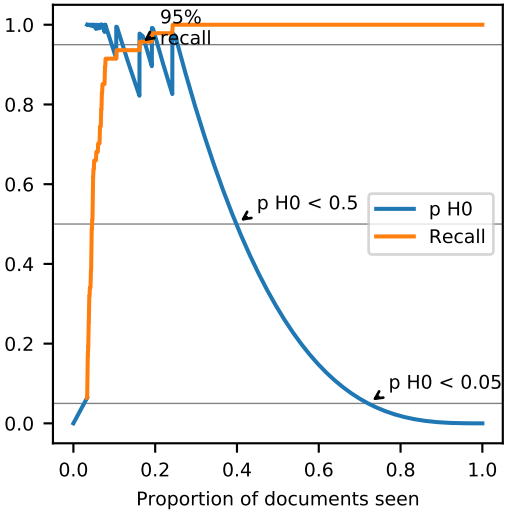
Marshall, I. J. and Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1):1–10.

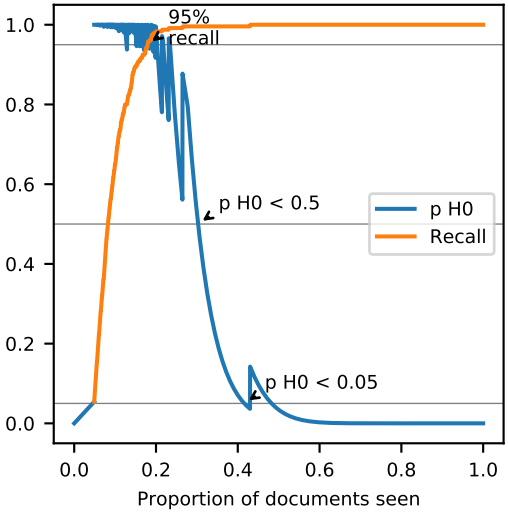
O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1):1–22.

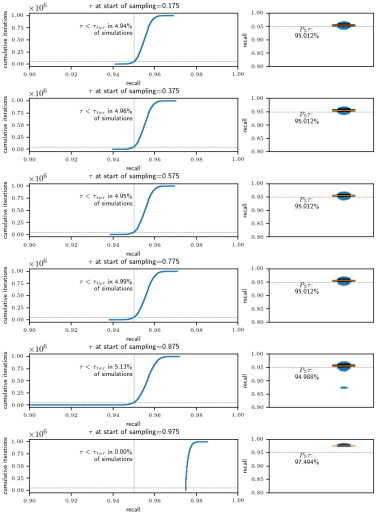
Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M. A., McNaught, J., von Elm, E., Nolan, K., and Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods*, 9(3):470–488.

Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., and Thomas, J. (2014). Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1):31–49.

Yu, Z. and Menzies, T. (2019). FAST 2 : An intelligent assistant for finding relevant papers. *Expert Systems with Applications*, 120:57–71.


95%
recall
p H0 < 0.5
p H0 < 0.05
p H0
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of documents seen


95%
recall
p H0 < 0.5
p H0 < 0.05
p H0
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of documents seen



During review, potential theoritical problems were raised around sequential testing.

- We show that the test performs well over a million simulations in different scenarios