Statistical stopping criteria for automated screening in systematic reviews

Max Callaghan, Finn Müller-Hansen

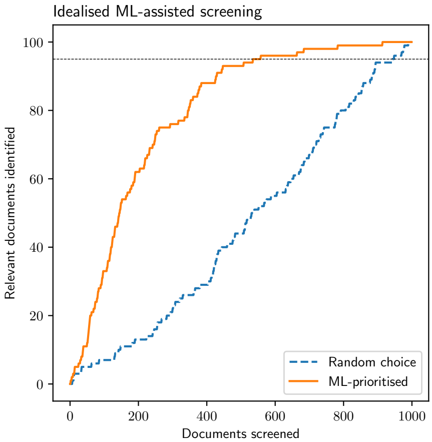
20 October 2022

Researcher in the Loop process

- A large literature (O'Mara-Eves et al., 2015) has developed "human-in the loop" machine learning applications which "overcome the manual and time-consuming screening of large numbers of studies by prioritizing relevant studies via active learning" van de Schoot et al. (2021)

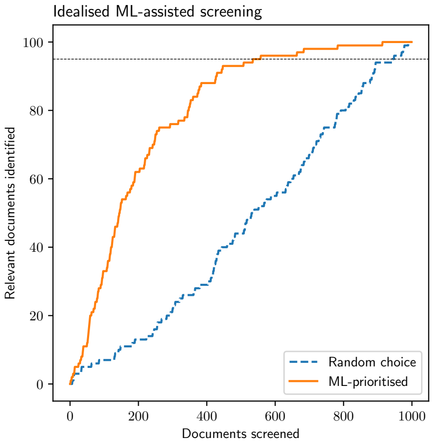
Researcher in the Loop process

- A large literature (O'Mara-Eves et al., 2015) has developed "human-in the loop" machine learning applications which "overcome the manual and time-consuming screening of large numbers of studies by prioritizing relevant studies via active learning" van de Schoot et al. (2021)
- Identifying 95% of relevant documents if screening manually would mean screening 95% of all documents



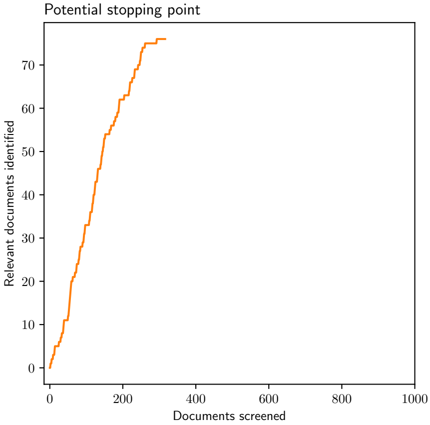
Researcher in the Loop process

- A large literature (O'Mara-Eves et al., 2015) has developed "human-in the loop" machine learning applications which "overcome the manual and time-consuming screening of large numbers of studies by prioritizing relevant studies via active learning" van de Schoot et al. (2021)
- Identifying 95% of relevant documents if screening manually would mean screening 95% of all documents
- If we can use machine learning to prioritise documents likely to be relevant, we can achieve high levels of recall with low levels of effort.



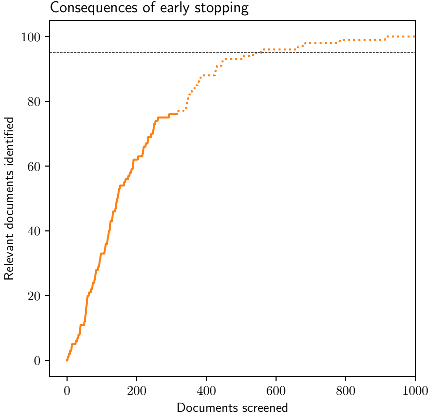
Researcher in the Loop process

- A large literature (O'Mara-Eves et al., 2015) has developed "human-in the loop" machine learning applications which "overcome the manual and time-consuming screening of large numbers of studies by prioritizing relevant studies via active learning" van de Schoot et al. (2021)
- Identifying 95% of relevant documents if screening manually would mean screening 95% of all documents
- If we can use machine learning to prioritise documents likely to be relevant, we can achieve high levels of recall with low levels of effort.
- BUT, given we do not know *a priori* the true number of relevant documents, we need strategies to stop screening and bank the work savings



Researcher in the Loop process

- A large literature (O'Mara-Eves et al., 2015) has developed "human-in the loop" machine learning applications which "overcome the manual and time-consuming screening of large numbers of studies by prioritizing relevant studies via active learning" van de Schoot et al. (2021)
- Identifying 95% of relevant documents if screening manually would mean screening 95% of all documents
- If we can use machine learning to prioritise documents likely to be relevant, we can achieve high levels of recall with low levels of effort.
- BUT, given we do not know *a priori* the true number of relevant documents, we need strategies to stop screening and bank the work savings
- Getting these wrong can mean missing our target



A stopping criterion

- Our stopping criterin works by treating documents as if they were white (not relevant) and red (relevant) marbles drawn from an urn without replacement.

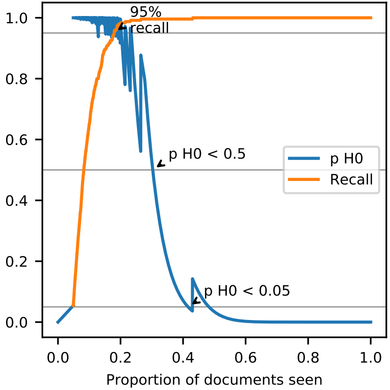
A stopping criterion

- Our stopping criterin works by treating documents as if they were white (not relevant) and red (relevant) marbles drawn from an urn without replacement.
- The hypergeometric distribution tells us the probability of observing **k** red marbles in a sample of **n** marbles, given an urn with **N** marbles, of which **K** are relevant

A stopping criterion

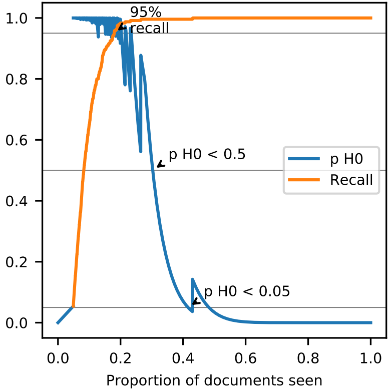
- Our stopping criterin works by treating documents as if they were white (not relevant) and red (relevant) marbles drawn from an urn without replacement.
- The hypergeometric distribution tells us the probability of observing **k** red marbles in a sample of **n** marbles, given an urn with **N** marbles, of which **K** are relevant
- We reformulate this to generate a null hypothesis H_0 that a given recall target has been **missed**.

A stopping criterion



- Our stopping criterin works by treating documents as if they were white (not relevant) and red (relevant) marbles drawn from an urn without replacement.
- The hypergeometric distribution tells us the probability of observing **k** red marbles in a sample of **n** marbles, given an urn with **N** marbles, of which **K** are relevant
- We reformulate this to generate a null hypothesis H_0 that a given recall target has been **missed**.
- We calculate a p-score for our null hypothesis, and if this is low enough, we reject H_0 and stop screening.

A stopping criterion



- Our stopping criterin works by treating documents as if they were white (not relevant) and red (relevant) marbles drawn from an urn without replacement.
- The hypergeometric distribution tells us the probability of observing **k** red marbles in a sample of **n** marbles, given an urn with **N** marbles, of which **K** are relevant
- We reformulate this to generate a null hypothesis H_0 that a given recall target has been **missed**.
- We calculate a p-score for our null hypothesis, and if this is low enough, we reject H_0 and stop screening.

Note, ML-prioritisation means documents are not drawn at random, which makes our test conservative as long as ML works as well as or better than random chance.

Results

We test our criteria against other commonly used criteria on 20 complete systematic review datasets

- *Potential* work savings (if we already knew when to stop) varied widely (higher for larger datasets - blue dots)

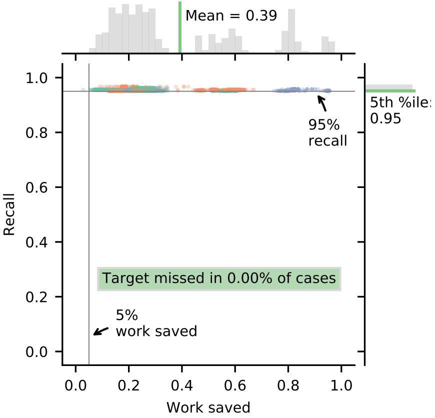


Figure: A priori knowledge

Results

We test our criteria against other commonly used criteria on 20 complete systematic review datasets

- *Potential* work savings (if we already knew when to stop) varied widely (higher for larger datasets - blue dots)
- Existing stopping criteria can result in catastrophic errors

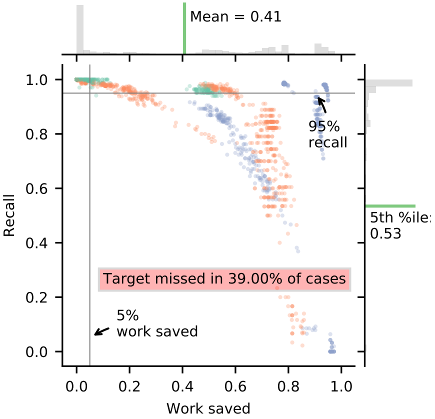


Figure: 50 consecutive irrelevant articles

Results

We test our criteria against other commonly used criteria on 20 complete systematic review datasets

- *Potential* work savings (if we already knew when to stop) varied widely (higher for larger datasets - blue dots)
- Existing stopping criteria can result in catastrophic errors

Figure: Estimating baseline inclusion rate

Results

We test our criteria against other commonly used criteria on 20 complete systematic review datasets

- *Potential* work savings (if we already knew when to stop) varied widely (higher for larger datasets - blue dots)
- Existing stopping criteria can result in catastrophic errors
- Our criteria generated work savings with reliably conservative performance wrt our recall target.

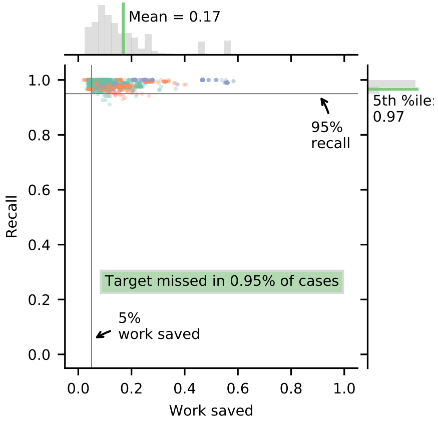
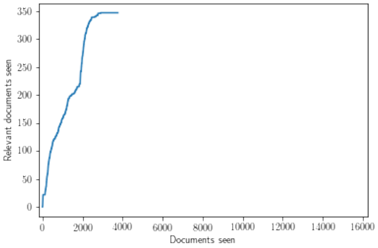


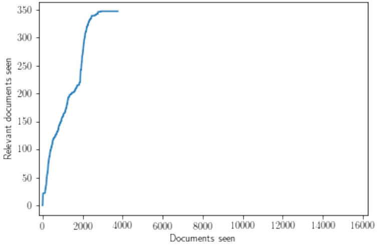
Figure: Our criterion

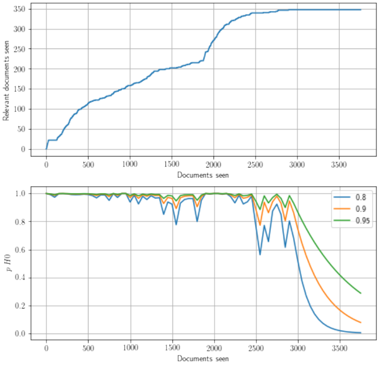
Applications and extensions



- We have used the stopping criteria to generate massive savings (77%) in real projects

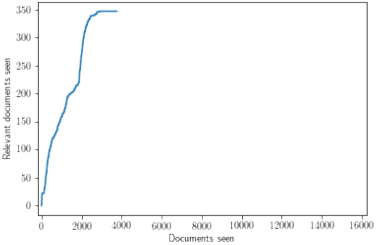
Applications and extensions

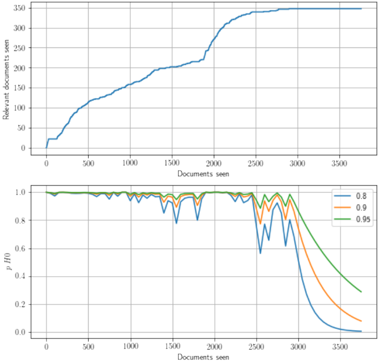




- We have used the stopping criteria to generate massive savings (77%) in real projects
- If rejecting our H_0 was less labour intensive we could have saved around 82%

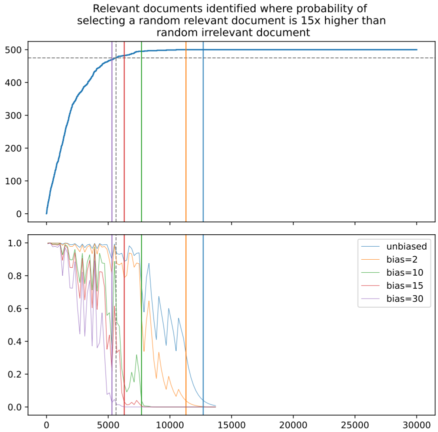
Applications and extensions





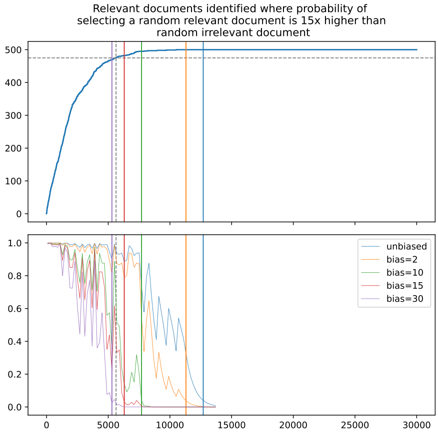
- We have used the stopping criteria to generate massive savings (77%) in real projects
- If rejecting our H_0 was less labour intensive we could have saved around 82%
- Using a biased urn could help create a more precise criterion

Biased urns



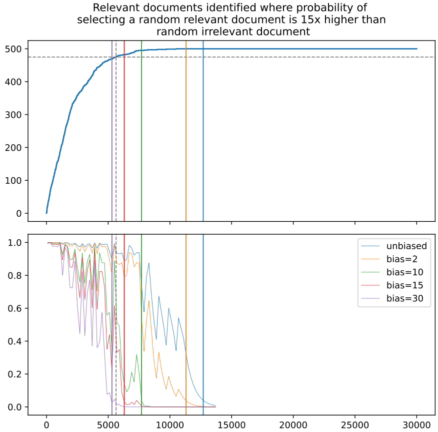
- Our original criteria treats documents as if they were drawn at random from an urn

Biased urns



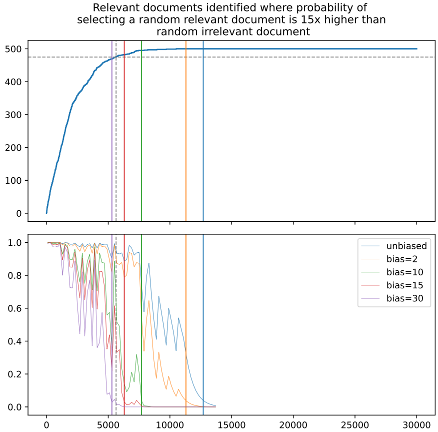
- Our original criteria treats documents as if they were drawn at random from an urn
- In fact, they are drawn in descending order of their predicted relevance

Biased urns



- Our original criteria treats documents as if they were drawn at random from an urn
- In fact, they are drawn in descending order of their predicted relevance
- Using a non-central hypergeometric distribution (Fog, 2008), we can input a **bias** parameter indicating how much more likely we are to draw a random relevant than a random non-relevant document.

Biased urns



- Our original criteria treats documents as if they were drawn at random from an urn
- In fact, they are drawn in descending order of their predicted relevance
- Using a non-central hypergeometric distribution (Fog, 2008), we can input a **bias** parameter indicating how much more likely we are to draw a random relevant than a random non-relevant document.
- Estimating this parameter is empirically non-trivial!

Realistic empirical evaluations for AI

To know when it is safe to use AI systems in future evidence synthesis projects, we need to evaluate them on past data ***under realistic conditions***

In living evidence applications:

Realistic empirical evaluations for AI

To know when it is safe to use AI systems in future evidence synthesis projects, we need to evaluate them on past data ***under realistic conditions***

In living evidence applications:

- How long can we trust classifier evaluations without labelling new data?

Realistic empirical evaluations for AI

To know when it is safe to use AI systems in future evidence synthesis projects, we need to evaluate them on past data ***under realistic conditions***

In living evidence applications:

- How long can we trust classifier evaluations without labelling new data?
- How does topic model fit decay over time?

Realistic empirical evaluations for AI

To know when it is safe to use AI systems in future evidence synthesis projects, we need to evaluate them on past data ***under realistic conditions***

In living evidence applications:

- How long can we trust classifier evaluations without labelling new data?
- How does topic model fit decay over time?
- How do we incorporate new topics?
- How frequently *would* LRs have been updated? Is this predictable?

Conclusion

We provide a stopping criteria that works on any model, with any tool:
https://github.com/mcallaghan/rapid-screening/blob/master/analysis/hyper_criteriaR.md.

Work savings in practice with large datasets are large!

Future work will identify how biased our urn is, in order to use a noncentral hypergeometric distribution, which should give a more precise, less conservative criterion.

Thanks!

Contact: mueller-hansen@mcc-berlin.net, callaghan@mcc-berlin.net

References

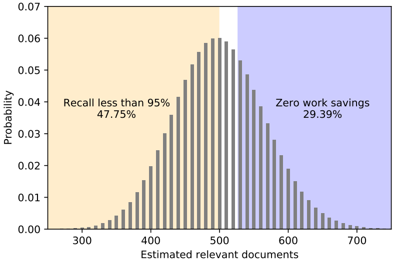
Fog, A. (2008). Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution. *Communications in Statistics - Simulation and Computation*, 37(2):258–273.

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):5.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., and Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133.

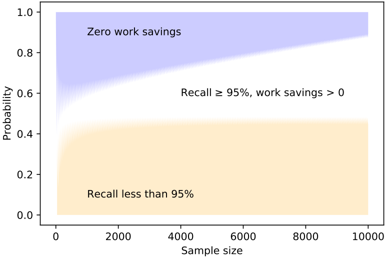
Baseline Inclusion Rate

- If we try to estimate the baseline inclusion rate, we will get it wrong most of the time. Overestimating results in 0 work savings, while underestimating results in less than target recall.



Baseline Inclusion Rate

- If we try to estimate the baseline inclusion rate, we will get it wrong most of the time. Overestimating results in 0 work savings, while underestimating results in less than target recall.
- Wrongness decreases with larger sample sizes, but bad outcomes remain most frequent.



Theory I

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

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

To operationalise this, we come up with a hypothetical value of K which is the lowest value compatible with our null hypothesis

$$K_{tar} = \lfloor \frac{\rho_{seen}}{\tau_{tar}} - \rho_{AL} + 1 \rfloor \tag{2}$$

In other words, if there were K_{tar} or more relevant documents in the urn when sampling began, the ρ_{al} relevant we identified before sampling, and the k we drew from the urn would not be enough to meet our target recall level.

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

$$p = P(X \leq k), \text{ where } X \sim Hypergeometric(N, K_{tar}, n) \tag{3}$$