Statistical stopping criteria for automated screening in systematic reviews

Max Callaghan, Finn Müller-Hansen

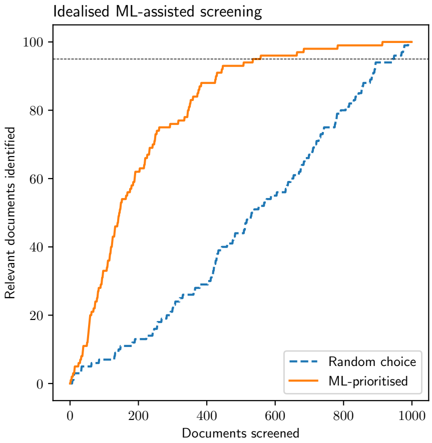
20 October 2022

ML-assissted document screening

- Growing number of "researcher-in-the-loop" machine learning applications for screening documents for systematic reviews (O'Mara-Eves et al., 2015; van de Schoot et al., 2021).

ML-assissted document screening

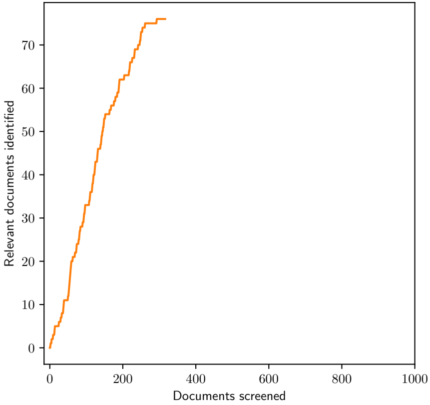
- Growing number of "researcher-in-the-loop" machine learning applications for screening documents for systematic reviews (O'Mara-Eves et al., 2015; van de Schoot et al., 2021).
- By using machine learning to prioritise documents likely to be relevant, we can achieve high levels of recall without screening all documents.



ML-assissted document screening

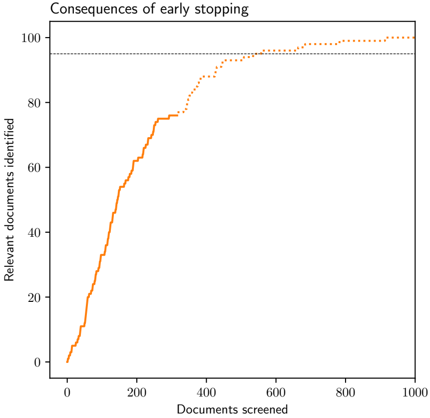
- Growing number of "researcher-in-the-loop" machine learning applications for screening documents for systematic reviews (O'Mara-Eves et al., 2015; van de Schoot et al., 2021).
- By using machine learning to prioritise documents likely to be relevant, we can achieve high levels of recall without screening all documents.
- BUT, given we do not know *a priori* the true number of relevant documents, we need criteria when to stop screening.

Potential stopping point



ML-assissted document screening

- Growing number of “researcher-in-the-loop” machine learning applications for screening documents for systematic reviews (O’Mara-Eves et al., 2015; van de Schoot et al., 2021).
- By using machine learning to prioritise documents likely to be relevant, we can achieve high levels of recall without screening all documents.
- BUT, given we do not know *a priori* the true number of relevant documents, we need criteria when to stop screening.
- Stopping too early can lead to huge biases in reviews.



A statistical stopping criterion

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

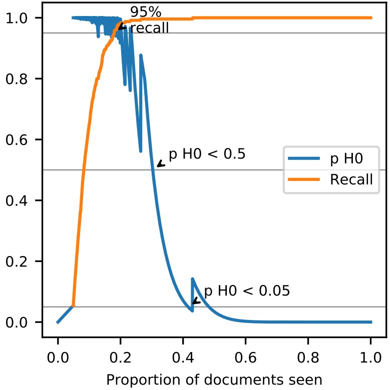
A statistical stopping criterion

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

A statistical stopping criterion

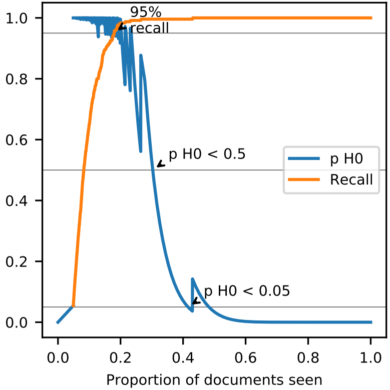
- Our stopping criterion works by treating documents as if they were white (not relevant) and red (relevant) marbles drawn from an urn without replacement.
- The hypergeometric distribution describes the probability of observing **k** red marbles in a sample of **n** marbles, given an urn with **N** marbles, of which **K** are red.
- We formulate a null hypothesis H_0 that a given recall target (e.g. 95% of relevant documents) has been **missed**.

A statistical stopping criterion



- Our stopping criterion works by treating documents as if they were white (not relevant) and red (relevant) marbles drawn from an urn without replacement.
- The hypergeometric distribution describes the probability of observing **k** red marbles in a sample of **n** marbles, given an urn with **N** marbles, of which **K** are red.
- We formulate a null hypothesis H_0 that a given recall target (e.g. 95% of relevant documents) has been **missed**.
- We calculate a p-score for H_0 and stop screening if this falls below a selected threshold.

A statistical stopping criterion



- Our stopping criterion works by treating documents as if they were white (not relevant) and red (relevant) marbles drawn from an urn without replacement.
- The hypergeometric distribution describes the probability of observing **k** red marbles in a sample of **n** marbles, given an urn with **N** marbles, of which **K** are red.
- We formulate a null hypothesis H_0 that a given recall target (e.g. 95% of relevant documents) has been **missed**.
- We calculate a p-score for H_0 and stop screening if this falls below a selected threshold.

Note: ML-prioritisation means documents are not drawn at random, which makes our test conservative.

Results

We run simulations on 20 complete systematic review datasets to test our criterion.

- *Theoretically achievable* work savings in the datasets varies widely (higher for larger datasets - blue dots)

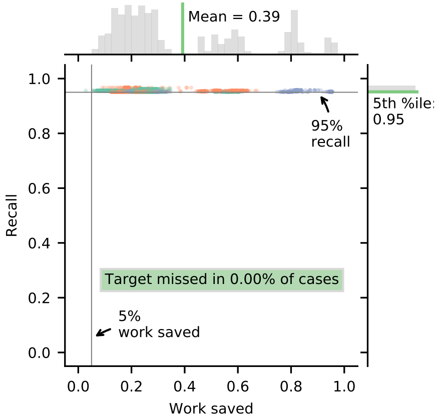


Figure: A priori knowledge

Results

We run simulations on 20 complete systematic review datasets to test our criterion.

- *Theoretically achievable* work savings in the datasets varies widely (higher for larger datasets - blue dots)
- Existing stopping criteria can result in very low recall (examples: 50 consecutive irrelevant articles, using baseline inclusion rate)

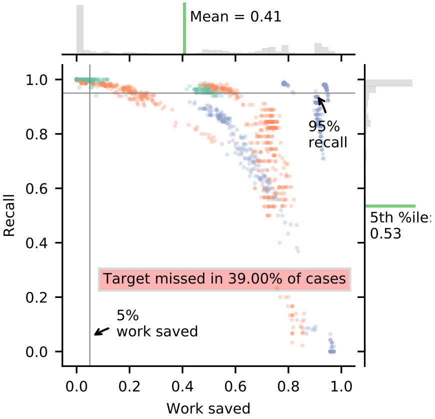


Figure: 50 consecutive irrelevant articles

Results

We run simulations on 20 complete systematic review datasets to test our criterion.

- *Theoretically achievable* work savings in the datasets varies widely (higher for larger datasets - blue dots)
- Existing stopping criteria can result in very low recall (examples: 50 consecutive irrelevant articles, using baseline inclusion rate)

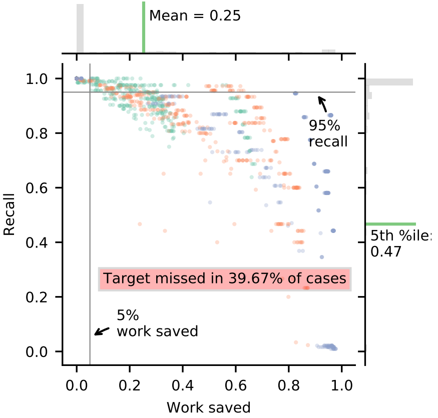


Figure: Estimating baseline inclusion rate

Results

We run simulations on 20 complete systematic review datasets to test our criterion.

- *Theoretically achievable* work savings in the datasets varies widely (higher for larger datasets - blue dots)
- Existing stopping criteria can result in very low recall (examples: 50 consecutive irrelevant articles, using baseline inclusion rate)
- Our criterion generated work savings with reliably conservative performance wrt our recall target.

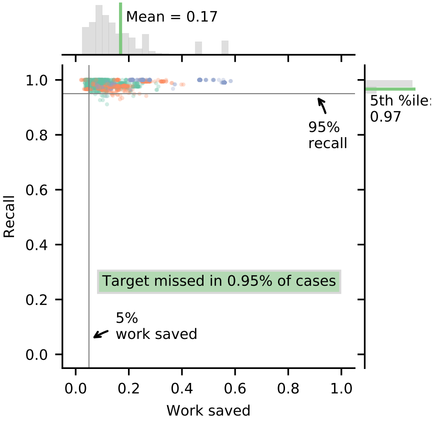


Figure: Our criterion

Conclusion

We provide a stopping criterion that works on any model and can be included in any tool: `https://github.com/mcallaghan/rapid-screening/blob/master/analysis/hyper_criteriaR.md`.

In practice, we see huge work savings for large datasets!

Future work: use noncentral hypergeometric distribution to generate a less conservative test and save more work.

Thanks!

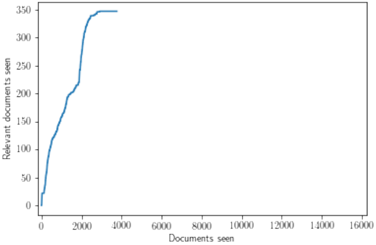
Contact: callaghan@mcc-berlin.net, mueller-hansen@mcc-berlin.net
Twitter: https://twitter.com/MaxCallaghan5

References

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.

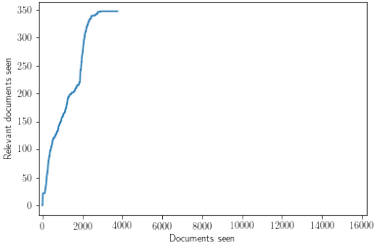
van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., and Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent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

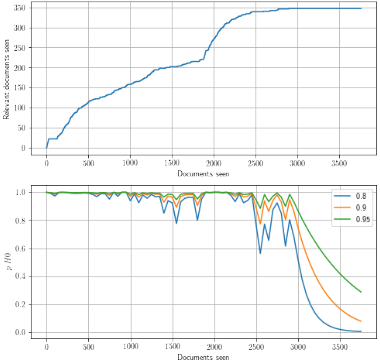
Applications and extensions



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

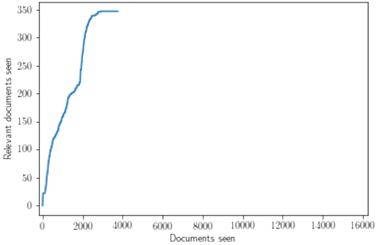
Applications and extensions

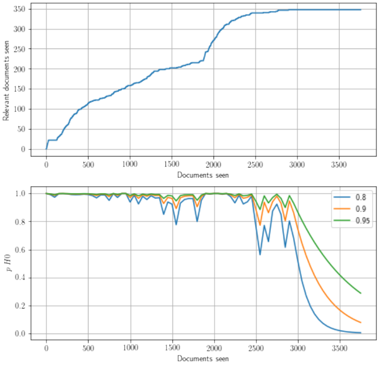




- We have used the stopping criterion to generate massive savings (77%) in real projects
- If rejecting our H_0 was less labour intensive we could have saved around 82%

Applications and extensions

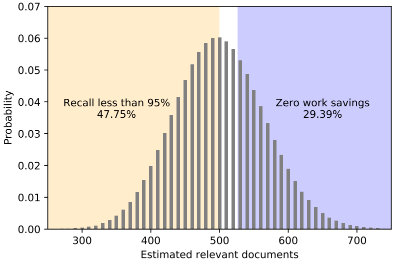




- We have used the stopping criterion to generate massive savings (77%) in real projects
- If rejecting our H_0 was less labour intensive we could have saved around 82%
- Using a biased urn could help create a more precise criterion

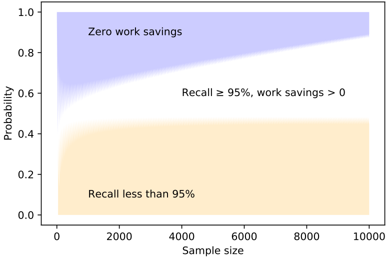
Baseline Inclusion Rate

- 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

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