A Simulation Study Assessing a Stopping Rule to Reduce Literature Review Title and Abstract Screening Burden

MSR163

Dolin O,[1] Langford B,[1] Masselot P,[2] Zhang H,[1] Gonçalves-Bradley DC,[1]

[1]Symmetron Limited, London, United Kingdom; [2]Environment & Health Modelling (EHM) Lab, Department of Public Health Environments & Society, London School of Hygiene and Tropical Medicine, London, United Kingdom • Poster inquiries: odolin@symmetron.net • www.symmetron.net • Presented at ISPOR EU 2024 Barcelona Annual Meeting

Introduction

- There is wide interest in using new artificial intelligence (AI) tools to speed up the literature review process in health economics and outcomes research (HEOR).
- One approach, prioritised screening, aims to make title and abstract (T&A) screening more efficient. Prioritised screening pushes relevant records to the front of the screening queue.
- Assuming it is only necessary to capture the majority of relevant records, substantial work savings could be achieved if an appropriate stopping rule is used with prioritised screening.
- The statistical stopping rule of Lewis and colleagues provides a rigorous, mathematical justification to stop screening early **(see Figure 1)** that ensures most relevant records have been found.[1]
- Work savings (the number of records in a dataset left un-screened) depend on prioritised screening performance and are difficult to assess *a priori*.

Objective:
Perform a simulation study to estimate the work savings produced in practice by Lewis and colleagues' stopping rule.

Methods

Simulation approach

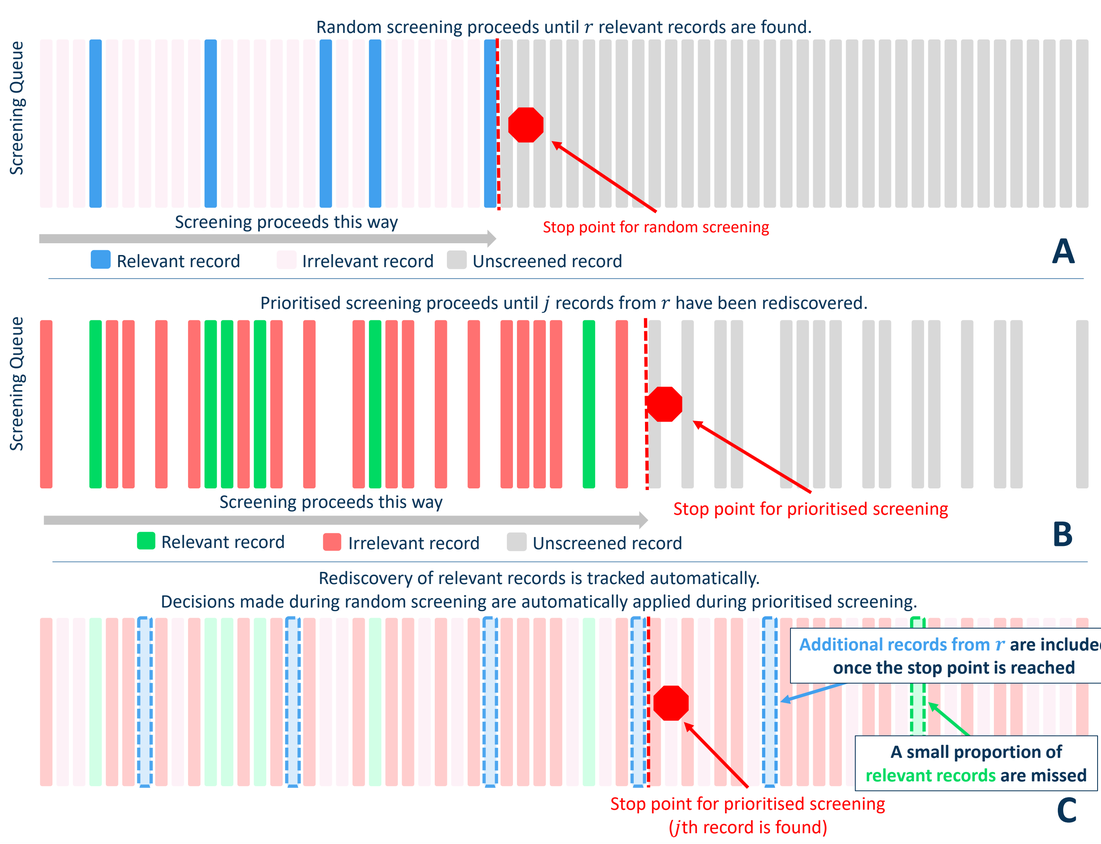
- 95% and 99% recall thresholds (t) were examined, at two different conventional levels of certainty ($1-\alpha$; 95% and 99%). These thresholds place a high demand on recall.
- In addition to these four combinations of t and $(1-\alpha)$, all possible values for r were examined.
- For every dataset of screening decisions and unique combination of t, $(1-\alpha)$, and r, a complete workflow **(Figure 1)** was simulated 2000 times.
- The prioritised screening stage **(Figure 1B)** was simulated using the ASReview Python package.[2] Default settings for the prioritised screening algorithm were used.
- Datasets of T&A screening decisions were gold standard: blind dual-screening with conflicts resolved via discussion **(Table 1).** Several review types were examined (economic evaluation, health care resource utilisation, randomised controlled trials) across several disease areas.

Table 1. Datasets: T&A screening decisions

Dataset ID	N	R	$\frac{R}{N}$ (%)
rheumatology1	1497	128	8.6%
nephrology1	2996	569	19.0%
nephrology2	1267	291	23.0%
dermatology1	468	118	25.2%
dermatology2	460	78	17.0%
dermatology3	3801	989	26.0%

Abbreviations: N, total number of records in dataset; R, total number of relevant records in dataset

Figure 1. Stopping rule workflow



Intuition for stopping rule: At any point during screening, the proportion of relevant records retrieved from a starting random sample of relevant records is indicative of the proportion of all relevant records that have been retrieved.

Random screening phase (Figure 1A)

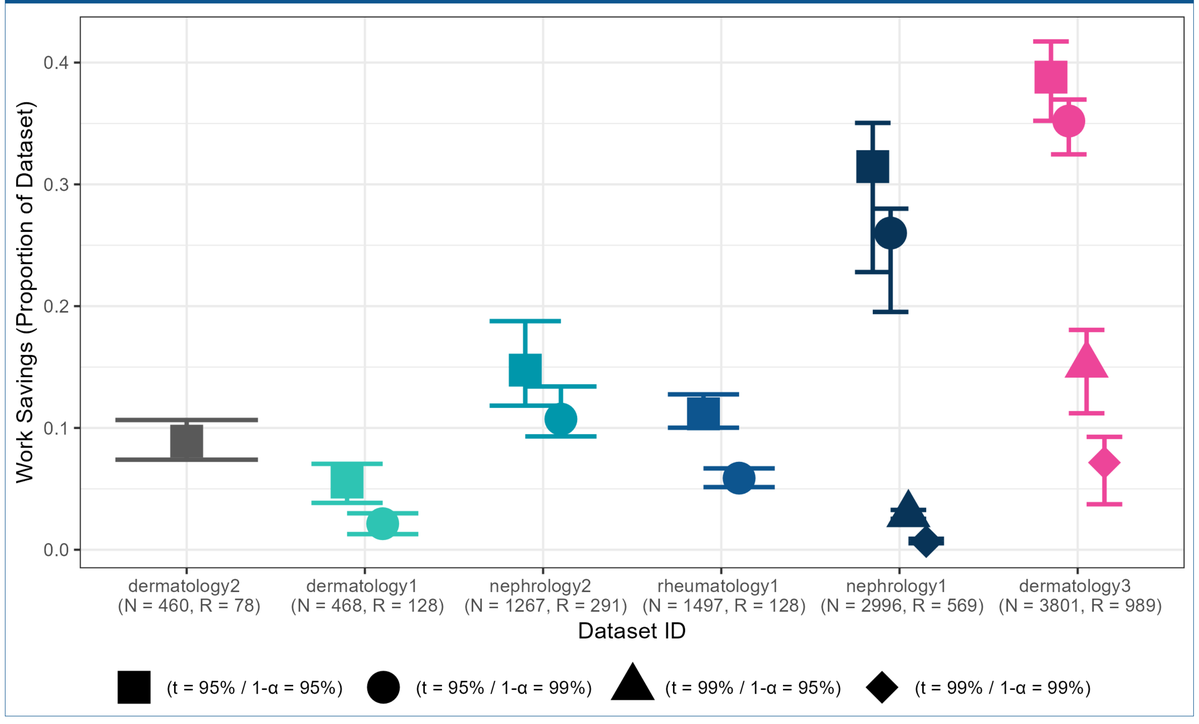
- The review manager begins by setting the following parameters:
- (t) – the target recall (desired percentage of relevant studies retrieved when the stopping rule is triggered).
- $(1-\alpha)$ – the proportion of cases where the target recall is met when the stopping rule is triggered. The target recall is not met in α% of cases.
- (r) – the number of relevant records that must be drawn during random sampling phase before proceeding to the prioritised screening phase
- Records are screened randomly until r relevant records are retrieved. After noting the relevant records in this set, this sample is 'released' back into the screening queue, which is then shuffled.

Prioritised screening phase (Figure 1B)

- Reviewers begin screening the dataset. They must not use any information gained during the initial random screening phase to decide the order in which records will be screened.
- Records screened during the initial random screening phase are not manually rescreened when they reappear. Instead, screening decisions are stored and automatically applied when these records reappear at the top of the screening queue during prioritised screening (**Figure 1C**).
- Screening continues until a threshold of j records from the initial random sample of relevant records are found. The threshold is then calculated using the stopping rule equations. They determine the number of records j that must be retrieved from the starting sample r to guarantee that t has been met, with $(1-\alpha)$ confidence.

Results

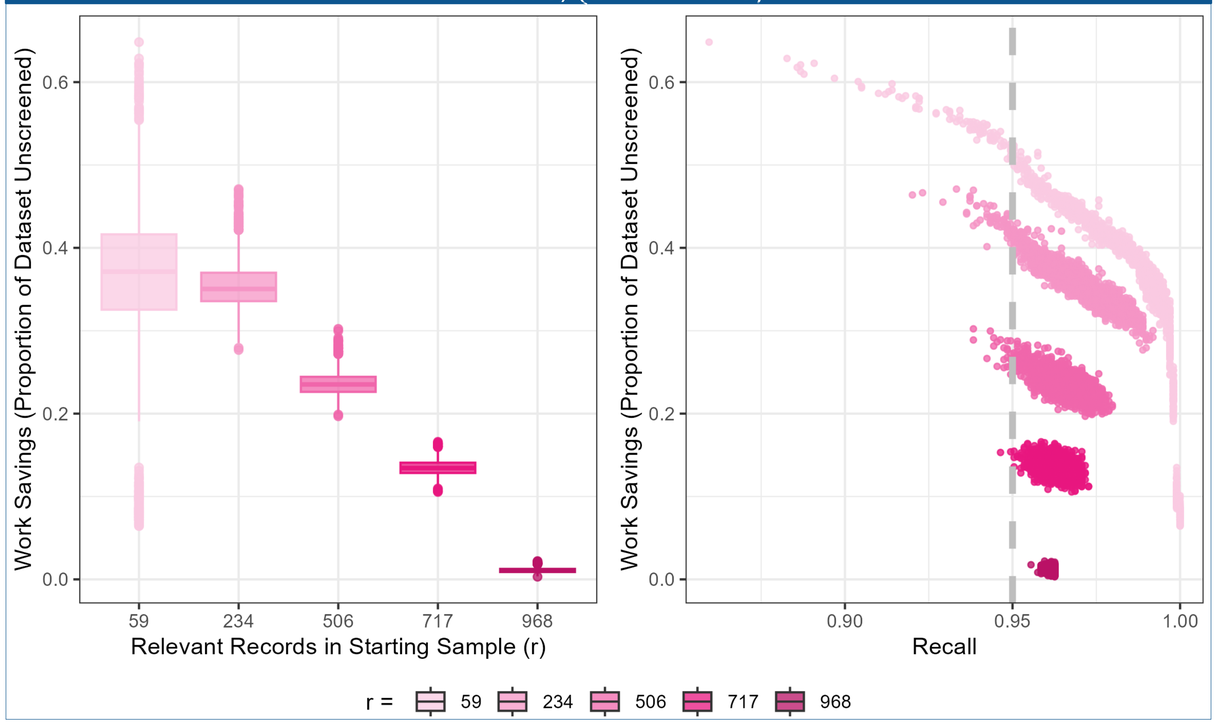
Figure 2. Median (IQR) work savings by dataset



Abbreviations: IQR, interquartile range; N, total # of records in dataset; R, total # of relevant records in dataset; t, target recall; $1-\alpha$, proportion of cases where true recall $\geq t$; r, number of relevant records drawn during the random sampling phase.

**Ordered from smallest to largest dataset (by N). Best-case results were selected from the specific value of r that produced the highest median work savings for each dataset and t, $(1-\alpha)$ combination. Not all datasets (e.g. dermatology1, dermatology2) contained enough relevant records to implement more stringent stopping rule configurations (e.g. where $t = 99\%$). In these cases, sampling to reach the required r would have continued until the entire dataset was screened and no work savings would have occurred.*

Figure 3. Relationship between r, work savings, and recall: dermatology3 dataset ($N = 3801$, $t = 0.95$, $(1-\alpha = 0.95)$



Abbreviations: r, the number of relevant records drawn during random sampling phase

**The work savings axis is identical in the plots on the left and right. In the plot on the left, the distribution of work savings for 5 different values of r are plotted as boxplots. The same 5 values of r are plotted in the plot on the right, linked by colour. In the plot on the right, the recall and work savings for each unique simulation (of 2000) are plotted for each of the 5 distinct values of r examined. The vertical grey line in the plot on the right represents the target recall (t) of 95%.*

Key findings: work savings and prioritised screening performance

- For most datasets (except dermatology1), prioritised screening was effective: 95% of relevant records were discovered after 50-75% of a dataset was screened.
- For lower recall targets ($t = 95\%$), median work savings were $> 25\%$ of N for the largest datasets but negligible for the smallest datasets. For higher recall targets ($t = 99\%$), work savings were minor even for large datasets **(Figure 2)**.

Key findings: relationship between r, work savings, and recall

- Smaller r values generated more work savings but higher variance in recall **(Figure 3)**. For these values of r, the highest work savings often fell well below the recall threshold.
- Larger values of r generated fewer work savings but were more reliable. In simulations where the true recall fell below the target recall, the discrepancy between target recall and true recall was far less in simulations that used larger r values.

Conclusions

Summary of findings and implications

- The stopping rule substantially reduced screening burden while maintaining high recall in several large datasets ($N = 2000+$; $R = 250+$). If prioritised screening performance is comparable, the stopping rule would likely reduce screening burden for other large datasets.
- Before using this workflow for literature reviews, researchers need to understand when excluding some relevant records during T&A screening would materially impact review quality.
- In many cases, reviews do not need to cover all available evidence (e.g. to quickly get a sense for a research area). Here, the proposed workflow is immediately usable.
- When reviews must identify all available evidence, implementing the proposed workflow is only feasible if reviewers know in advance that relevant records excluded during T&A screening will ultimately be excluded during full-text review. Additional research is required to determine when, if at all, this can be expected.

Limitations

- Work savings depend on the performance of the prioritised screening algorithm and may not generalise to other datasets.
- The importance of relevant records that are captured or lost is not considered.
- Findings from the examined workflow may not generalise to dual-screening workflows.
- Error rates in single human screening were not accounted for when considering recall at the stopping point.
- Recall thresholds are inappropriate for reviews that require retrieval of all relevant records.

References: *(1) Lewis DD, Yang E, Frieder O. Certifying one-phase technology-assisted reviews, In Proceedings of the 30th ACM international conference on information & knowledge management, 2021. (2) ASReview LAB developers. (2023). ASReview LAB - A tool for AI-assisted systematic reviews (v1.3rc1). Zenodo. https://doi.org/10.5281/zenodo.8297019*

Declaration of funding: *This project has been funded in full by Symmetron Limited.*