Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEACH ARTICLE

Don’t stop me now, cause I’m having a good time screening: Evaluation of stopping methods for safe use of priority screening in systematic maps and reviews

Tim Repke[1,3] | Francesca Tinsdeall[2] | Diana Danilenko[1] | Sergio Graziosi[3] | Finn Müller-Hansen[1] | Lena Schmidt[3,5] | James Thomas[3] | Gert van Valkenhoef[4]

[1]Climate Economics and Policy, Potsdam Institute for Climate Impact Research (PIK), Brandenburg, Germany

[2]Centre for Clinical Brain Sciences, University of Edinburgh, United Kingdom

[3]Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI), University College London (UCL), United Kingdom

[4]IT development and infrastructure, Cochrane Central Executive, United Kingdom

[5]NIHR Innovation Observatory, Newcastle University, United Kingdom

Correspondence

Corresponding author is Tim Repke.

Email: tim.repke@pik-potsdam.de

Funding Information

This research was supported by the Wellcome Trust (PATHFINDER2, grant no. 313586/Z/24/Z and DESTinY, grant no. 227165/Z/23/Z).

Abstract

Introduction

Priority screening has the potential to reduce the number of records that need to be annotated in systematic literature reviews. So-called technology-assisted reviews (TAR) use machine-learning with prior include/exclude annotations to continuously rank unseen records by their predicted relevance to find relevant records earlier. In this article, we present a systematic evaluation of methods to determine when it is safe to stop screening when using prioritisation.

Methods

We implement an open-source evaluation framework that features a novel method to generate rankings and simulate priority screening processes for 86 real-world datasets. We use these simulations to evaluate 12 statistical or rule-based (heuristic) stopping methods, testing a range of hyper-parameters for each.

Results

The work-saving potential and performance of stopping criteria heavily rely on ‘good’ rankings, which is typically not achieved by a single ranking algorithm across the entire screening process. Our evaluation shows that the existing stopping methods either fail to reliably stop without missing relevant records or fail to utilise the full potential work-savings. Only one method reliably met the set recall target but stops conservatively.

Conclusions

Many digital evidence synthesis tools provide priority screening features that are already used in many research projects. However, the theoretical work-savings demonstrated in retrospective simulations of prioritisation can only be unlocked with safe and reproducible stopping criteria. Our results highlight the need for improved stopping methods and guidelines on how to responsibly use priority screening. We also urge screening platforms to provide indicators and authors to transparently report metrics when automating (parts of) their synthesis.

KEYWORDS

stopping methods, systematic maps, systematic reviews, priority screening, technology-assisted reviews, digital evidence synthesis

1 | INTRODUCTION

The published scientific literature is is growing at an impressive rate[1,2,3]. Although additional evidence may be good overall, it poses a considerable challenge for evidence synthesists, decision-makers, and other users of evidence[4] as exhaustive identification of relevant evidence an essential requirement most methodologies. The sheer volume prohibits the use of conventional systematic map and review methods and requires (partial) automation[5,6].

In this article, we focus on priority screening which is growing in popularity and has large work-saving potentials[7,8]. Priority screening supports the process of deciding whether abstracts that were found by a database search are actually relevant for the current study. Therefore, unseen records are continuously ranked to always show potentially most relevant records next using machine-learning models trained on prior include/exclude annotations. However, this approach can only save work if the ranking is good and if there is a reliable criterion to decide when to stop screening because all relevant records are already found[9]. Furthermore, we need to ensure that those stopping methods are safe and robust enough to use responsibly while also being able to quantify remaining uncertainties.

To this end, we present a empirical evaluation of existing stopping methods on screening decisions from 86 systematic reviews to determine i) which stopping methods work and under which conditions, ii) how stopping methods are influenced by the priority ranking of unseen records, and iii) how the choice of user-defined hyper-parameters impact the performance of stopping methods.

There are already evaluation frameworks to (partially) address some of these research questions. For example, both, Yang et al. and Bron et al., published python libraries that each implement five stopping methods and abstract classification pipelines for ranking[10,11]. Li et al. published a Python library as part of their proposed stopping framework that implements six stopping methods and has three ranking algorithms (BM25, logistic regression, SVM) as well as CLEF and TREC datasets built-in for evaluation[12].

However, these analyses are often part of the proposal of a new stopping method, use synthetic data, improperly apply competing stopping methods, or make unrealistic assumptions. Reviewing the literature, there appears to be no consensus on which stopping methods are reliable and how to validate that. Priority screening is already supported in several major digital evidence synthesis tools such as covidence, EPPI-reviewer[13], NACSOS[14], Abstrackr[15], or Rayyan[16], Only NACSOS implements a statistical stopping criterion for users to track the progress and determine when it is safe to stop. tools Some popular platforms do not even provide the relevant information to enable users to use external tools for computing stopping criteria.

With this work we publish an easily extensible open-source framework that already implements a wide range of stopping methods in such a way that they can be easily re-used elsewhere.. Furthermore, the framework allows anyone effortless evaluations of simulations on their datasets. Here, we present the results from thousands of simulations using a wide range of stopping-method-hyper-parameters and optimised ranking models based on real-world datasets. theOur analysis shows that only one method, BUSCAR, is safe to use but is not using the full work-saving potential. The knee method is safe to use most of the time in our simulations but it highly dependent on the performance of the ranking algorithm and choice of hyper-parameters. We hope to inspire future research to improve stopping methods and for platform developers to integrate guidelines and indicators that empower the safe and responsible use of machine-learning-assisted screening tools.

2 | METHODS

For the evaluation of stopping methods we set up an open-source modular and extensible evaluation framework.[†] Typically, priority screening is a combination of two main mechanisms. The first mechanism uses machine-learning models trained on existing inclusion/exclusion annotations to then rank unseen records in descending order of relevance. In this way, annotators are not screening records at random but get to see the potentially most relevant first. At some point, all relevant records are found and the remaining unseen records do not need to be screened. However, determining this point without complete knowledge is not trivial and the purpose of the second mechanism, the stopping criterion. Stopping criteria are rules that use indicators of the prior screening decisions, prediction scores of the ranking model, or other indicators to determine when it is safe to stop screening.

Figure 1 is a composite of all inclusion curves, also known as the gain curves, that illustrate how many included records are found as more records are screened. In an initial random sampling baseline (blue area) the slope typically follows a diagonal where relevant records are occasionally found. After switching to prioritisation (orange are), most screened records are relevant and the curve is much steeper than before and begins flattens again once fewer relevant records remain in the unscreened set until it eventually converges.

Our framework decouples these mechanisms and simulates the screening process including priority ranking but without stopping. This is a computationally very expensive process and was pre-computed on the high-performance cluster at the Potsdam Institute for Climate Impact Research. In the second step, we test when each stopping criterion determines a safe point to stop and measure how close this is to the theoretically optimal point.

2.1 | Priority ranking

Conventionally, all abstracts need to be screened and the next record to be annotated is chosen at random. Some technology-assisted review tools use the ranking function from the retrieval system that measures

[†] https://github.com/destiny-evidence/stopping-methods

TABLE 1 Overview of stopping methods currently implemented in our evaluation framework and their usage in related work

Method	Type	Hyper-parameters	Proposed or evaluated in
APRIORI	Target-agnostic	recall threshold, window size	10
BATCHPRECISION	Target-agnostic	precision threshold, window size	17,18,19
BUSCAR (CMH with biased urns)	Target-aware, uncertainty-aware	recall target, confidence level, sampling bias	20,21,18,19,22
CURVE_FITTING	Target-aware, uncertainty-aware	recall target, confidence level, windows	23
HEURISTIC_FIX (consecutive excludes)	Target-agnostic	fixed number of consecutive excludes	24,25,26,27,20,28,22,29
HEURISTIC_FRAC (consecutive excludes)	Target-agnostic	fraction of consecutive excludes	30
HEURISTIC_RANDOM	Target-aware, uncertainty-agnostic	recall target	None
HEURISTIC_SCORES	Target-aware, uncertainty-agnostic	recall target	None
KNEE	Target-agnostic	slope ratio, curve threshold	31,32,25,24,33,34,21,35,36,37,38,39,18,22,23
METHOD2399	Target-agnostic	scaling factor	36,18,19
QUANT	Target-aware, uncertainty-agnostic	recall target	18,21,22
QUANT_CI (QUANT with confidence interval)	Target-aware, uncertainty-aware	recall target, confidence level	18,21,22

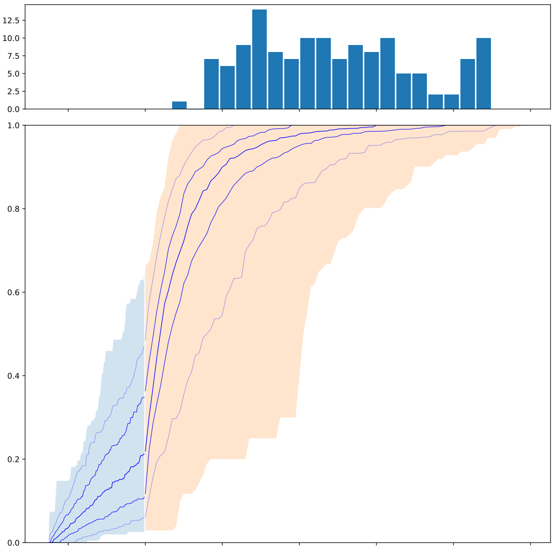

FIGURE 1 Composite of aligned gain curves from all screening simulations. Lines mark the mean and 5th, 25th, 75th, 95th percentiles and the shaded areas mark the full range. The blue area marks the initial random samples and the orange areas the remainder of the datasets. The histogram depicts where a theoretical 99% recall target is reached.

how well an abstract matches the search query. A commonly used function is Okapi BM25[40] and has some parameters that can be tuned once some data was annotated. This approach has the benefit that abstracts are ranked right from the start without the need for training data as the query serves as a prior. Other commonly used ranking models are based on support vector machines (SVMs) or logistic regression models that are trained on sparse vector representations of the abstracts and existing inclusion/exclusion annotations. To the best of our knowledge, transformer models are not yet used in digital evidence synthesis tools or stopping method evaluations and the use of large language models is subject to ongoing research[41,42,43,44].

In regular intervals, the ranking model is re-trained as additional records are annotated. Each synthesis tool offering priority screening handles this differently. Whereas some start training models right away and re-trained at relatively small fixed intervals, others require a larger initial sample of records screened at random and have adaptive batch sizes. Most synthesis tools are not (fully) open-source, but from the limited information available there appear to be no adaptive model selection strategies or tuning.

The mechanism for pre-computing ranking simulations uses a fixed initial set of 500 randomly sampled abstracts[‡] after which the model is re-trained in adaptive intervals. The target batch size is within a lower (25 records) and upper bound (200 records), whereas the target starts at the lower bound and is growing at a rate of 10% after each iteration. The actual batch size might be smaller or larger within the bounds if at least two new included records are found. In this way we can reduce the computational overhead by only re-training when it is actually sensible and sufficient amount of additional training data is available.

We implemented transformer-based models and four basic ranking models (logistic regression, support vector machine, gradient descent

[‡] This value was determined experimentally, see discussion for details.

learning, and light gradient-boosting machine) based on sparse TF-IDF vector representations. Each model can be used with pre-determined hyper-parameters or the hyper-parameters can be optimised with Optuna[45]. Early experiments have shown that hyper-parameters should be adjusted as the available set of training data is growing. Furthermore, we found that the rankings can be drastically improved by choosing the best kind of model at each re-training iteration. To this end, we use an ensemble of models with optimised hyper-parameters and use prediction scores of the best performing model to rank unseen documents. Early tests have shown that the setup does not change with every re-training step, so we re-used the same model and hyper-parameters for four iteration. We pre-compute three priority screening simulations per dataset with a different initial random sample and different random seeds. Figure 1 shows a composite of all simulations.

2.2 | Stopping methods

The selection of stopping methods implemented for this evaluation as listed in Table 1 is informed by an ongoing systematic review[46]. Several methods are actually just variations one another and may appear with different names in the literature. In general, stopping criteria can be categorised as *target-agnostic* if the decision to stop is not based on reaching a pre-defined metric such as recall or precision or *target-aware* otherwise. A *target-aware* criterion is also *uncertainty-aware* if it offers a measure of statistical uncertainty about reaching the set target, for example using confidence intervals or variance of an estimated recall or formal hypothesis testing procedures. Conversely, *uncertainty-agnostic* stopping criteria solely base the decision to stop on the point estimate of the performance metric compared to the pre-defined target. For our evaluation, we chose representative methods from each category. Where possible, we adapted existing implementations but also rewrote and adjusted some aspects of them to make them work more reliably. We iteratively apply each stopping method with a range of different hyper-parameters in fixed regular intervals of 15 records on the pre-computed ranked screening simulations. Note that these batch sizes are independent of the re-training batches from the screening simulation. Once a stopping criterion was fulfilled, we consider the end of the first batch when as the stopping point for that method and that hyper-parameter combination.

In the following, we explain the basic underlying principles for each of the included stopping methods. *APRIORI*[10] measures the recall by comparing annotator ratings and the classifications used in the ranking and stops once the recall is above the set threshold. *BATCHPRECISION*, also known as marginal precision[17,18], stops when the precision between classifier prediction and human annotation in the latest N screened records falls below the set threshold. *BUSCAR*, also referred to as CMH in the literature, is a statistical stopping criterion that calculates the p-score for the null hypothesis, that the set recall target is missed and stops once the desired confidence target is met.[20] We use an extended version of the original criterion with a sampling bias parameter. The *CURVE_FITTING*[23] fits a negative exponential curve to the smoothed inclusion curve to estimate the total number of relevant records and derives a recall estimate from that. *HEURISTIC_RANDOM* uses the inclusion rate in the initial random sample to extrapolate the overall number of relevant records to be able to stop once an estimated recall target is reached. *HEURISTIC_SCORES*, also known as *Quant*[18], use the ranking model scores by estimating the overall number of included records and stopping at a set recall target. *QUANT_CI*[18] extends this idea by adjusting this estimation by using the variance of model scores to derive a confidence interval. *HEURISTIC_FIXED*[24] and *HEURISTIC_FRAC*[30] are commonly used methods to stop screening once a fixed number (or proportion of the dataset) of consecutive excludes have been observed under the assumption that this indicates the convergence of the inclusion curve. The *KNEE* method[31] is based on kneedle[32] that identifies the point where the inclusion curve begins to flatten and stops when the ratio of the slope before and after that point is above a certain threshold. In the original implementation this worked very unreliably, so we changed the curve smoothing mechanism by fitting configurable polynomials and introducing a minimum threshold for the distance between the inclusion curve and the diagonal connecting the start- and endpoint. The *2399 METHOD*[31] is triggered once the number of screened records is above at least 2,399 excluding the number of included records found so far adjusted by a factor.

We do not include the target method[31] which assumes prior knowledge of some relevant records to hide within the screening process to estimate when a pre-determined recall target is met. Other variations of the above-mentioned are also not included in this analysis but will be contained in the open-source framework and an interactive companion website as part of future work and the ongoing review of stopping methods[46]. Additional methods, such as RLStop[38], point-based processes[47], Chao's estimator[22], and confidence sequences[48], needed additional validation and will be added during the revision of this article.

2.3 | Datasets

Our evaluation is based on 86 real-world datasets of fully manually screened abstracts for systematic reviews across various research areas that are commonly used to evaluate technology-assisted review methods. We selected those from commonly used collections of systematic review annotations, namely from CLEF-TAR 2017–2019[49,50,51] (reviews in empirical medicine), TREC 2010, 2015, 2016 tracks on total recall and legal[52,53,54], the SYNERGY collection (reviews in psychology, medicine, biology, computer science, and maths) which extends the Cohen collection[55,56], and CSMeD[57] which extends some of the previously mentioned collections and others[58,59]. Additionally, we use unpublished datasets that were kindly provided by users of the EPPI-reviewer platform.

For smaller datasets the benefits of safe stopping methods at higher recall targets with good certainty are minor and simulations on low inclusion rates would mostly depend on the capability of the ranking method which is beyond the scope of this work. To this end, we excluded many datasets because of their size (only using datasets with at least 1,000

TABLE 2 The proportion of relevant records missed, actual recall at stopping (with percentile range), and additional work as a fraction of the dataset for each stopping method and target recall (if applicable).

Method	Target	% missed	Recall	Overshoot
			[mean (10–90q)]	(mean)
APRIORI	0.80	16.79	0.9 (0.45–1.0)	0.47
APRIORI	0.90	5.11	0.98 (1.0–1.0)	0.54
APRIORI	0.95	0.00	1.0 (1.0–1.0)	0.51
APRIORI	0.99	0.00	1.0 (1.0–1.0)	0.37
BATCHP.	–	–	0.78 (0.2–1.0)	–
BUSCAR	0.80	0.00	0.98 (0.96–1.0)	0.26
BUSCAR	0.90	0.00	0.99 (0.98–1.0)	0.32
BUSCAR	0.95	0.00	1.0 (0.99–1.0)	0.36
BUSCAR	0.99	0.00	1.0 (1.0–1.0)	0.34
CURVE.	0.80	23.85	0.83 (0.18–1.0)	0.37
CURVE.	0.90	18.58	0.87 (0.26–1.0)	0.43
CURVE.	0.95	15.83	0.89 (0.34–1.0)	0.39
CURVE.	0.99	14.68	0.9 (0.41–1.0)	0.26
H._FIX	–	–	0.3 (0.0–1.0)	–
H._FRAC	–	–	0.3 (0.0–1.0)	–
H._SCORES	0.80	17.52	0.92 (0.66–1.0)	0.49
H._SCORES	0.90	17.52	0.93 (0.69–1.0)	0.45
H._SCORES	0.95	17.52	0.94 (0.74–1.0)	0.39
H._SCORES	0.99	17.52	0.95 (0.76–1.0)	0.26
KNEE	–	–	0.98 (0.97–1.0)	–
M.2399	–	–	0.99 (0.97–1.0)	–
QUANT_CI	0.80	3.16	0.96 (0.89–1.0)	0.24
QUANT_CI	0.90	2.68	0.98 (0.95–1.0)	0.33
QUANT_CI	0.95	3.89	0.99 (0.98–1.0)	0.37
QUANT_CI	0.99	5.84	1.0 (1.0–1.0)	0.31

records) or their extremely low inclusion rate (below 1%). Of the 93 remaining, we excluded another seven datasets where we were not able to retrieve at least 90% of the abstracts through OpenAlex, the Web of Science, Scopus, dimensions.ai, or PubMed.

3 | RESULTS

The following analysis of stopping method performance is based on 258 simulated prioritised screening runs for 86 real-world datasets of systematic reviews or other relevance decisions. On average, datasets contain 3,600 records (1,000–13,095; STD=2,570) with on average 167 (5%) relevant records (16–1,957; STD=240 or 0.8%–37%; STD=5%). In total, we recorded 3.2M stopping decisions across 13,773 combinations of dataset, repeated screening simulations, stopping criteria, and hyper-parameters.

Figure 1 shows a composite of all inclusion curves to give an indication of the quality of rankings. Ideally, a ranking model would have a slope of one (every screened record is included) after the initialisation sample until all relevant records are found and then sharply transition to a zero-slope. Early experiments with simple machine-learning models have produced curves that only continue along the slope of random sampling which does not allow for any early stopping. Using our continuous optimised ensemble approach, almost all simulations offer potential work savings when paired with an optimal stopping criterion. In the first quartile, only 32% of the dataset needed to be screened before reaching the 99% recall target with 49% in the second and 66% in the third quartile. Additional statistics at different recall targets are included in the supplemental information (Figure A1). Another observation from early experiments is that the curves often have intermediate plateaus where no new relevant records are discovered for a while. Most stopping methods detect such 'steps' as a false signal and stop too early. It is our hypothesis that this pattern is caused by the ranking model being stuck in a local minimum and that the inclusion curve is temporarily growing again as a new cluster of relevant records is found. Increasing the size of the initial random sample appears to be a good strategy to mitigate this stepping pattern.

Common metrics in the literature to measure the performance of a stopping method are variations of the recall measured at the stopping point and the work saved compared to screening the complete dataset. In Figure 2, we provide an overview of all results, whereas each dot represents a decision to stop screening by a method with a given set of hyper-parameters for a screening simulation. We use colours and marker symbols to denote results for different recall targets where available. For methods without a target recall parameter, we group results by the most influential hyper-parameter and replicate the result with different recall targets as the theoretically optimal stopping point as a reference. The x-axis is split in two halves. The left half of the lower axis shows the proportion of the records that should have been included for the given recall target but were missed due to stopping too early. The right side shows what proportion of records after the optimal stopping point for a given recall target were screened until the method decided it is safe to stop, where 0% would coincide with the optimal point and 100% means the criterion was never fulfilled. This same metric is shown on the y-axis as absolute values. Additional in-depth results per method and tables listing the values of the distributions shown are available in the supplementary information.

All stopping methods, except for BUSCAR, stop too early in many occasions. Particularly heuristic methods that use the number of consecutive excludes tend to miss most of the relevant records. Conversely, CURVE_FITTING, APRIORI, and recall estimation (HEURISTIC_SCORES) stop far too late or never while still showing several cases where they stop before reaching a specific recall. Due to the nature of Method2399, it always screens 37 of our datasets completely as they contain fewer than 2399 records but also never stops in more cases than that. BATCHPRECISION and KNEE have no clear point where they stop far too early or late and are spread across the entire range independent of the hyper-parameter settings. They do however tend to stop too early more often the larger the dataset is. The estimation of the number of included documents using the inclusion rate of the initial random sample in HEURISTIC_RANDOM tends to under-estimate the recall target and misses records in half of all cases or almost never stops otherwise. These results indicate that these methods cannot be used to reliably determine when it is safe to stop prioritised screening.

The QUANT method uses ranking model scores to estimate a recall similar to HEURISTIC_SCORES, but also has variations to compute a

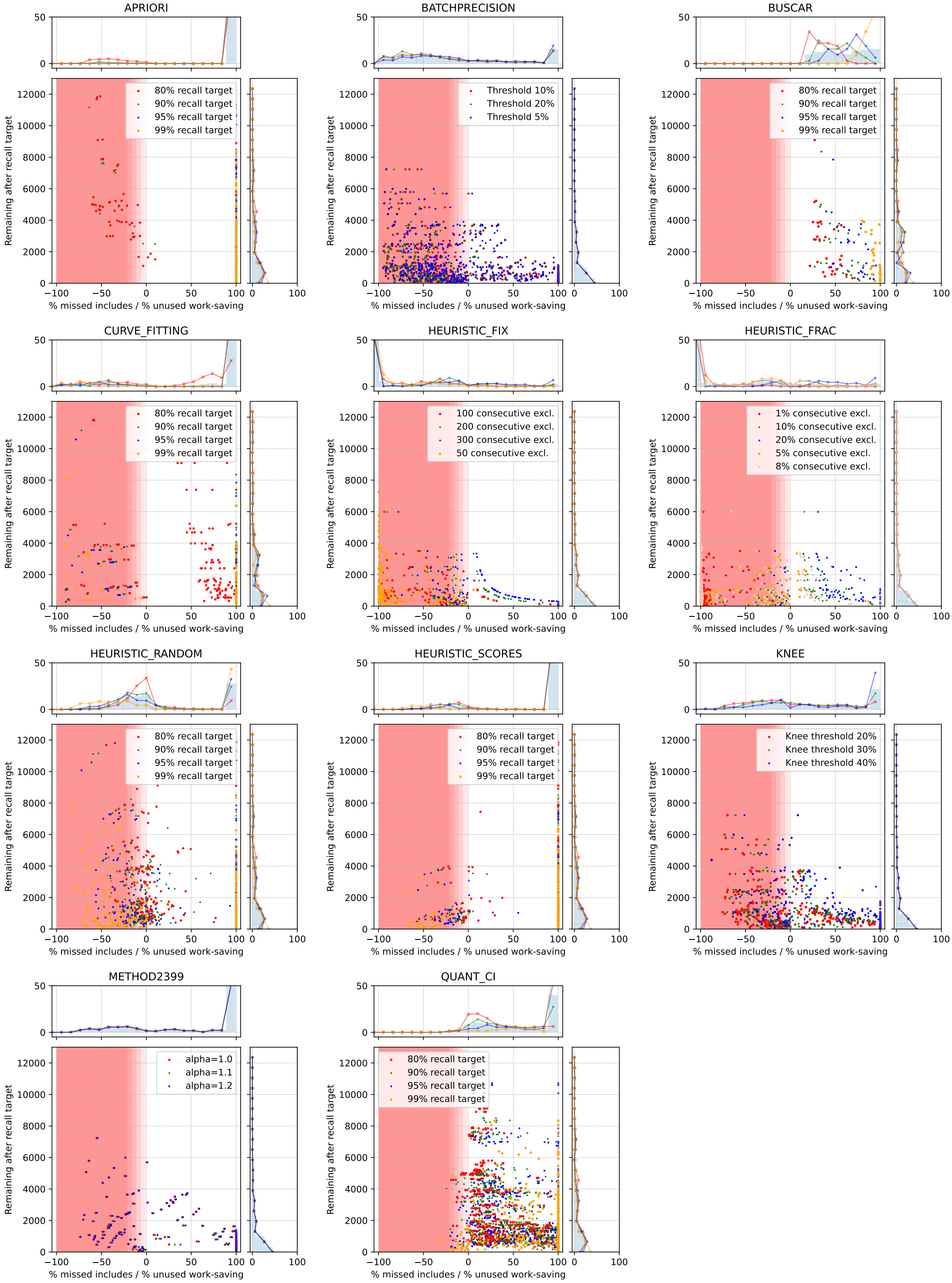

FIGURE 2 Trade-off between work saved and missed relevant records by stopping early. The ideal stopping method would have all points along a vertical line at x=0, where no relevant records were missed and the stopping rule invoked at the theoretically perfect time with no additional work. Results are grouped by a dominant hyper-parameter or recall target where available. The histograms show the distribution of points along each axis. Histogram lines correspond to hyper-parameter groups, whereas bars show the overall distribution.

confidence interval around the estimate. Doing so makes this method more conservative and overshoots the recall target by less than 20% in 26% of the cases and by less than 50% in 44% of the cases. For about 40% of the cases, more than 95% of the records after the theoretically optimal stopping point were screened, which is more pronounced for higher recall targets. In about 4% of cases, up to 50% of relevant records were missed by stopping too early.

Only one method, BUSCAR, never stops before the set recall target.Overall, it tends to be more conservative for higher recall targets and less for larger datasets, thus foregoing large shares of potential work savings in these settings. This method overshoots by less than 20% in 9% of the cases and in less than 50% in 46% of the cases. By adjusting the assumptions about the underlying sampling distribution using biased urns, the method can be tuned to get closer to the theoretical recall target.

4 | DISCUSSION

In this article, we analysed the performance of a wide range of stopping methods on their ability to reliably determine the optimal point to stop screening in a ranked set of abstracts. The results show that almost all stopping methods cannot be used without a high risk of missing relevant records and no method is able to reliably achieve optimal work savings that are theoretically possible for a given recall target. This highlights a trade-off between having relative certainty about achieving a recall target and optimal work savings.

When evaluating the utility of stopping methods, we argue that a method is unreliable and cannot be used safely when the functionality depends on fine-tuning hyper-parameters or on an optimal ranking. In a real-world screening scenario, reviewers do not have perfect knowledge to retrospectively choose the best hyper-parameters or distinguish a ranking an imperfect ranking from actual convergence that would indicate that all relevant records were found and it is safe to stop annotating abstracts. To this end, we focus less on reporting aggregated performance metrics but present the entire distribution of results across a wide range of datasets, repeated ranking simulations, and hyper-parameter settings. In the same way, we argue that metrics such as work-saved-over-sampling can be misleading when judging the effectiveness of stopping methods, as this is also influenced by the ability of the ranking method to find relevant records based on prior include/exclude decisions. Therefore, we focus on the deviation from the theoretically optimal stopping point on a given inclusion curve. However, this requires setting recall targets for target-unaware methods as well, with the choice of these targets impacting our evaluation metrics.

To align results from such a wide range of datasets means that we have to choose a common frame of reference. This has the side-effect that some implications are hidden. For example, the interpretation of a reported percentage of 50% unused work-saving-potential is vastly different if the recall target is reached after screening 10% or 90% of a dataset. Particularly for larger datasets this can substantively impact the person-hours required to complete a systematic evidence synthesis.

As discussed above, we found that the choice of ranking models strongly influenced the potential to save work. Furthermore, even if a ranking method is able to find all relevant records early on in the process, the shape of the inclusion curve towards that point can have strong impacts on the performance of stopping methods. Particularly plateaus from local minima are often wrongly interpreted as a signal to stop, and some methods use window sizes or other smoothing strategies to mitigate this. The best way to mitigate is to screen a large enough random sample in the beginning. We chose a fixed size for the initial sample based on anecdotal evidence. Future work is required to systematically explore adaptive mechanisms to reduce that size based on dataset characteristics to switch to prioritised screening as late as needed but as soon as possible. The introduced uncertainties by stopping early and the apparent need for a large enough training dataset to train a good ranking model imply that priority screening with early stopping should only be used in larger reviews. For this analysis we chose 1,000 to be the lower bound.

Using classifier scores to rank unseen records may pose issues that require further research. In doing so, we need to make the assumption that slight differences in scores actually carry some meaning with regard to how relevant a record is compared to another one given prior screening decisions. However, the scores of classification models might describe the distance to a decision boundary, feature similarity to previously seen records, or others. The score distribution of some models is also tightly grouped around zero (or negative one) and one and allow no true fine-grained ranking. If all scores only vary in the seventh decimal and are close to zero as the inclusion curve starts to converge, it is not clear how that ranking actually differs from random sampling at that point. Similarly, after each re-training of the model as the inclusion curve converges, there always tend to be a few records with higher scores that reveal a few more relevant records that might not have been found at that point without re-training. This suggests that the frequency of re-training might also have an impact on the utility of a ranking to determine a safe point to stop screening.

Stopping methods cannot be used without a clear and reliable guide on how to choose hyper-parameters in an ongoing priority screening project without full knowledge. Based on our wide range of tested hyper-parameters, we were not able to derive reproducible rules-of-thumb to determine safe and reliable settings. With increasing number of technology-assisted reviews that use automation in some or all steps of the process, uncertainties have a compounding effect on the final outcomes. For example, an imperfect screening recall implies some records might be missed and their reported effect sizes will not be be included in the final meta-analysis. Being able to estimate how many records are affected allows the authors if the analysis to adjust the reported confidence interval. Researchers developing stopping methods should therefore focus on uncertainty-aware methods to enable an estimation of those compounding effects. Furthermore, the wider research community and users of evidence need to form a common understanding of which uncertainties are acceptable in which context.

5 | CONCLUSION

Safe use of AI in evidence synthesis projects requires rigorous evaluation to ensure robust results. Priority screening is a widely adopted method to reduce time for identifying relevant articles for a systematic map or review. However, it is crucial to combine machine-learning based ranking with statistically sound stopping methods.

The modular architecture of our open-source evaluation framework makes it easy to add stopping methods without tightly integrating them into framework utilities so they can also easily be used elsewhere by tool developers or users of digital evidence synthesis tools. The framework serves as a good foundation for future benchmarking across a growing set of reference datasets, stopping methods, and ranking models. At the moment, some existing stopping methods are not yet part of the framework and adding larger fully-screened datasets from a wide range of different research areas for evaluation would be desirable.

Our empirical evaluation on 86 real-world datasets reveals that none of the existing stopping methods are fit-for-purpose. Only one method has is able to never miss any relevant records at the expense of stopping too conservatively. Another method has a similar overshoot characteristic but also misses up to half the relevant records in a few cases. Future work is needed to develop safe and robust stopping methods that can reliably save work by stopping early at a set confidence level.

AUTHOR CONTRIBUTIONS

Tim Repke: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing (original draft), Writing (review & editing). **Francesca Tinsdeall:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing (original draft), Writing (review & editing). **Diana Danilenko:** Conceptualization, Data curation, Investigation, Methodology, Validation, Writing (review & editing). **Ella Flemyng:** Supervision, Writing (review & editing). **Sergio Graziosi:** Data curation, Investigation, Resources, Validation, Writing (review & editing). **Finn Müller-Hansen:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Writing (review & editing). **Lena Schmidt:** Conceptualization, Investigation, Methodology, Writing (review & editing). **James Thomas:** Data curation, Funding acquisition, Resources, Writing (review & editing). **Gert van Valkenhoef:** Formal analysis, Writing (review & editing).

ACKNOWLEDGMENTS

DD, JT, LS, SG, and TR are supported by the Wellcome Trust (PATHFINDER2, grant no. 313586/Z/24/Z and DESTinY, grant no. 227165/Z/23/Z). FMH acknowledges funding from the GENIE project by the European Research Council under the Horizon 2020 programme (grant agreement no. 951542). The authors gratefully acknowledge the Ministry of Research, Science and Culture (MWFK) of Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research (Grant No. 22-Z105-05/002/001).

ETHICS STATEMENT

The authors have nothing to report.

CONSENT

The authors have nothing to report.

CONFLICTS OF INTEREST

Tim Repke, Sergio Graziosi, and James Thomas are developers of popular screening and review platforms (NACSOS, EPPI-reviewer). Otherwise, the authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The code and raw simulation data is available at https://doi.org/zenodo/will/be/added/later and on GitHub at https://github.com/destiny-evidence/stopping-methods. Unfortunately, we are not able to share the fully annotated data including abstracts due to copyright restrictions.

REFERENCES

1. Minx JC, Callaghan M, Lamb WF, Garard J, Edenhofer O. Learning about climate change solutions in the IPCC and beyond. *Environmental Science & Policy.* 2017;77:252–259.
2. Park M, Leahey E, Funk RJ. Papers and patents are becoming less disruptive over time. *Nature.* 2023;613(7942):138–144.
3. Chu JS, Evans JA. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences.* 2021;118(41):e2021636118.
4. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and Meta-analyses. *Research Synthesis Methods.* 2019;10(3):330–342. doi: 10.1002/jrsm.1354
5. Schmidt L, Cree I, Campbell F, group WEM. Digital Tools to Support the Systematic Review Process: An Introduction. *Journal of evaluation in clinical practice.* 2025;31(3):e70100.
6. Bond M, Khosravi H, Bergdahl N, et al. Digital evidence synthesis tools in educational technology research: A systematic mapping review. *Pre-print,.* 2024:1–26.
7. Kusa W, Lipani A, Knoth P, Hanbury A. An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. *Intelligent Systems with Applications.* 2023;18:200193. doi: https://doi.org/10.1016/j.iswa.2023.200193
8. Rose CJ, Meneses-Echavez JF, Muller AE, et al. Artificial Intelligence and Machine Learning to Improve Evidence Synthesis Production Efficiency: An Observational Study of Resource Use and Time-to-Completion. *Cochrane Evidence Synthesis and Methods.* 2025;3(3):e70030.
9. Callaghan M, Müller-Hansen F, Bond M, et al. Computer-assisted screening in systematic evidence synthesis requires robust and well-evaluated stopping criteria. *Systematic Reviews.* 2024;13(1):284.
10. Bron M. Python Package python-allib. https://doi.org/10.5281/zenodo.10869682; 2024
11. Yang E, Lewis DD, Frieder O. On Minimizing Cost in Legal Document Review Workflows. In: ACM. 2021.
12. Li D, Kanoulas E. When to Stop Reviewing in Technology-Assisted Reviews: Sampling from an Adaptive Distribution to Estimate Residual Relevant Documents. *ACM Trans. Inf. Syst..* 2020;38(4). doi: 10.1145/3411755
13. Thomas J, Graziosi S, Brunton J, et al. EPPI-reviewer: advanced software for systematic reviews, maps and evidence synthesis. 2022. *Google Scholar Google Scholar Reference.* 2022;1.

14. Repke T, Callaghan M. NACSOS-nexus: NLP Assisted Classification, Synthesis and Online Screening with New and EXtended Usage Scenarios. *arXiv preprint arXiv:2405.04621*. 2024.
15. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In: IHI '12. ACM. Association for Computing Machinery 2012; New York, NY, USA:819–824
16. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for Systematic Reviews. *Systematic Reviews*. 2016;5(1). doi: 10.1186/s13643-016-0384-4
17. Cormack GV, Grossman MR. Multi-faceted recall of continuous active learning for technology-assisted review. In: ACM. 2015:763–766.
18. Yang E, Lewis DD, Frieder O. Heuristic stopping rules for technology-assisted review. In: ACM. 2021:1–10.
19. Yang E. *Cost Reduction and Modeling of Technology-Assisted Review*. Georgetown University, 2021.
20. Callaghan MW, Müller-Hansen F. Statistical stopping criteria for automated screening in systematic reviews. *Systematic Reviews*. 2020;9:1–14.
21. Molinari A, Esuli A. SALτ: efficiently stopping TAR by improving priors estimates. *Data Mining and Knowledge Discovery*. 2024;38(2):535–568.
22. Bron M, Der Heijden vPG, Feelders A, Siebes A. Using Chao's estimator as a stopping criterion for technology-assisted review. *ACM Transactions on Information Systems*. 2025;43(3):1–51.
23. Sneyd A, Stevenson M. Modelling stopping criteria for search results using poisson processes. *arXiv preprint arXiv:1909.06239*. 2019.
24. Yu Z, Menzies T. FAST2: An intelligent assistant for finding relevant papers. *Expert Systems with Applications*. 2019;120:57–71.
25. König L, Zitzmann S, Fütterer T, Campos DG, Scherer R, Hecht M. An evaluation of the performance of stopping rules in AI-aided screening for psychological meta-analytical research. *Research Synthesis Methods*. 2024;15(6):1120–1146.
26. Dijk vSH, Brusse-Keizer MG, Bucsán CC, Palen v. dJ, Doggen CJ, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ open*. 2023;13(7):e072254.
27. Oude Wolcherink M, Pouwels X, Dijk vS, Doggen C, Koffijberg H. Can artificial intelligence separate the wheat from the chaff in systematic reviews of health economic articles?. *Expert review of pharmacoeconomics & outcomes research*. 2023;23(9):1049–1056.
28. Boetje J, Schoot v. dR. The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic reviews*. 2024;13(1):81.
29. Quan Y, Tytko T, Hui B. Utilizing ASReview in screening primary studies for meta-research in SLA: A step-by-step tutorial. *Research Methods in Applied Linguistics*. 2024;3(1):100101.
30. Campos DG, Fütterer T, Gfrörer T, et al. Screening smarter, not harder: A comparative analysis of machine learning screening algorithms and heuristic stopping criteria for systematic reviews in educational research. *Educational Psychology Review*. 2024;36(1):19.
31. Cormack GV, Grossman MR. Engineering quality and reliability in technology-assisted review. In: ACM. 2016:75–84.
32. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a kneedle in a haystack: Detecting knee points in system behavior. In: IEEE. 2011:166–171.
33. Piramuthu OB. Multiple choice online algorithms for technology-assisted reviews. In: ACM. 2023:639–645.
34. Bianco GD, Duarte D, Gonçalves MA. Reducing the user labeling effort in effective high recall tasks by fine-tuning active learning. *Journal of Intelligent Information Systems*. 2023;61(2):453–472.
35. Cormack GV, Grossman MR. Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2017.. *CLEF (working notes)*. 2017;11:1–15.
36. Cormack GV, Grossman MR. Waterloo (Cormack) Participation in the TREC 2015 Total Recall Track.. In: NIST. 2015:1–6.
37. Li D, Kanoulas E. When to stop reviewing in technology-assisted reviews: Sampling from an adaptive distribution to estimate residual relevant documents. *ACM Transactions on Information Systems (TOIS)*. 2020;38(4):1–36.
38. Bin-Hezam R, Stevenson M. RLStop: A Reinforcement Learning Stopping Method for TAR. In: ACM. 2024:2604–2608.
39. Sneyd A, Stevenson M. Stopping criteria for technology assisted reviews based on counting processes. In: ACM. 2021:2293–2297.
40. Jones KS, Walker S, Robertson SE. A probabilistic model of information retrieval: development and comparative experiments. *Information processing & management*. 2000;36(6):809–840.
41. Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Systematic reviews*. 2024;13(1):158.
42. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Systematic reviews*. 2024;13(1):219.
43. Wang S, Scells H, Zhuang S, Potthast M, Koopman B, Zuccon G. Zero-shot generative large language models for systematic review screening automation. In: Springer. 2024:403–420.
44. Wang S, Scells H, Koopman B, Potthast M, Zuccon G. Generating natural language queries for more effective systematic review screening prioritisation. In: ACM. 2023:73–83.
45. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: ACM. 2019:2623–2631.
46. Tinsdeall F. Automation to reduce human effort in the document selection stage of systematic evidence syntheses: a systematic review of current approaches.. Protocol; 2024
47. Stevenson M, Bin-Hezam R. Stopping Methods for Technology-assisted Reviews Based on Point Processes. *ACM Transactions on Information Systems*. 2023;42(3):1–37.
48. Lewis DD, Gray L, Noel M. Confidence Sequences for Evaluating One-Phase Technology-Assisted Review. In: ACM. 2023:131–140.
49. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. In: CEUR-WS.org. 2017.
50. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. In: CEUR-WS.org. 2018:1–10.
51. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. In: CEUR-WS.org. 2019:267–274.
52. Roegiest A, Cormack GV, Clarke CL, Grossman MR. TREC 2015 Total Recall Track Overview.. In: CEUR-WS.org. 2015.
53. Grossman MR, Cormack GV, Roegiest A. TREC 2016 Total Recall Track Overview.. In: CEUR-WS.org. 2016.
54. Cormack GV, Grossman MR, Hedin B, Oard DW. Overview of the TREC 2010 legal track.. In: CEUR-WS.org. .
55. De Bruin J, Ma Y, Ferdinands G, Teijema J, Schoot V. dR. SYNERGY - Open machine learning dataset on study selection in systematic reviews. In: DataverseNL. 2023
56. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*. 2006;13(2):206–219.
57. Kusa W, E. Mendoza O, Samwald M, Knoth P, Hanbury A. CSMeD: Bridging the Dataset Gap in Automated Citation Screening for Systematic Literature Reviews. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S. , eds. *Advances in Neural Information Processing Systems*. 36. Curran Associates, Inc. 2023:23468–23484.

58. Alharbi A, Stevenson M. A dataset of systematic review updates. In: ACM. 2019:1257–1260.
59. Scells H, Zuccon G, Koopman B, Deacon A, Azzopardi L, Geva S. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In: ACM. 2017:1237–1240.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Repke, T., Tinsdeall, F., Danilenko, D., Flemyng, E., Graziosi, S., Müller-Hansen, F., Schmidt, L., Thomas, J. and Valkenhoef, G.. Don't stop me now, cause I'm having a good time screening: Evaluation of stopping methods for safe use of priority screening in systematic maps and reviews *Cochrane Evidence Synthesis and Methods* 2025;00(00):1–18.

APPENDIX

A EXTENDED RESULTS SECTION

In Table A1 and Table A2 we list the distribution of the x/y-axis histograms presented in Figure 2. With Figure A2 we also present an alternative grid of scatter-plots that shows the relationship between the proportion of missed relevant records and distribution of dataset a ranking simulations (Figure A1)

B STOPPING METHOD PARAMETERS

In this section we list all tested hyper-parameter settings for each stopping method. Where available, we use the same set of recall targets (80%, 90%, 95%, 99%) for each method. Additionally, we use all combinations of

- APRIORI: recall targets, inclusion threshold = 0.5
- BATCHPRECISION: batch size ∈ $\{500, 1000, 2000\}$, threshold∈ $\{0.05, 0.1, 0.2\}$
- BUSCAR: recall targets, bias∈ $\{., 2., 5., 10.\}$, confidence level = 0.99
- CURVE_FITTING: recall targets, windows∈ $\{10, 50\}$, confidence level∈ $\{0.8, 0.95\}$
- HEURISTIC_FIX: consecutive includes∈ $\{50, 100, 200, 300\}$
- HEURISTIC_FRAC: proportional consecutive includes∈ $\{1\%, 5\%, 7.5\%, 10\%, 20\%\}$
- HEURISTIC_RAND: recall targets
- HEURISTIC_SCORES: recall targets
- KNEE: window size for smoothing∈ $\{500\}$, polynomial∈ $\{1\}$, threshold for slope ratio∈ $\{2, 3, 4, 7\}$, threshold for distance to diagonal ∈ $\{0.2, 0.3, 0.4\}$
- METHOD2399: alpha∈ $\{1.0, 1.1, 1.2\}$
- QUANT_CI: recall targets, confidence interval scaling∈ $\{0, 1, 2\}$

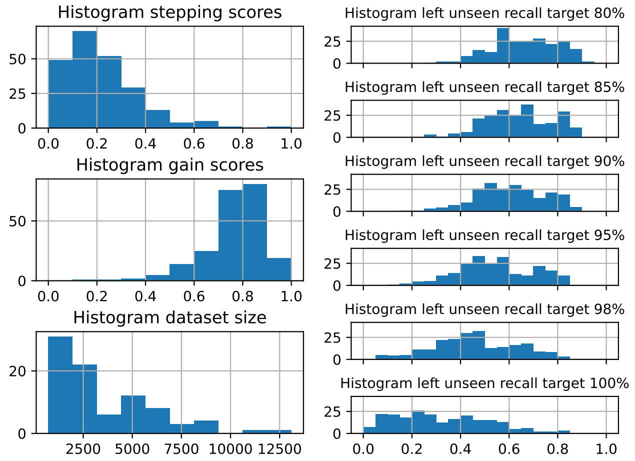

FIGURE A1 Histograms of where different theoretical recall targets are reached

TABLE A1 Distribution of values shown in x-axes of Figure 2

APRIORI	BATCHPRECISION	BUSCAR	CURVE_FITTING	HEURISTIC_FIX	HEURISTIC_FRAC	HEURISTIC_RANDOM	HEURISTIC_SCORES	KNEE	METHOD2399	QUANT_CI	% missed / % overshoot
0.0	0.2	0.0	0.0	56.9	58.7	0.0	0.0	0.0	0.0	0.0	-105
0.0	6.4	0.0	2.5	5.4	3.8	0.0	0.0	0.5	0.0	0.0	-94
0.0	5.2	0.0	1.5	2.0	1.7	0.3	0.0	0.2	0.2	0.0	-84
0.1	9.9	0.0	3.2	1.6	1.2	1.7	0.3	2.6	2.5	0.0	-73
1.0	8.0	0.0	2.7	1.3	0.8	2.6	1.4	3.5	3.8	0.0	-63
1.6	9.5	0.0	3.8	2.8	1.0	3.3	1.5	4.6	2.9	0.0	-52
1.5	8.5	0.0	3.5	4.0	2.9	5.7	3.0	6.5	5.7	0.0	-42
1.3	7.2	0.0	2.2	5.2	4.8	8.1	3.1	7.8	5.7	0.1	-31
0.8	5.7	0.0	2.3	4.7	5.3	13.8	3.9	7.8	6.1	1.3	-21
0.7	4.5	0.0	1.3	4.4	4.0	13.8	3.3	9.1	4.0	2.1	-10
0.5	2.9	0.0	0.8	1.0	0.6	16.3	1.7	4.6	1.7	8.1	0
0.0	3.2	0.0	0.0	1.7	2.8	3.5	0.5	5.5	1.2	10.0	10
0.0	2.4	9.4	0.0	1.8	2.7	1.3	0.3	5.1	2.8	8.7	21
0.0	2.8	12.5	0.2	1.6	2.1	0.9	0.1	3.6	3.4	6.2	31
0.0	1.8	13.3	0.7	0.8	1.2	0.3	0.0	3.5	1.7	5.4	42
0.0	1.8	10.9	1.3	0.8	1.2	0.3	0.1	4.3	1.8	5.4	52
0.0	1.7	10.2	2.5	0.7	0.9	0.1	0.0	4.2	0.3	4.1	63
0.0	1.8	13.3	3.5	0.3	0.8	0.2	0.0	2.3	2.3	4.5	73
0.0	0.8	14.8	2.3	0.3	0.6	0.1	0.0	2.7	2.2	4.3	84
92.5	15.6	15.6	65.6	2.5	2.9	27.5	80.6	21.7	51.8	39.7	94

TABLE A2 Distribution of values shown in y-axes of Figure 2

APRIORI	BATCHPRECISION	BUSCAR	CURVE_FITTING	HEURISTIC_FIX	HEURISTIC_FRAC	HEURISTIC_RANDOM	HEURISTIC_SCORES	KNEE	METHOD2399	QUANT_CI	Size remaining
20.9	45.2	25.0	24.5	45.2	45.2	21.8	20.9	45.2	45.2	20.9	0
27.1	26.7	25.0	25.9	26.7	26.7	26.8	27.1	26.7	26.7	27.1	650
15.2	6.0	7.0	8.3	6.0	6.0	15.2	15.2	6.0	6.0	15.2	1300
5.3	8.8	7.8	7.9	8.8	8.8	5.2	5.3	8.8	8.8	5.3	1950
6.7	3.7	10.9	9.3	3.7	3.7	6.5	6.7	3.7	3.7	6.7	2600
8.1	5.1	15.6	13.0	5.1	5.1	7.8	8.1	5.1	5.1	8.1	3250
4.8	0.9	2.3	2.8	0.9	0.9	4.7	4.8	0.9	0.9	4.8	3900
3.9	1.4	3.1	2.8	1.4	1.4	3.9	3.9	1.4	1.4	3.9	4550
1.5	1.4	0.8	0.5	1.4	1.4	1.6	1.5	1.4	1.4	1.5	5200
0.6	0.5	0.0	0.5	0.5	0.5	0.7	0.6	0.5	0.5	0.6	5850
1.2	0.0	0.0	0.5	0.0	0.0	1.2	1.2	0.0	0.0	1.2	6500
2.3	0.5	0.0	0.9	0.5	0.5	2.2	2.3	0.5	0.5	2.3	7150
1.0	0.0	1.6	1.4	0.0	0.0	1.0	1.0	0.0	0.0	1.0	7800
0.3	0.0	0.8	0.5	0.0	0.0	0.3	0.3	0.0	0.0	0.3	8450
0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	9100
0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	9750
0.2	0.0	0.0	0.5	0.0	0.0	0.2	0.2	0.0	0.0	0.2	10400
0.5	0.0	0.0	0.5	0.0	0.0	0.4	0.5	0.0	0.0	0.5	11050
0.2	0.0	0.0	0.5	0.0	0.0	0.2	0.2	0.0	0.0	0.2	11700
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12350

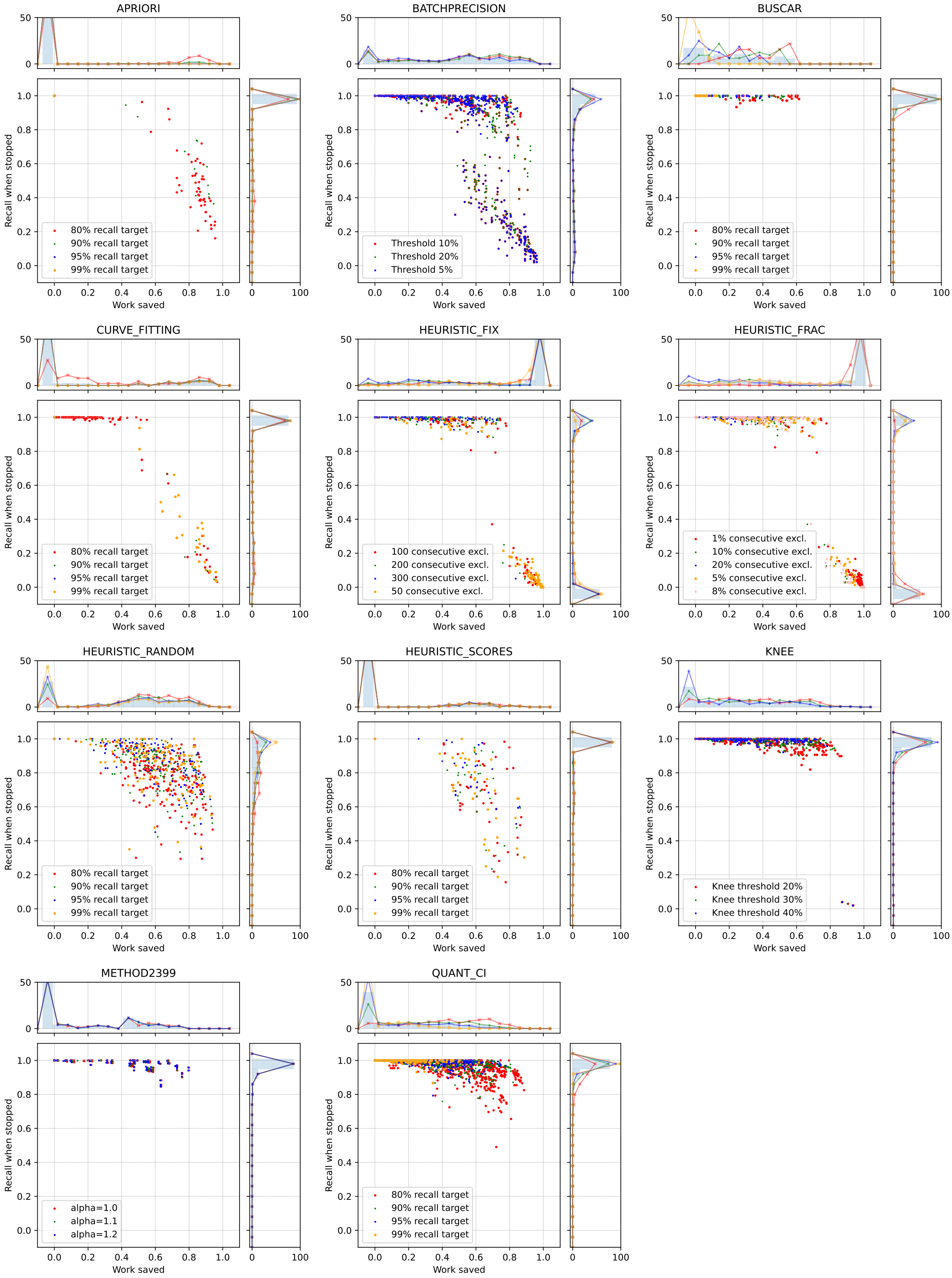

FIGURE A2 Distribution of trade-off between work saved thanks to early stopping and missed relevant records. The ideal stopping method would have all points at (0, 0), meaning that no relevant records were missed and the stopping rule invoked at the theoretically perfect time with no additional work. Results are grouped by stopping method and respectively one key hyper-parameter. Shaded red area marks 80–95% recall in 5%p increments. The histograms help to reveal the distribution of points for each axis which is otherwise obfuscated due to over-plotting in the scatterplot. Histogram lines correspond to points of each hyper-parameter grouping whereas bars show the overall distribution.

```
File: main.v3.tex
Encoding: utf8
Sum count: 5420
Words in text: 5067
Words in headers: 14
Words outside text (captions, etc.): 325
Number of headers: 13
Number of floats/tables/figures: 8
Number of math inlines: 14
Number of math displayed: 0
Subcounts:
  text+headers+captions (#headers/#floats/#inlines/#displayed)
  356+4+0 (5/0/0/0) _top_
  567+1+0 (1/0/0/0) Section: Introduction
  284+1+75 (1/2/0/0) Section: Methods
  489+2+9 (1/0/0/0) Subsection: Priority ranking
  683+2+0 (1/0/1/0) Subsection: Stopping methods
  205+1+111 (1/2/0/0) Subsection: Datasets
  910+1+0 (1/0/0/0) Section: Results
  897+1+0 (1/0/0/0) Section: Discussion
  676+1+130 (1/4/13/0) Section: Conclusion
```