







# An evaluation of the performance of stopping rules in AI-aided screening for psychological meta-analytical research

Lars König<sup>1</sup>  | Steffen Zitzmann<sup>2</sup>  | Tim Fütterer<sup>3</sup>  | Diego G. Campos<sup>4</sup>  |  
Ronny Scherer<sup>4</sup>  | Martin Hecht<sup>1</sup> 

<sup>1</sup>Helmut Schmidt University, Hamburg, Germany

<sup>2</sup>Medical School Hamburg, Hamburg, Germany

<sup>3</sup>University of Tübingen, Tübingen, Germany

<sup>4</sup>University of Oslo, Oslo, Norway

## Correspondence

Lars König, Helmut Schmidt University, Hamburg, Germany.  
Email: [lars.koenig@hsu-hh.de](mailto:lars.koenig@hsu-hh.de)

## Funding information

Research Council of Norway, Grant/Award Number: 331640; European Commission under the Horizon Europe scheme, Grant/Award Number: 101132474

## Abstract

Several AI-aided screening tools have emerged to tackle the ever-expanding body of literature. These tools employ active learning, where algorithms sort abstracts based on human feedback. However, researchers using these tools face a crucial dilemma: When should they stop screening without knowing the proportion of relevant studies? Although numerous stopping rules have been proposed to guide users in this decision, they have yet to undergo comprehensive evaluation. In this study, we evaluated the performance of three stopping rules: the knee method, a data-driven heuristic, and a prevalence estimation technique. We measured performance via sensitivity, specificity, and screening cost and explored the influence of the prevalence of relevant studies and the choice of the learning algorithm. We curated a dataset of abstract collections from meta-analyses across five psychological research domains. Our findings revealed performance differences between stopping rules regarding all performance measures and variations in the performance of stopping rules across different prevalence ratios. Moreover, despite the relatively minor impact of the learning algorithm, we found that specific combinations of stopping rules and learning algorithms were most effective for certain prevalence ratios of relevant abstracts. Based on these results, we derived practical recommendations for users of AI-aided screening tools. Furthermore, we discuss possible implications and offer suggestions for future research.

## KEYWORDS

literature screening, machine learning, meta-analysis, stopping rules, systematic reviews

## Highlights

### What is already known?

- AI-aided screening tools hold the potential to expedite literature searches in systematic reviews and meta-analyses. These tools employ active learning to prioritize abstracts based on their expected relevance, drawing from previously categorized abstracts.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Research Synthesis Methods* published by John Wiley & Sons Ltd.

- A substantial challenge for users of AI-aided screening tools lies in determining a reliable stopping point for terminating the screening process.
- While several stopping rules have been devised, their effectiveness remains inadequately evaluated within the context of psychological literature.
- Typically, these rules are evaluated without controlling for features of the screening tools (i.e., type of learning algorithm) or the attributes of abstract collections (i.e., prevalence of relevant abstracts).
- The performance of stopping rules is often measured using metrics that are unfamiliar to psychologists.

#### **What is new?**

- The performance of stopping rules (including the knee method, data-driven heuristic, and prevalence estimation) is measured in terms of sensitivity, specificity, and screening cost (percentage of records needed to be screened).
- The performance of stopping rules depends on the prevalence of relevant abstracts.
- Specific combinations of stopping rules and learning algorithms outperform others.

#### **Potential impact for *Research Synthesis Methods* readers?**

- Receive guidance on choosing stopping rules and learning algorithms tailored to specific prevalence rates of relevant abstracts.
- Researchers receive recommendations regarding aspects of AI-aided screening.
- Researchers can access code that can be utilized stopping rules in R.

## **1 | INTRODUCTION**

The evolving scientific knowledge results in an expanding body of research findings. Synthesizing evidence via systematic reviews and meta-analyses in a particular research domain consequently demands a greater investment of time and effort.<sup>1</sup> Moreover, these syntheses typically necessitate the collaborative efforts of multiple researchers.<sup>2</sup> Especially the literature search and screening of abstracts constitute a substantial portion of the workload.<sup>3,4</sup> Recommendations, such as (a) giving preference to broadly defined search terms over narrowly focused ones, (b) systematically expanding the search by examining the references cited within the identified articles (backward snowballing), and (c) searching for articles that cited the identified articles (forward snowballing), underscore the substantial effort required to identify relevant articles for research syntheses.<sup>5–7</sup> Many tools have been developed to assist researchers in conducting research synthesis for various tasks involved in the process.<sup>8</sup> Among these, some tools enhance the search process by automating forward and backward search techniques.<sup>9,10</sup> While this automation might improve the quality of research syntheses by identifying more potentially relevant articles, it also requires

additional time and resources to screen the abstracts of these articles for eligibility. For instance, Wallace et al.<sup>11</sup> noted that screening an abstract typically requires about 30 s. Consequently, screening 5000 abstracts demands approximately 40 h of skilled labor, and identifying an additional 1000 potentially relevant abstracts would extend the screening time by another 8 h. However, similar to the literature search, the screening of abstracts can be accelerated by using modern, innovative AI-aided screening (AI-AS) tools.<sup>12–14</sup> These tools primarily utilize machine learning techniques to reorder abstracts based on their predicted relevance to the researcher.<sup>8,15</sup> This altered order should enable researchers to identify all relevant abstracts, and therefore articles, before screening all of them. Thus, the time required for manual screening could diminish while maintaining the number of identified relevant articles. Notably, identifying 95% of the relevant literature using AI-assisted screening (AI-AS) is typically considered sufficient performance. This is based on findings that conducting meta-analyses without these last-to-find studies did not impact the main findings.<sup>16</sup> Moreover, traditional random screening can also result in the misclassification of around 10% of the abstracts due to factors such as fatigue.<sup>17</sup> While fatigue can also affect the quality of any AI-AS, the smaller number of items

requiring screening could reduce this risk.<sup>8,18</sup> In addition, aiming for 100% rather than 95% of relevant abstracts can increase the screening time in AI-AS (see Reference 19), thereby reducing its benefits and increasing the risk of fatigue.

In a recent review, Burgard and Bittermann<sup>15</sup> summarized and compared the effectiveness of 15 tools supporting the screening of abstracts utilizing machine learning algorithms to order abstracts based on their predicted probability of being relevant. They observed that these tools helped to identify at least 95% of relevant articles after screening a median of approximately 40% of the abstracts. Moreover, in their study, 25% of the summarized evaluation studies reported screening up to 14% of the abstracts was sufficient to achieve this identification rate. In comparison, another 25% reported that at least 70% needed to be screened to identify a minimum of 95% of the relevant literature. This considerable performance variation was observed between different AI-AS tools and within the same tool. The authors concluded that reasons for such heterogeneous findings include differences in data (i.e., abstracts from different literature searches, hereafter referred to as abstract collections) and the prediction models employed.<sup>15</sup> Thus, performance seems to depend on a specific model-data combination, limiting the generalizability of the reported screening savings. Consequently, despite their potential, users of AI-AS tools face difficulties in determining how many abstracts need to be screened to identify at least 95% of the relevant articles.

To tackle the challenge of determining a reliable stopping point when screening abstracts utilizing AI-AS tools, a variety of stopping rules with different levels of complexity have been developed.<sup>11,20–25</sup> Unfortunately, the performance of most stopping rules in AI-AS remains uncertain, and the existing literature lacks a comprehensive evaluation of factors that could impact their performance.<sup>15</sup> Moreover, the generalizability of existing findings to different research domains is limited, as current evaluation studies primarily rely on literature reviews from clinical, educational, or nonpsychological-related research domains.<sup>11,21,23–26</sup> Given these limitations, additional research is needed to develop reliable stopping rules and to derive suggestions for their implementation across various research contexts. Without sufficient performance (i.e., identifying at least 95% of the relevant literature) of these rules, users may remain uncertain about the number of potentially missed relevant articles. Missing relevant articles, in turn, might introduce bias into the results of their quantitative syntheses. Consequently, it is crucial to advance the understanding and development of robust stopping rules that identify at least 95% of the relevant literature.

Therefore, the present study's primary objective is to systematically evaluate factors that may influence the performance of stopping rules in AI-AS. Specifically, we focus on three aspects: the generalizability of previous findings to psychological research domains, the influence of the prevalence of relevant articles, and the influence of the learning algorithm on the performance of stopping rules. To achieve this, we conducted a hybrid simulation study that leveraged real data to simulate the AI-AS process. First, we manipulated the prevalence of relevant articles in abstract collections sourced from diverse psychological research domains. Second, using the open-source AI-AS software "ASReview,"<sup>19</sup> we applied different learning algorithms to each manipulated abstract collection to simulate AI-AS. We assessed the performance of three different stopping rules across these conditions: a prevalence estimation technique<sup>6</sup> that estimated the number of relevant abstracts, a heuristic stopping rule<sup>24</sup> that stopped after  $n$  consecutive irrelevant abstracts, and the knee method<sup>22</sup> a statistical stopping procedure. Performance was evaluated in terms of sensitivity (i.e., the percentage of relevant abstracts among all screened abstracts), specificity (i.e., the percentage of irrelevant articles among all unscreened articles), and cost (i.e., the percentage of articles screened until the stopping rule is met). Whereas sensitivity and specificity assessed the accuracy of the stopping rule, the cost measure described the efficiency of AI-AS compared to traditional screening. By combining these measures, researchers can achieve an optimal balance between accuracy and efficiency in the AI-AS process.

## 2 | THEORETICAL BACKGROUND

Screening abstracts and titles is a key step in any systematic review.<sup>5</sup> However, it is also one of the most time-consuming components and typically requires the work of several researchers.<sup>3,4</sup> Numerous scientists have already advocated for enhancing the efficiency of this process.<sup>3,12–14</sup> Among the solutions discussed in the literature, AI-AS is receiving increasing attention.<sup>15</sup>

### 2.1 | AI-aided screening

In their review, Burgard and Bittermann<sup>15</sup> identified 15 AI-AS tools that aim to speed up the screening of abstracts and titles. These tools employ various approaches to achieve this objective, with 11 of them utilizing active learning. In the context of AI-AS, active learning describes a process in which a machine learning algorithm orders abstracts on the basis of a relevance

estimate and then refines this estimate through feedback from a human screener. This iterative, semiautomated approach allows the algorithm to learn from the feedback of a human screener and continuously improve its performance in identifying relevant abstracts. Two essential processes are required to utilize an active learning algorithm for AI-AS: First, the algorithm extracts relevant information, such as phrases, keywords, and patterns from all abstracts. This process is performed by algorithms known as *feature extractors*, such as *Term Frequency-Inverse Document Frequency* (TF-IDF<sup>27</sup>), *doc2vec*,<sup>28</sup> and *Sentence-Bidirectional Encoder Representations from Transformers* (SBERT<sup>29</sup>). The extracted elements are then compared across relevant and irrelevant abstracts and weighted to determine their importance in predicting the relevance of unseen abstracts. Second, using this information, *classifiers* calculate inclusion probabilities for abstracts that have not yet been screened. These probabilities indicate the likelihood of an abstract being relevant, guiding the order in which the abstracts should be presented for screening. Typical classifiers in AI-AS are *Random Forest* (RF<sup>30</sup>), *Support Vector Machine* (SVM<sup>31</sup>), *Fully Connected Neural Network with 2 hidden layers* (nn-2-layer<sup>32</sup>), *Logistic Regression* (LR<sup>33</sup>), and *Naïve Bayes* (NB<sup>34</sup>).

However, in the following sections, we will refer to the combination of both feature extractors and classifiers as a *learning algorithm*. Such learning algorithms require a training set to create the initial ranking of relevancy, including at least one abstract labeled as relevant and one labeled as irrelevant. All manually screened and labeled abstracts are then added to the training set, providing more information to the learning algorithm to update the ranking. In this sense, active learning is not an entirely automated process of labeling abstracts. Active learning requires the active involvement and feedback of a human screener. For a comprehensive review of active learning, we refer readers to Settles.<sup>35</sup> Besides this semiautomatic approach, some AI-AS tools employ fully automated methods that only require training data to categorize abstracts as relevant or irrelevant. However, their performance is insufficient, whereas semiautomatic AI-AS performed well on the same abstract collections (<sup>18</sup>; see also<sup>8</sup>).

## 2.2 | The AI-AS tool “ASReview”

ASReview is an AI-AS tool that integrates all the feature extractors and classifiers utilized in this study.<sup>36</sup> Whereas other AI-AS tools include some of the learning algorithms that ASReview contains (e.g., SVMs), the flexibility to choose from a wide range of learning algorithms is

unique to ASReview (see Reference 15 for a comparison of the tools). In addition, ASReview is an open-source software that can be accessed via a user interface, command line, and Python Application Programming Interface (API). Being open-source, ASReview allows users to access and modify its source code and ensure transparency in the review process. Moreover, ASReview offers an intuitive simulation mode, which can simulate the AI-AS process on pre-labeled data. In essence, AI-AS is simulated by imitating a human screener. After a training set of at least one relevant and one irrelevant abstract is provided, all abstracts are ordered according to their predicted probability of belonging to the relevant category. The abstract with the highest likelihood of being relevant is then selected and labeled according to the associated labels previously provided by a human screener. Thus, the simulation mode consistently aligns with the decisions made by the researchers who initially screened the abstracts for their research synthesis.<sup>36</sup>

ASReview has consistently demonstrated satisfactory performance across various evaluation studies.<sup>19</sup> observed that screening only 8–33% of abstracts led to the identification of 95% of all relevant abstracts. These positive outcomes were further supported by Ferdinands,<sup>37</sup> who observed that screening 10% of all abstracts led to the identification of over 80%, and screening 20% enabled the discovery of 95% of relevant publications.<sup>38</sup> However, it is essential to recognize that studies displayed varying performances across different learning algorithms and abstract collections. For instance, in a study conducted by Harmsen et al.,<sup>39</sup> the authors identified 95% of the relevant abstracts by screening anywhere from 2 to 70% of all abstracts while employing the same learning algorithm. Moreover, they highlighted the critical role of accurate abstract classification, emphasizing its substantial influence on the algorithm's performance. Higher-quality labels yielded better performance. To assess classification quality, individuals with varying skills labeled the abstracts as relevant or irrelevant. The labeled abstracts were then used to simulate using the AI-AS tool ASReview. Another study delved deeper into the variations in performance among different learning algorithms. While keeping the abstract collection constant, Teijema et al.<sup>16</sup> observed considerable differences in the order of abstracts, which is determined by the learning algorithms. The authors also explored whether switching to more complex learning algorithms after meeting stopping rules could enhance the performance of the overall AI-AS process. Their results suggested that additional relevant articles can be identified more quickly by switching to a different learning algorithm. Nonetheless, this effect was less pronounced when using the LR + SBERT algorithm (see Supporting

Information), which outperformed other algorithms with and without switching.<sup>16,26</sup> Another technique to increase the overall performance of the AI-AS screening process was proposed by Boetje and van de Schoot.<sup>40</sup> The authors developed a four-step screening procedure for AI-AS, which comprises a random screening phase, an AI-AS screening phase, another AI-AS screening phase with a different learning algorithm, and a quality control phase. The latter summarizes an AI-AS of the previously excluded abstracts to control for misclassification. This step is similar to a recently proposed noisy label filter framework, which aims to reconstruct and update previously conducted literature searches.<sup>41</sup>

Although these results imply the immense potential of ASReview, it is important to emphasize that most evaluation studies have primarily focused on clinical, educational, or non-psychology-related literature.<sup>16,19,26,38,39</sup> Consequently, performance variations may arise when using AI-AS for abstract collections from other research domains. Such variations might be attributed to differences in the number of abstracts, the prevalence of relevant abstracts, and the specific keywords utilized in each domain.<sup>26</sup> Moreover, as some stopping rules depend on the occurrences of consecutive nonrelevant abstracts during the screening process, we contend that variations in screening orders among different learning algorithms could potentially influence the effectiveness of these stopping rules.

## 2.3 | Stopping AI-AS

Determining the optimal stopping point in AI-AS is a critical aspect of the AI-AS process, ultimately influencing the identification rate of relevant literature. In traditional literature screening, approximately 10% of relevant articles are overlooked.<sup>17</sup> However, as Teijema et al.<sup>16</sup> reported, missing 5–10% of relevant articles may not significantly impact the results of a meta-analysis.<sup>16</sup> Therefore, researchers should aim to achieve at least a similar performance when utilizing AI-AS. Moreover, it is essential to find stopping rules that can achieve such performance with the lowest screening cost, that is, with the fewest abstracts needed to be screened. However, the lack of systematic evaluations of the performance of various stopping rules makes it challenging to select a stopping rule that can achieve this goal. Whereas some AI-AS tools have integrated estimation of the number of relevant studies to aid users in this decision-making,<sup>23,25</sup> many tools, including ASReview, do not provide explicit guidance on choosing a stopping point. In our review of meta-analyses and systematic reviews referencing ASReview, we found that many studies did not disclose their

stopping criteria for the AI-AS process. This lack of disclosure may be attributed to the absence of initial guidelines for AI-AS tools. Thus, we urge users of these tools to adhere to both conventional systematic review and meta-analysis guidelines, as well as new guidelines to ensure transparency and reproducibility of the AI-AS process.<sup>42,43</sup> Several studies have utilized the stopping rule proposed by Ros et al.,<sup>24</sup> which involves stopping the AI-AS process after screening  $n$  consecutive irrelevant abstracts. The value of  $n$  varied between studies, ranging from 20 to 500.<sup>44,45</sup> This finding, coupled with the observation that some authors opted to screen all identified abstracts despite using ASReview,<sup>46,47</sup> underscores the uncertainty surrounding the selection of a stopping rule and highlights the need for a systematic evaluation.

### 2.3.1 | Stopping rules

Several stopping rules have been developed to address the challenge of determining a reliable stopping point in AI-AS. Among these rules, *heuristic stopping rules* provide a straightforward and practical approach. Two prominent examples are the time-based approach, where the screening should be stopped after screening a certain percentage of the abstracts,<sup>11</sup> and the data-driven approach,<sup>24</sup> which stops the screening after  $n$  consecutive irrelevant abstracts. Whereas Wallace's time-based approach relies on screening a fixed percentage, Ros et al.'s<sup>24</sup> data-driven method is more adaptive, allowing adaptation to a specific context, such as different prevalence ratios. Other stopping rules adopt a more statistically grounded approach to determine the optimal stopping point. For instance, prevalence estimation methods estimate the number of relevant abstracts in a given collection. The screening can then be stopped once the estimate equals the number of identified relevant abstracts or a specified proportion of it. One such method is included in the AI-AS tool SWIFT-Active Screener.<sup>23</sup> It estimates the number of relevant abstracts based on the assumption that their occurrence during AI-AS follows a negative binomial distribution. Moreover, the estimate is derived from the number of relevant abstracts found between the latest screened abstract and the  $X$ th previously identified relevant abstract. Consequently, this method requires implementation in the AI-AS tool to update the estimate with each screening decision (see Reference 23 for additional information). However, due to its complexity insufficient information regarding its implementation, this estimation technique cannot be applied outside of the AI-AS tool SWIFT-Active Screener. Therefore, Callaghan and Müller-Hansen<sup>21</sup> proposed an alternative estimation technique that operates similarly

but is based on the assumption that relevant abstracts follow a hypergeometric distribution. While the authors provide all necessary information to use this technique, it also needs to be integrated into the screening process. Another recently published estimation technique which also requires implementation in the AI-AS process, utilizes Chao's Population Size Estimator to estimate the number of relevant articles during the AI-AS process.<sup>20</sup> In contrast to these techniques, van Haastrecht et al.<sup>6</sup> proposed a simple technique that can be used to estimate the prevalence of relevant articles before starting the AI-AS process. The authors suggested estimating the number of relevant abstracts based on a random sample from the identified and deduplicated results from the literature search (abstract collection). Users should randomly screen abstracts until a predefined number of relevant studies have been detected or a specific percentage of the abstracts have been screened. For instance, a researcher can randomly screen 50 abstracts out of 1000 identified articles and estimate the prevalence based on this 5% subset, as shown in Equation 1. In this equation,  $N$  is the number of abstracts in a given abstract collection,  $r$  is the number of relevant abstracts identified by random screening, and  $i$  is the number of irrelevant abstracts screened randomly.

$$R \approx N \left( \frac{r}{r+i} \right) \quad (1)$$

To prevent overestimation due to oversampling relevant abstracts, the estimated number  $R$  should further be multiplied by a factor of 0.95. After this estimation, the remaining 950 abstracts can be screened using an AI-AS tool. The screening can be stopped after the  $0.95 \cdot R$  relevant abstracts are identified.

Besides estimating the number of relevant articles, another statistically grounded technique is the *knee method*.<sup>22</sup> This method calculates the ratio between the slope before ( $slope_{<r}$ ) and after a critical inflection point ( $slope_{>r}$ ) in the gain curve (see Figure B1, Appendix B), which is referred to as the “knee.”<sup>48</sup> The gain is represented by the connection between the number of screened abstracts (x-axis) and the number of identified relevant abstracts (y-axis). As the learning algorithm improves its performance in identifying relevant abstracts during screening, the gain curve typically becomes steeper at the beginning. However, at some stage in the AI-AS process, the gain curve levels off because the algorithm encounters difficulties in identifying the remaining relevant abstracts that tend to be less similar to the initially screened, relevant abstracts. This point of inflection is the “knee.” The calculation of a simple version of the slope ratio in the knee method, adapted from van

Haastrecht,<sup>49</sup> is shown in Equation 2. With  $r$  representing a given rank below the total number of the screened abstracts  $s$ . The number of the relevant abstracts before rank  $r$  is represented by  $rel_{<r}$ , the number of abstracts screened at rank  $r$  is  $i$ , and the number of relevant abstracts among all screened abstracts is represented by  $rel_{total}$ . The 1 in the denominator is added for smoothing, that is, to avoid edge cases in which  $i$  is close to  $s$ , and no more relevant articles are identified after rank  $r$ .

$$slope\ ratio = \frac{slope_{<r}}{slope_{>r}} = \frac{\frac{rel_{<r}}{i}}{\frac{1+rel_{total}-rel_{<r}}{s-i}} \quad (2)$$

To stop the AI-AS process, the slope ratio is compared to a predefined value. When it exceeds this threshold, the AI-AS process can be stopped. In an ideal scenario, a perfect algorithm would order the unlabeled abstracts so that all relevant abstracts are placed at the top and, thus, at the beginning of the screening. In this case, the  $slope_{<r}$  would be 1, and the  $slope_{>r}$  would be 0. However, in a more realistic example, the  $slope_{<r}$  will be below 1, and the  $slope_{>r}$  will only approach 0. For instance, in a dataset with 1000 abstracts, 50 of which are relevant, a  $slope_{<r}$  of 0.7 could be attained when 35 relevant abstracts are within the first 50 screened abstracts. In this example, the knee would be reflected by the rank 50. The  $slope_{>r}$ , on the other hand, would be equal to 0.1 when the next relevant abstract would appear after screening an additional 10 abstracts. Note that the knee method adds 1 to the number of relevant abstracts identified after the knee for smoothing. Thus, a slope of 0.1 would be attained after screening an additional 20 abstracts, in which 1 would be labeled as relevant. However, the authors recommend that the slope ratio should be greater than or equal to 6 to stop the AI-AS process. In our example, this would have led to stopping the AI-AS.

Nevertheless, the authors also propose that even higher values than 6 might be more suitable in certain contexts. Unfortunately, the knee method has not been incorporated in any of the 15 tools reviewed by Burgard and Bittermann.<sup>15</sup> To apply this method, the AI-AS must be paused to extract the ordered list of all currently labeled abstracts. Using this data, the knee can be calculated. If the knee value remains below 6, the user must continue the screening, pausing, and knee calculation process. Thus, the knee method's unique characteristic might limit its practical use in the eyes of many users.

In addition to the presented stopping rules, various other methods have been developed.<sup>21,22,50,51</sup> Nonetheless, after reviewing studies citing ASReview and following an ongoing discussion on GitHub,<sup>52</sup> we identified the explained stopping rules as the most commonly used and discussed methods within the community of ASReview.

### 2.3.2 | Performance of the stopping rules

As noted earlier, systematic evaluations of the performance of the stopping rules are still rare. Specifically, there is no evaluation study regarding the prevalence estimation technique proposed by van Haastrecht et al.<sup>6</sup> However, the estimation procedure integrated in the AI-AS tool SWIFT-Active Screener has shown promising results, identifying 91.5–100% of all relevant abstracts across 26 clinical abstract collections.<sup>23</sup> Additionally, certain implementations of the estimation technique based on Chao's Population Size Estimator have been found to perform well across a variety of different abstract collections from various research domains, with identification rates of relevant abstracts above 95%.<sup>20</sup> Also, the estimation technique of Callaghan and Müller-Hansen<sup>21</sup> resulted in similar identification rates regardless of the performance of learning algorithms. Nonetheless, such estimation techniques bear the risk of under- and overestimating the proportion of relevant abstracts. The latter results in not stopping the AI-AS until all abstracts are screened.<sup>53</sup> Moreover, as these techniques require information such as the rank at which an abstract was screened, they need to be implemented in the AI-AS tool or the screened data must be downloaded to calculate the predicted number of relevant abstracts. Thus, they lack practical usefulness.<sup>54</sup> The simple estimation technique of van Haastrecht et al.<sup>6</sup> overcomes this shortcoming.

The evidence regarding the performance of the heuristic approach proposed by Ros et al.<sup>24</sup> is mixed. For example, stopping after 50 consecutive irrelevant abstracts resulted in finding 76–97% of the relevant abstracts across four abstract collections from the field of computer science.<sup>25</sup> However, Callaghan and Müller-Hansen<sup>21</sup> found that the same procedure failed to identify 95% of all relevant abstracts in 39% of the cases when assessed across 20 clinical abstract collections. The performance improved when stopping after 200 consecutive irrelevant abstracts. This implementation of the data-driven heuristic stopping rule resulted in finding 95% of the relevant abstracts after screening 17.14% of them. This finding aligns with other recent investigations of this method, which revealed that the performance of stopping after screening 50 consecutive irrelevant abstracts varied strongly across different medical-related abstract collections, with identification rates ranging roughly between 30 and 100%. Stopping after 100 consecutive irrelevant abstracts resulted in better performance, with identification rates ranging from around 65–100%.<sup>54</sup> Moreover, in this study, the difference in performance across cut-off values was more pronounced when using the AI-AS tool Rayyan<sup>55</sup> compared to ASReview.<sup>19</sup> However, the median performance with a cut-off value of 100 was higher for

Rayyan. When comparing cut-off values of 50, 100, 150, and 200 for an abstract collection from the field of health economics, other interesting findings emerged.<sup>56</sup> While the median performance increased notably when the cut-off was set to 100 instead of 50, it did not further increase for larger cut-off values. Nonetheless, the variability in performance across replication runs with varying training sets diminished when at least 150 consecutive irrelevant abstracts needed to be screened before stopping the AI-AS. Another unique result from this research is that, while all cut-off values resulted in missing roughly 20% of the potentially relevant abstracts, they still identified almost 100% of the studies deemed relevant after full-text screening. In addition, when five instead of one randomly selected relevant and irrelevant abstract were used to train the learning algorithm (NB + TF-IDF), the variability in performance was considerably reduced for both potentially relevant and actually relevant abstracts. In contrast to these studies, Campos et al.<sup>26</sup> evaluated the data-driven heuristic with an adaptive cut-off value. They determined the number of consecutive irrelevant abstracts as percentages of the total number of abstracts. Thus, a cut-off value of 5% in an abstract collection of 1000 abstracts translates to stopping after 50 consecutive irrelevant abstracts, while 100 consecutive irrelevant abstracts need to be screened for an abstract collection of 2000. When averaged across different learning algorithms and abstract collections from educational and educational psychology research syntheses, setting this value to 7% resulted in detecting 95% of relevant abstracts in all abstract collections.

As for the data-driven heuristic, results regarding the time-based heuristic are mixed. For instance, while screening 50% of all abstracts resulted in identifying up to 100% of relevant abstracts in some studies,<sup>11,57</sup> others required screening 70%–achieve the same result.<sup>26,54</sup> Moreover, in a recent study by Oude Wolcherink et al.,<sup>56</sup> stopping after screening 7.5% of the abstracts resulted in identifying less than 75% of potentially relevant articles, while nearly all actually relevant articles after full-text screening were identified. The authors also observed that the variability in performance across simulation runs with different training sets diminished when using five instead of one randomly chosen abstract to train the learning algorithm. However, these results are based solely on a single abstract collection from the field of health economics. Notably, combining the data-driven and time-based heuristics revealed promising results. In a recent study by Campos et al.,<sup>26</sup> 95% of the relevant records were detected after screening 20% of the abstracts and then stopping after 5% consecutive irrelevant abstracts. Note that other implementations of this combination resulted in the same detection rate.

Apart from the heuristic stopping rules, the knee method identified 93–99% of all relevant records across 10 diverse text collections, including hacker forums and clinical databases.<sup>22</sup> These results were also confirmed by Yu and Menzies,<sup>25</sup> who observed that the knee method found 88–97% of all relevant abstracts in four computer science-related text collections.

In conclusion, the results of these evaluation studies highlight considerable variability in the performance of the described stopping rules across different research contexts and AI-AS tools, thus compromising their generalizability. Furthermore, as the majority of evaluation studies were based on abstract or text collections from nonpsychological research domains, the performance of these rules in the field of psychology remains unknown.<sup>11,21,23–25,54,57</sup> Given the variation in publishing guidelines, keywords, and phrases across different research areas, we anticipate that these distinctions may impact the learning algorithm's capacity to effectively formulate a relevance ranking. Another limitation is that none of the evaluation studies manipulated or standardized the abstract collections, making it challenging to trace performance variations across abstract collections and studies back to specific factors, such as the abstract collection itself or the prevalence of relevant abstracts. Finally, different AI-AS tools with varying learning algorithms were used for the performance evaluations, further complicating the comparison of results. Therefore, the primary goal of this study is to address these limitations and determine whether the variations in performance are influenced by the abstract collections themselves, the learning algorithms utilized, or the discrepancies in the prevalence of relevant abstracts.

## 2.4 | The present study

As noted earlier, AI-AS tools can considerably reduce the time and effort needed to conduct a systematic review.<sup>15</sup> However, at the current stage of research, it is not yet possible to make any reliable statements regarding the actual performance of such tools. A critical component in utilizing AI-AS is the determination of a stopping point. The stopping point not only determines how much time can be saved but also how many of the relevant abstracts can be detected.<sup>26</sup> To date, our knowledge regarding the performance of stopping rules in diverse conditions remains scarce, particularly within the context of psychology-related abstract collections. Furthermore, a systematic evaluation of the prevalence of pertinent abstracts remains absent from the literature. Additionally, there is a notable gap in our understanding regarding how different learning algorithms may influence the effectiveness of various stopping rules.

The present study aims to overcome these shortcomings by assessing the performance of three different stopping rules with abstract collections sourced from various psychological research domains. We standardized the prevalence of relevant abstracts across abstract collections and manipulated it within each. Then, we employed different learning algorithms for each manipulated abstract collection and applied several stopping rules. Through this comprehensive and systematic approach, we sought to shed light on the optimal utilization of AI-AS in psychological research and provide valuable insights for the development of efficient and accurate screening processes. We chose ASReview for our study because it offers various learning algorithms—a feature allowing us to assess differences between them. To evaluate the efficiency and accuracy of the stopping rules, we focused on key measures, such as sensitivity, specificity, and cost. We addressed the following research questions (RQs) concerning the performance of stopping rules in AI-AS:

**RQ1.** How do the three stopping rules (i.e., prevalence estimation, knee method, and data-driven heuristic) perform under different abstract collections, prevalence ratios, and learning algorithms?

**RQ2.** How does the prevalence ratio moderate the performance of the stopping rules?

**RQ3.** How does the learning algorithm moderate the performance of the stopping rules?

**RQ4.** How does the learning algorithm moderate the interaction between abstract collection and prevalence ratio?

**RQ5.** How does the abstract collection moderate the effect of the stopping rules?

## 3 | METHODS

The present study is best described as a hybrid simulation approach combining real-world data collection with simulation techniques. We gathered abstract collections from meta-analyses published in six different psychological research domains. We then manipulated the abstract collections regarding the prevalence ratio of relevant to irrelevant abstracts. All manipulated datasets were subsequently used in ASReview to simulate abstract screening while using different stopping rules. We preregistered each step of the present study and adjusted the preregistration to address the challenges we encountered when



drawing meta-analyses from the different research domains (<https://osf.io/ucz8d>). Given the exploratory nature of our study, we did not preregister any hypotheses, and all changes of the preregistration were documented and justified (see Supporting Information). Furthermore, we have made all relevant materials, including the data and analytic code, available via the Open Science Framework (<https://osf.io/7yhrq>). However, we did not receive permission to share all abstract collections.

### 3.1 | Data collection

We carefully designed the data collection procedure to create diverse abstract collections covering various psychological research domains: *Applied Psychology*, *Social Psychology*, *Biological Psychology*, *Clinical Psychology*, *Developmental Psychology*, and *Educational Psychology*. To ensure that the project remains within manageable boundaries, we predefined the number of meta-analyses from which to request the data to 180. We distributed the request as equally as possible across research domains. To ensure the feasibility of our data manipulation, we established eligibility criteria for the meta-analyses. Specifically, an eligible meta-analysis had to encompass a minimum of 50 relevant abstracts out of at least 1000 screened abstracts. In addition, the meta-analysis had to be documented in either English or German. To account for missing meta-analyses in a specific domain, we requested additional meta-analyses from one of the other five domains (for more details, please see the Supporting Information). This resulted in the following distribution of data requests: Applied Psychology ( $n = 36$ ), Biological Psychology ( $n = 14$ ), Clinical Psychology ( $n = 36$ ), Developmental Psychology ( $n = 35$ ), Educational Psychology ( $n = 29$ ), and Social Psychology ( $n = 30$ ). Ultimately, we acquired data from 28 meta-analyses, 21 of which met the eligibility criteria. To be considered eligible, the datasets had to include at least abstracts and the respective screening decisions. Unfortunately, we did not receive any eligible data from the domain of Biological Psychology. However, the data we received was relatively evenly distributed across the other domains: Applied Psychology ( $n = 4$ ), Clinical Psychology ( $n = 3$ ), Developmental Psychology ( $n = 5$ ), Educational Psychology ( $n = 5$ ), and Social Psychology ( $n = 4$ ). All received abstract collections were cleaned by deleting records with missing abstracts. As most received abstract collections lacked a digital object identifier for each abstract, we deduplicated the abstracts based on the title using the R package *revtools*.<sup>58</sup> In some cases, we obtained the data in a text format and then separated titles from abstracts. Descriptive statistics for the datasets are presented in Table 1.

### 3.2 | Data manipulation

Our sample consisted of 21 *original abstract collections* (OACs) from 21 meta-analyses. Each OAC was reassembled into four *artificial abstract collections* (AACs) with manipulated prevalence ratios of relevant and irrelevant abstracts (0.5, 1, 5, and 10%), yielding  $4 \cdot 21 = 84$  AACs. To address the issue of selection bias during the construction of the AACs, we implemented a resampling approach in which each collection was resampled 1000 times. Through this method, we obtained  $84 \cdot 1,000 = 84,000$  *replicated AACs* (RAACs), which were the unit of analysis (please see the Supporting Information for more information).

### 3.3 | Simulation design

The simulation was conducted with the Python API from ASReview.<sup>36</sup> Our analytic code was written in R<sup>80</sup> with the help of the R package *reticulate*,<sup>81</sup> which integrates Python code into R. Each of the 84,000 RAACs was used to simulate AI-AS with each of nine learning algorithms (i.e., LR + doc2vec, LR + SBERT, LR + TFIDF, NB + TFIDF, nn2layer + doc2vec, nn2layer + SBERT, RF + doc2vec, RF + TFIDF, and SVM + TFIDF), resulting in 756,000 simulation runs. In each run, one relevant and one irrelevant abstract were randomly selected to train the learning algorithm. It is important to note that the authors of the respective meta-analyses solely determined the classification of abstracts as relevant or irrelevant. We did not independently classify abstracts ourselves—instead, we relied on the authors' classification. Across all simulation runs, we set up ASReview with the default balancing strategy “dynamic resampling” and the default, certainty-based query strategy (see Supporting Information for additional information regarding ASReview). We used the same seed to simulate the data as the one we used to create the RAAC so that the same seed was used for all AACs of the same replication run. This way, we ensured homogeneity within each replication run and introduced heterogeneity between these runs. We recalculated inclusion probabilities after every 10 newly labeled abstracts to reduce computation time. As a result, we received 756,000 ordered abstract collections, in which each abstract was labeled as either relevant or irrelevant in the order they would have been screened. To reduce computational time, the simulation process stopped after all relevant articles had been found. We then applied each of the three stopping rules on each ordered RAAC. Overall, we simulated the application of each stopping rule 756,000 times, which resulted in 2268,000 data points for each performance measure.

TABLE 1 Descriptives of the original and artificially constructed abstract collections.

Meta-analysis	Original abstract collection			Artificial abstract collection			
	D.	N; $n_{relevant}$	Ratio	N; $n_{relevant}$			
				0.5%	1%	5%	10%
Vermillet et al. <sup>59</sup>	D	1875; 206	12.34	1407; 7	1515; 15	1659; 79	1738; 158
Bottema-Beutel et al. <sup>60</sup>	D	6283; 743	13.41	5226; 26	5252; 52	5523; 263	5786; 526
Khazanov et al. <sup>61</sup>	C	3423; 250	7.88	3015; 15	3030; 30	3150; 150	2607; 237
Reimer and Sengupta <sup>62</sup>	S	2661; 219	8.97	2211	2323; 23	2415; 115	2288; 208
Hall et al. <sup>63</sup>	E	13,531; 544	4.19	12,261; 61	12,423; 123	10,836; 516	5676; 516
Simonsmeier et al. <sup>64</sup>	E	9768; 1507	18.24	7839; 39	7878; 78	8232; 392	8624; 784
Hsieh et al. <sup>65</sup>	S	2343; 107	4.79	2010; 10	2121; 21	2121; 101	1111; 101
Alden et al. <sup>66</sup>	A	1486; 59	4.13	1206; 6	1313; 13	1176; 56	616; 56
Liu et al. <sup>67</sup>	C	1585; 579	57.55	804; 4	909; 9	987; 47	1045; 95
Tang et al. <sup>68</sup>	E	2053; 53	2.65	1809; 9	1919; 19	1050; 50	550; 50
Ober et al. <sup>69</sup>	E	6124; 718	13.28	5025; 25	5151; 51	5376; 256	5643; 513
Castro-Alonso et al. <sup>70</sup>	E	2351; 217	10.17	2010; 10	2020; 20	2121; 101	2222; 202
Bourke et al. <sup>71</sup>	D	9308; 158	1.73	8643; 43	8686; 86	3150; 150	1650; 150
Karabinski et al. <sup>72</sup>	A	1840; 70	3.95	1608; 8	1616; 16	1386; 66	726; 66
Estevez Cores et al. <sup>73</sup>	A	1784; 227	14.58	1407; 7	1414; 14	1533; 73	1617; 147
Schindler et al. <sup>74</sup>	S	2272; 414	22.28	1608; 8	1717; 17	1848; 88	1936; 176
Endendijk et al. <sup>75</sup>	S	1271; 274	27.48	804; 4	909; 9	987; 47	1034; 94
Dailey and Bergelson <sup>76</sup>	D	4992; 203	4.24	4422; 22	4545; 45	4032; 192	2112; 192
Woods et al. <sup>77</sup>	A	5955; 265	4.66	5427; 27	5454; 54	5271; 251	2761; 251
Leijten et al. <sup>78</sup>	C	4382; 262	6.36	3819; 19	3939; 39	4095; 195	2728; 248
Zaneva et al. <sup>79</sup>	D	7266; 89	1.24	6834; 34	6868; 68	1764; 84	924; 84
Min.		1271; 53	1.24	804; 4	909; 9	987; 47	550; 50
Max.		13,531; 1507	57.55	12,261; 61	12,423; 123	10,836; 516	8624; 784
First quartile		1875; 158	4.19	1608; 8	1616; 16	1533; 73	1045; 95
Median		2444; 227	7.88	2010; 10	2121; 21	2121; 101	1936; 176
Third quartile		4396.95; 341.10	11.66	3771.14; 18.76	3847.62; 38.10	3262.00; 155.33	2542.05; 231.10

Note: N = total number of abstracts;  $n_{relevant}$  = number of relevant abstracts; D. = research domain; ratio = prevalence ratio; D = developmental, C = clinical; S = social; E = educational; A = applied psychology.

### 3.3.1 | Implementation of the stopping rules

We implemented the data-driven heuristic,<sup>24</sup> a prevalence estimation method,<sup>6</sup> and the knee method<sup>22</sup> as stopping rules, following the recommendations of the authors who introduced them. However, because the knee method required a minimum of two relevant abstracts to work, we paired each stopping rule with the rule that at least two relevant abstracts needed to be found. Thus, the worst performance was indicated by finding at least two relevant articles. However, apart from this specification, we followed the implementation of the stopping rules as recommended. For the prevalence estimation technique,<sup>6</sup> we randomly sampled 10% of all

abstracts from the unordered RAACs. On the basis of these 10%, we calculated the prevalence of relevant abstracts and multiplied it by 0.95. The estimate was consistently rounded off. The screening stopped once the number of relevant abstracts was equal to or greater than the estimate. Considering an abstract collection of 3800 abstracts, 380 abstracts were screened randomly. If, for example, 10 of these 380 abstracts were marked as relevant, the prevalence ratio would have been 2.63%. Consequently, we would have estimated that  $(3,800 - 380) \cdot 0.0263 \cdot 0.95 \approx 85$  of the unscreened abstracts were relevant. Thus, the screening would have stopped after identifying 85 relevant abstracts. If the random sampling resulted in finding zero relevant abstracts,

we randomly sampled a prevalence to simulate a guessing researcher. We drew this sample from a truncated normal distribution with a lower bound of 0.01%, upper bound of 1%, mean of 0.5%, and standard deviation of 0.1%, using the R package *truncnorm*.<sup>82</sup>

We stopped for the data-driven heuristic<sup>24</sup> after 2.5% of all abstracts had been marked as irrelevant in a row. We decided to use the percentage corresponding to the number of consecutive irrelevant abstracts, as recommended by Ros et al.<sup>24</sup> As Ros and colleagues used a dataset with 1939 abstracts, 50 consecutive, irrelevant abstracts reflect roughly 2.5%. Because our OACs had large variation in the total number of abstracts, we used percentages instead of fixed numbers. Thus, considering an abstract collection with 3800 abstracts, we stopped the screening after  $3,800 \cdot 0.025 = 95$  consecutive irrelevant abstracts.

To implement the knee method,<sup>22</sup> we adopted the code provided by van Haastrecht.<sup>49</sup> This algorithm calculates the slope ratio after each newly labeled abstract. Two parameters had to be specified: the number of relevant articles identified before the algorithm starts and the slope ratio, which needs to be attained to stop the screening. We defined the algorithm as starting after finding two relevant articles. Furthermore, the screening was stopped once the slope ratio was equal to or greater than 6, or all abstracts were screened. The slope ratio was calculated as shown in Equation 2.

### 3.3.2 | Performance measures

The performance of each stopping rule was assessed in three different ways. Once a screening stops, the abstracts could be divided into a screened and unscreened section. We labeled all screened relevant abstracts *true positives* (TP) and all screened but irrelevant abstracts *false positive* (FP). Similarly, we labeled all relevant but unseen abstracts *false negative* (FN) and all unseen irrelevant abstracts *true negative* (TN). Performance measures were the sensitivity (often referred to as recall), specificity, and cost:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad (3)$$

$$\text{Specificity} = \frac{TN}{FP + TN}, \quad (4)$$

$$\text{Cost} = \frac{TP + FP}{TP + FP + TN + FN} \quad (5)$$

We chose to introduce performance measures that are familiar to psychologists for the sake of clarity and ease of understanding.

## 3.4 | Statistical analysis

To analyze the results of each performance measure (sensitivity, specificity, cost) separately, we conducted three  $9$  (learning algorithms)  $\times 4$  (prevalence ratios)  $\times 3$  (stopping rules)  $\times 21$  (OACs) mixed-effects ANOVAs with prevalence ratios as between-subject and stopping rules and learning algorithms as within-subject factors. As the RAACs were the unit of analysis and to account for the fact that multiple RAACs were derived from the same OAC, we incorporated OAC as a fixed effect in our model. Consequently, our model specification included 12 parameter estimates: the main effect of stopping rule (1), learning algorithm (2), prevalence ratio (3), OAC (4), the two-way interactions stopping rule  $\times$  prevalence ratio (5), stopping rule  $\times$  learning algorithm (6), learning algorithm  $\times$  prevalence ratio (7), stopping rule  $\times$  OAC (8), and learning algorithm  $\times$  OAC (9), the three-way interactions stopping rule  $\times$  learning algorithm  $\times$  prevalence ratio (10), stopping rule  $\times$  learning algorithm  $\times$  OAC (11), and the residual variance (12). Our experimental design consisted of 108 groups, each comprising 21,000 observations. When including the control factor OAC, we had 2268 groups with 1000 observations each. All calculations were done in R,<sup>80</sup> and we conducted the ANOVAs using the R package *afex*.<sup>83</sup> In line with the explorative nature of this study, we examined the main and interaction effects. However, no post-hoc tests were computed. Instead, we provide bar plots with means and bootstrapped confidence intervals for each effect that exceeded an effect size of  $\eta^2 \geq .01$ . The effect sizes were interpreted in line with J. Cohen<sup>84</sup> as  $\eta^2 = .01$  (small),  $\eta^2 = .06$  (medium), and  $\eta^2 = .14$  (large).

Outliers were detected using the R package *rstatix*.<sup>85</sup> This package detects outliers as data points outside  $Q_{25\%} - 1.5 \cdot IQR$  and  $Q_{75\%} + 1.5 \cdot IQR$ , respectively. Extreme outliers are detected by multiplying the Interquartile range by a factor of 3. We checked the assumption of normality using QQ-Plots constructed using the R package *ggpubr*.<sup>86</sup> Finally, the assumption of sphericity was checked using the *afex* R package.<sup>83</sup> If the assumption was violated, we applied a Greenhouse–Geisser correction.<sup>87</sup> The testing for statistical significance was conducted based on a significance level of  $\alpha = 0.05$ .

## 4 | RESULTS

### 4.1 | Descriptive statistics

The literature search resulted in eligible abstract collections from 21 meta-analyses, as detailed in Table 1. Originally, the number of abstracts within the 21 received

abstract collections ranged from 1271 to 13,531, with a median of  $Mdn = 2,444$  ( $IQR = 2,522$ ). The number of relevant abstracts ranged from 53 to 1507, with a median of  $Mdn = 227$  ( $IQR = 183$ ). The original prevalence ratios ranged from 1.24 to 57.55%, with a median of  $Mdn = 7.88\%$  ( $IQR = 7.47\%$ ). It is important to note that the number of abstracts and the prevalence of relevant abstracts were based on the data we received. However, we manipulated the OACs so that their sample size and number of relevant studies mirrored the values of the AACs presented in Table 1.

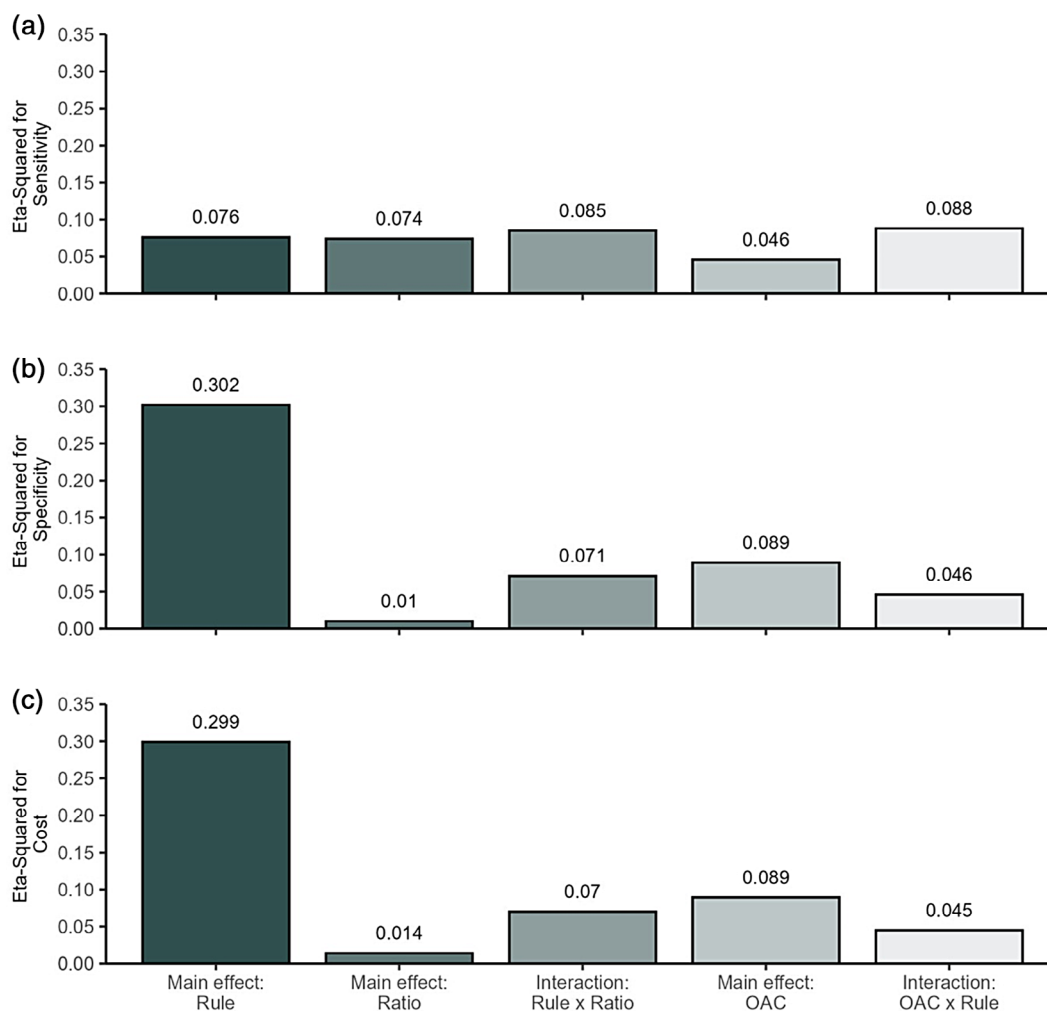
## 4.2 | Main analyses

Testing for outliers, sphericity, and normal distribution revealed that all assumptions of the mixed-effects ANOVAs had been violated (see Supporting Information for more details). However, given that some evidence

suggests the robustness of mixed-effects linear models against violations of their assumptions,<sup>88</sup> we continued computing the ANOVAs for sensitivity, specificity, and cost (Supporting Information, Tables S1, S2, and S3, respectively). All effect sizes ( $\eta^2$ ) larger than or equal to 0.01 are summarized in Figure 1.

### 4.2.1 | Sensitivity

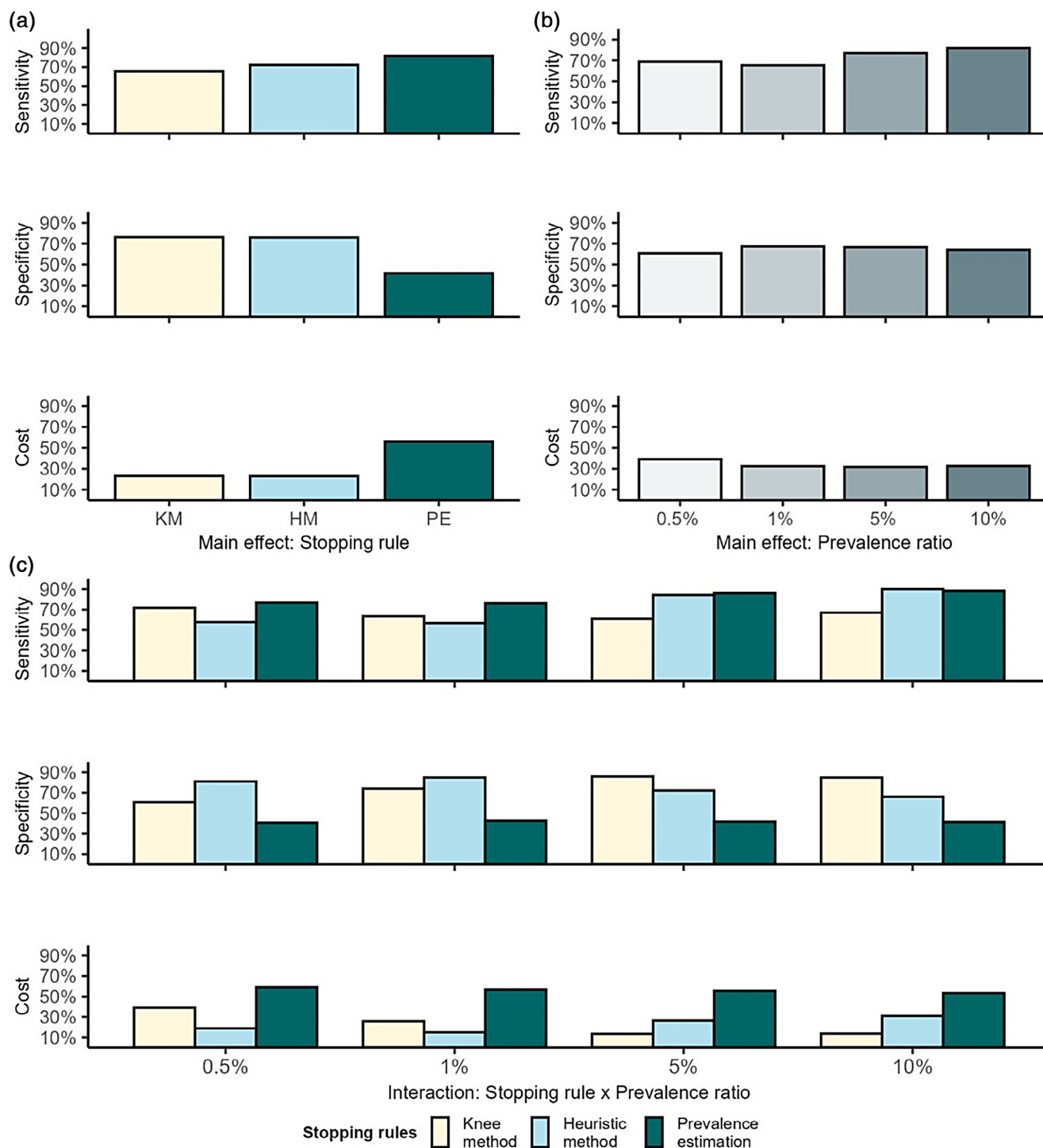
The mixed-effects ANOVA for sensitivity revealed considerable variation in effect sizes (Supporting Information, Table S1). All means and 95% CIs concerning the design factors stopping rule, learning algorithm, and prevalence estimation are presented in Table A1 (Appendix A). The main effect of the learning algorithm and all interactions involving the learning algorithm were negligible ( $\eta^2 < 0.01$ ). Means and 95% CIs regarding the interaction between the stopping rule and the learning algorithm are



**FIGURE 1** Bar-plot pertaining to the effect sizes from the mixed-effects ANOVAs. Effect sizes of factors and interactions in  $\eta^2$  for (a) sensitivity, (b) specificity, and (c) screening cost. Only effect sizes  $\eta^2 \geq 0.01$  are visualized. Rule = Stopping rule, Ratio = Prevalence ratio.

provided in Table S4 (Supporting Information). In contrast, we observed a medium-strong main effect of the stopping rule ( $F(1.81, 151584.57) = 38921.78$ ,  $p < 0.001$ ,  $\eta^2 = 0.076$ ). As Figure 2a illustrates, the prevalence estimation stopping rule achieved the highest

sensitivity ( $M = 81.87$ , 95% CI [81.82, 81.93]), followed by the data-driven heuristic stopping rule ( $M = 72.27$ , 95% CI [72.21, 72.32]). The knee method performed least favorably  $M = 65.71$ , 95% CI [65.64, 65.78]). The main effect of the prevalence ratio also exhibited a medium



**FIGURE 2** Bar-plot pertaining to the main effects and interaction of stopping rule and prevalence ratio. This figure shows the (a) main effect of stopping rule, (b) main effect of prevalence ratio, and (c) interaction of stopping rule and prevalence ratio, separately for sensitivity, specificity, and cost. As the CIs were too small to be visible, we did not include them in the figure. HM, heuristic method; KM, knee method; PE, prevalence estimation.

effect size ( $F(3,83976) = 17214.32$ ,  $p < 0.001$ ,  $\eta^2 = 0.074$ ). The 5% ( $M = 77.21$ , 95% CI [77.15, 77.28]) and 10% ( $M = 81.75$ , 95% CI [81.69, 81.81]) prevalence ratio conditions outperformed the 0.5% ( $M = 68.77$ , 95% CI [68.69, 68.85]) and 1% ( $M = 65.40$ , 95% CI [65.32, 65.48]) conditions. In addition, we observed a moderate interaction between the prevalence ratio and the stopping rule, as shown in Figure 2c and Table A2 (Appendix A). The knee method achieved higher sensitivity values than the heuristic stopping rule in the low-prevalence ratio conditions (0.5 and 1%). Conversely, the heuristic stopping rule outperformed the knee method in high-prevalence ratio conditions (5 and 10%). However, the prevalence estimation stopping rule outperformed the other rules in all but the 10% ratio condition, performing equally well as the data-driven heuristic.

#### 4.2.2 | Specificity

The mixed-effects ANOVA focusing on specificity revealed all effects to be significant. As for sensitivity, all effects involving the learning algorithm were below 0.01. The respective ANOVA results are summarized in Table S2 (Supporting Information). Means and 95% CIs regarding the main effects of the design factors stopping rule, learning algorithm, and prevalence ratio are presented in Table A1 (Appendix A). Means and 95% CIs regarding the interaction between the stopping rule and learning algorithm are summarized in Table S4 (Supporting Information). The main effect of the stopping rule was substantial ( $F(1.74, 146384.54) = 1503.49$ ,  $p < 0.001$ ,  $\eta^2 = 0.302$ ), with the knee method ( $M = 76.40$ , 95% CI [76.35, 76.46]) demonstrating a slight advantage over the heuristic stopping rule ( $M = 76.08$ , 95% CI [76.05, 76.12]). Conversely, the prevalence estimation stopping rule ( $M = 41.69$ , 95% CI [41.61, 41.78]) exhibited poorer performance than the other two methods (see Figure 2a). The main effect of the prevalence ratio on specificity was small ( $F(3, 83976) = 2529.87$ ,  $p < 0.001$ ,  $\eta^2 = 0.010$ ). As for sensitivity, higher prevalence ratios were associated with slightly improved specificity values (see Figure 2b). Examining the interaction between the prevalence ratio and stopping rule revealed that the knee method's performance improved with increasing prevalence ratios, whereas the heuristic stopping rule's performance declined as prevalence ratios increased (see Table A2, Appendix A). In contrast, the prevalence estimation method maintained consistent performance across all conditions (see Figure 2c).

#### 4.2.3 | Cost

The mixed-effects ANOVA for the screening cost, conducted with a design of 9 (learning algorithms)  $\times$  4 (prevalence ratios)  $\times$  3 (stopping rules)  $\times$  21 (OACs), also revealed significant effects (see Table S3, Supporting Information). The bootstrapped means and 95% CIs regarding the main effects of the design factors stopping rule, prevalence ratio, and learning algorithm are summarized in Table A1 (Appendix A). As with the other two performance measures, all effects associated with the learning algorithm were negligible ( $\eta^2 < 0.01$ ). In contrast, the stopping rule exhibited a substantial effect ( $F(1.75, 146778.73) = 229568.29$ ,  $p < 0.001$ ,  $\eta^2 = 0.299$ ), with the heuristic stopping rule ( $M = 22.76$ , 95% CI [22.73, 22.80]) performing slightly better than the knee method ( $M = 22.99$ , 95% CI [22.94, 23.04]), and both outperforming the prevalence estimation ( $M = 56.10$ , 95% CI [56.01, 56.17]). On average, the prevalence estimation required more than twice as many abstracts to screen compared to the other methods (see Figure 2a).

The main effect of the prevalence ratio on screening cost was small ( $F(3, 83976) = 3554.08$ ,  $p < 0.001$ ,  $\eta^2 = 0.014$ ). Higher prevalence ratios were associated with slightly lower screening costs. However, this effect was inconsistent across all conditions (see Figure 2b). Furthermore, we observed an interaction between the prevalence ratio and stopping rule ( $F(5.24, 146778.73) = 13493.17$ ,  $p < 0.001$ ,  $\eta^2 = 0.070$ ). While the effect of the prevalence estimation remained consistent across different prevalence ratios, the knee method's performance improved with increasing prevalence ratios. The performance of the heuristic stopping rule decreased as the prevalence of relevant abstracts increased (see Figure 2c). The respective means and CIs are presented in Table A2 (Appendix A).

#### 4.3 | Secondary analyses

Beyond the experimental design factors, the control factor OAC exhibited a small-sized main effect for sensitivity ( $\eta^2 = 0.046$ ), a medium-sized effect for specificity ( $\eta^2 = 0.089$ ), and screening cost ( $\eta^2 = 0.089$ ; see Figure B2, Appendix B). All means and 95% CIs for the main effect of OAC are presented in Table S5 (Supporting Information) and visualized in Figure B2 (Appendix B). Moreover, the interaction between abstract collection and stopping rule resulted in the largest observed effect on sensitivity ( $\eta^2 = 0.088$ ) and considerable effects on specificity ( $\eta^2 = 0.046$ ) and cost ( $\eta^2 = 0.045$ ). As displayed in Figure B2 (Appendix B), the

performance of all three stopping rules varied across abstract collections. All means and 95% CIs regarding this interaction are presented in Table S6 (Supporting Information). We further explored this interaction by visually inspecting how the stopping rules performed for groups of abstract collections with diverging numbers of relevant abstracts. As shown in Figure B3 (Appendix B), the knee method performed poorly in terms of sensitivity for abstract collections with more than 200 relevant abstracts, while both the prevalence estimation and the heuristic stopping rule performed well for these abstract collections. For specificity and cost, this pattern reversed (see Figure B3, Appendix B).

As the data for each condition did not appear to follow a normal distribution, we calculated the 25, 50, and 75% quantiles for each combination of the design factors. In Table 2, we compiled a summary highlighting the top-performing combinations based on these statistics. A comprehensive list, including the performance of all combinations and additional statistics, is accessible within our repository (<https://osf.io/7yhrq>). When focusing solely on sensitivity, the prevalence estimation paired with the LR + SBERT learning algorithm outperformed the other combinations for the 0.5, 1, and 5% ratio conditions but not for the 10% ratio condition, where the combination of the data-driven heuristic and the LR + SBERT learning algorithm performed best. It is important to highlight that, for the prevalence estimation, all learning algorithms demonstrated similar sensitivity levels while displaying variations in terms of specificity and screening cost. However, as displayed in the quantiles and minimum values in Table 2, the prevalence estimation was associated with a high risk of over-sampling and undersampling, resulting in greater sensitivity, specificity, and screening cost variance. While the knee method also carries this risk, it was much less pronounced.

Furthermore, the prevalence estimation stopping rule consistently results in the highest screening cost. Thus, when considering all quality measures, we identified different top performers. The knee method paired with the LR + doc2vec learning algorithm outperformed the other stopping rules for the 0.5 and 1% ratio conditions. Note that while the LR + doc2vec learning algorithm achieved approximately 10% higher sensitivity values across all prevalence conditions compared to when combined with the other algorithms, it also required screening considerably more abstracts. However, for the 5 and 10% ratio conditions, the data-driven heuristic paired with the LR + SBERT learning algorithm outperformed the other stopping rule and learning algorithm combinations.

## 5 | DISCUSSION

In this study, we assessed the performance of three stopping rules employed in AI-AS tools for concluding the abstract screening process. These stopping rules comprised the data-driven heuristic,<sup>24</sup> a prevalence estimation method,<sup>6</sup> and the knee method,<sup>22</sup> each utilizing distinct techniques to determine when to halt the screening process. Our evaluation was conducted using abstract collections obtained from five distinct research domains within the field of psychology: Applied Psychology, Clinical Psychology, Developmental Psychology, Educational Psychology, and Social Psychology. We systematically manipulated the prevalence of relevant abstracts for each domain, representing prevalence ratios of 0.5, 1, 5, and 10%. With this extensive dataset, we proceeded to simulate the practical utility of stopping rules. Employing nine different learning algorithms, we conducted nine simulations for each manipulated abstract collection within the AI-AS tool ASReview.<sup>19</sup> Subsequently, we applied the stopping rules in each simulation run. To evaluate performance, we considered sensitivity (the proportion of relevant abstracts among all screened abstracts), specificity (the proportion of irrelevant articles among all unscreened articles), and screening cost (the proportion of articles screened until the stopping rule criteria were met).

### 5.1 | Performance results

The results of our analysis revealed notable performance differences between the three stopping rules (i.e., prevalence estimation, knee method, and data-driven heuristic). On average, the stopping rule that relied on a prevalence estimate<sup>6</sup> performed best in terms of sensitivity. However, this stopping rule was associated with significantly increased screening costs and reduced specificity compared to the knee method<sup>22</sup> and the data-driven heuristic.<sup>24</sup> Furthermore, the prevalence estimation carries the risk of both over-sampling and under-sampling.<sup>53</sup> Regarding the prevalence of relevant abstracts, we observed a moderately strong main effect, indicating that higher prevalence ratios were linked to higher sensitivity and lower specificity and screening cost. However, we also observed a moderately strong interaction between the prevalence ratio and the stopping rule, suggesting that meta-analysts should estimate the prevalence of relevant literature and customize their stopping rules accordingly (see Figure 3). The effect of the learning algorithm and all interactions involving the learning algorithm appeared negligible at first glance.

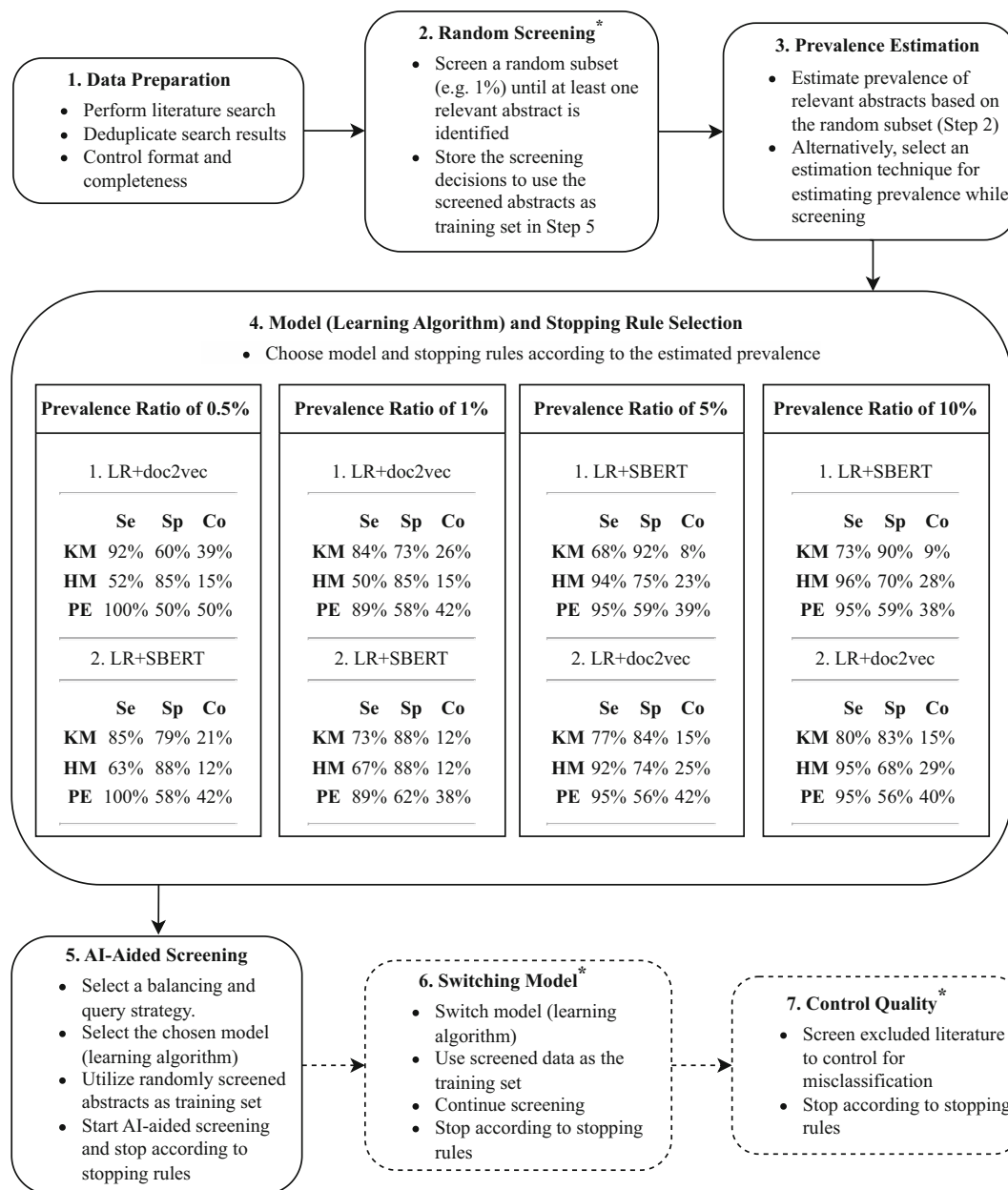
TABLE 2 Descriptives of the top-performing stopping rule × learning algorithm combination.

Prevalence ratio	Stopping rule	Learning algorithm	Min.	Max.	Median	Q <sub>25%</sub>	Q <sub>75%</sub>	IQR
<i>Sensitivity</i>								
0.5%	HM	LR + SBERT	7.14	100.00	62.50	42.86	88.88	46.03
	KM	LR + doc2vec	5.00	100.00	91.67	66.67	100.00	33.33
	PE	LR + SBERT	5.00	100.00	100.00	50.00	100.00	50.00
1%	HM	LR + SBERT	3.90	100.00	66.67	37.50	88.26	50.74
	KM	LR + doc2vec	2.46	100.00	84.09	59.70	95.00	35.30
	PE	LR + SBERT	3.90	100.00	89.41	56.00	100.00	44.00
5%	HM	LR + SBERT	1.18	100.00	93.88	85.46	97.37	11.91
	KM	LR + doc2vec	0.58	100.00	77.18	55.69	89.09	33.41
	PE	LR + SBERT	4.17	100.00	94.76	77.55	100.00	22.45
10%	HM	LR + SBERT	1.49	100.00	96.00	91.53	98.63	7.11
	KM	LR + doc2vec	0.38	100.00	80.00	60.57	90.32	29.75
	PE	LR + TFIDF	4.62	100.00	95.03	80.93	100.00	19.07
<i>Specificity</i>								
0.5%	HM	LR + SBERT	0.13	97.50	88.04	80.74	91.86	11.12
	KM	LR + doc2vec	0.00	99.94	60.49	24.30	83.99	59.69
	PE	LR + SBERT	0.00	89.99	58.45	0.00	81.38	81.38
1%	HM	LR + SBERT	16.35	97.48	87.88	81.09	91.49	10.40
	KM	LR + doc2vec	0.03	99.90	73.42	51.50	89.52	38.02
	PE	LR + SBERT	0.00	89.93	61.95	0.00	82.41	82.41
5%	HM	LR + SBERT	15.27	97.10	75.49	67.39	85.22	17.83
	KM	LR + doc2vec	0.11	99.91	84.39	75.29	92.82	17.53
	PE	LR + SBERT	0.00	89.61	59.07	0.00	78.68	78.68
10%	HM	LR + SBERT	10.55	96.99	69.68	60.35	81.67	21.32
	KM	LR + doc2vec	19.49	99.89	83.12	75.14	90.99	15.84
	PE	LR + TFIDF	8.64	97.19	69.56	61.06	80.06	19.00
<i>Cost</i>								
0.5%	HM	LR + SBERT	2.49	99.50	11.90	8.10	19.16	11.06
	KM	LR + doc2vec	0.06	99.63	39.32	15.94	75.36	59.42
	PE	LR + SBERT	9.96	99.63	41.36	18.54	99.53	80.99
1%	HM	LR + SBERT	2.50	82.91	12.01	8.43	18.73	10.30
	KM	LR + doc2vec	0.10	99.12	26.33	10.38	48.06	37.68
	PE	LR + SBERT	9.98	99.12	37.70	17.42	99.04	81.61
5%	HM	LR + SBERT	2.77	80.71	23.36	14.09	31.07	16.98
	KM	LR + doc2vec	0.08	95.33	14.87	6.84	23.54	16.70
	PE	LR + SBERT	9.90	95.33	39.00	20.32	95.27	74.94
10%	HM	LR + SBERT	2.74	81.33	27.56	16.68	36.06	19.38
	KM	LR + doc2vec	0.10	73.25	15.36	8.20	22.61	14.41
	PE	LR + TFIDF	9.53	91.06	40.13	21.15	90.95	69.80

Note: The order is based on best median sensitivity. In case of equal median performance, the order is based on the 25% quantile. Each statistic comprises 21,000 data points.

Abbreviations: IQR, interquartile range; Q, quantile.





**FIGURE 3** Flowchart pertaining nine steps of AI-aided screening with ASReview. Performance estimates reflect median performance. Literature suggestions for each step are provided within this work. \* = Adapted from the SAFE method proposed by Boetje and van de Schoot.<sup>40</sup> Se = Sensitivity, Sp = Specificity, Co = Cost, KM = Knee Method with rho equals 6,<sup>22</sup> HM = Data-driven Heuristic with 2.5% consecutive irrelevant abstracts,<sup>11</sup> PE = Prevalence Estimate based on a random sample of 10%.<sup>6</sup>

Nevertheless, a closer examination of the data revealed combinations of stopping rules and learning algorithms that outperformed others. For instance, in the 0.5% prevalence ratio condition, the knee method paired with the LR + doc2vec learning algorithm achieved a median sensitivity of 92%, while the second-best combination of this method with the LR + SBERT learning algorithm achieved a median sensitivity of 85% but also halved the screening cost. It is worth noting that, to the best of our knowledge, only ASReview incorporates all the learning algorithms utilized in this study. Nonetheless, several AI-

AS tools incorporate some of the learning algorithms used in this study (for an overview of AI-AS tools, see Reference 15). Besides these findings, we also advocate for considering several performance measures when selecting stopping rules and learning algorithms. While the prevalence estimation constantly achieved similar sensitivity across learning algorithms, we observed that combining this method with the LR + SBERT learning algorithm resulted in a lower screening cost. Furthermore, in the 1% prevalence ratio condition, the prevalence estimation was the only stopping rule that achieved a median

sensitivity of around 90%. However, when paired with the LR + doc2vec learning algorithm, the knee method offered a competitive alternative with a median sensitivity of around 84%, exhibiting less variance in sensitivity estimates and a lower screening cost. Similarly, when evaluating all performance measures, the combination of the data-driven heuristic with the LR + SBERT learning algorithm outperformed both methods in the 5 and 10% prevalence conditions. In both conditions, this combination achieved a median sensitivity above 93%.

In addition to these findings, our data incorporates several noteworthy observations, particularly compared to previous research in the field. For instance, the relatively minor influence of the learning algorithm found in our study contradicted recent findings.<sup>26</sup> One possible explanation for these differing results could be disparities in study design. Campos et al. conducted their simulations using the same abstracts to train the algorithms, while we replicated all simulation runs using different abstracts for training. As Teijema et al.<sup>16</sup> have demonstrated that the screening order of abstracts can vary among different learning algorithms, we posit that this variation could impact the performance of stopping rules reliant on the screening order. Consequently, when averaging across simulation runs with different screening orders, the effect of the learning algorithm may have been diminished or averaged out.

Furthermore, a previous study found that configuring the data-driven heuristic to halt after encountering 7% consecutive irrelevant abstracts resulted in sensitivity exceeding 95% across abstract collections with different prevalence ratios.<sup>26</sup> However, this modification also led to screening of approximately 70% of the abstracts. Our results suggest that comparable sensitivity can be achieved with reduced screening costs when the stopping rules are tailored to the prevalence of relevant abstracts. In the 10% prevalence ratio condition, for example, stopping the screening process after encountering 2.5% consecutive irrelevant records resulted in a median sensitivity of 95% with a screening cost below 30% when the LR + SBERT algorithm was employed. Additionally, we noticed that the knee method outperformed the data-driven heuristic for specific abstract collections and vice versa. This suggests that combining both stopping rules may potentially enhance overall sensitivity, which aligns with the findings of Campos et al.,<sup>26</sup> who observed that combining the data-driven heuristic with the time-based heuristic can reduce screening costs. However, based on our results, we believe that the selection of stopping rule combinations should also consider the prevalence ratio and learning algorithm to further improve sensitivity and reduce screening cost. Finally, upon closer examination, we identified interactions between the performance of the stopping rules and the number of relevant studies

within an abstract collection (see Figure 2). The prevalence estimation procedure and the heuristic stopping rule demonstrated superior performance for collections with many relevant abstracts, while the knee method's performance notably declined in such collections. Notably, our manipulation design did not allow for a clear differentiation between the number of relevant abstracts and the size of the abstract collection; both factors could have contributed to this finding. Nevertheless, comparing the variation in sample size and the number of relevant studies between conditions revealed much more variation in the number of relevant studies than in the sample sizes. However, the lower performance of the knee method for high prevalence ratio conditions and abstract collections with many relevant abstracts is particularly intriguing. This finding contrasts with the recommendation of Cormack and Grossman,<sup>22</sup> who suggested using the same cut-off value we employed for abstract collections with a high prevalence of relevant abstracts. Furthermore, they proposed adjusting the cut-off upward for abstract collections with low prevalence. Our findings suggest the opposite. Differences in the type and number of text documents could be one potential explanation for these inconsistent results. The highest number of abstracts in our psychological-related abstract collections reached 12,432, while the smallest text collections in Cormack and Grossman<sup>22</sup> included three times as many text documents.

## 5.2 | Limitations and future research directions

Our study has several limitations. First, while this study may be among the first to evaluate the performance of various stopping rules separate for different performance ratios, the range of prevalence ratios was not exhaustive. Future investigations could explore, for example, whether there are ceiling effects by considering prevalence ratios above 10%.

Second, our study implies that the characteristics of the abstract collection, such as sample size or the number of relevant studies, influences the performance of the stopping rules. To gain a deeper understanding of the performance variations among different abstract collections and stopping rules, researchers could consider manipulating the sample size of the abstract collections along with the prevalence ratio. However, human error could be an alternative explanation for the discrepancies in the performance of stopping rules across abstract collections. Given that humans typically misclassify between 5 and 10% of relevant abstracts, the established baseline of relevant abstracts in the collected collections might not be entirely accurate. Consequently, the performance of the learning algorithms

could have been affected, which in turn could have impacted the performance of the stopping rules.

Third, we mitigated any potential impact of the abstracts used to train the learning algorithm by averaging it out. We also trained the learning algorithms with only one relevant and one irrelevant abstract. Future research should investigate whether specific procedures to select training studies, such as sampling 1% at random<sup>40</sup> or the sheer number of training studies, impact the performance of the stopping rules and learning algorithms.

Fourth, the screening cost and specificity associated with the prevalence estimation technique may differ in a real-world scenario. We calculated the prevalence estimate after simulating the AI-AS. Consequently, the randomly screened abstracts were not used as training studies, which would likely have resulted in higher specificity and lower screening costs. However, this design decision did not impact sensitivity. Whereas this design decision was not optimal, it effectively reduced the computation time of our simulation by 50%, allowing us to evaluate the performance of all stopping rules on the same simulated datasets. Future research could investigate the cost and specificity of the prevalence estimation procedure of van Haastrecht et al.<sup>6</sup> when using the randomly screened abstracts as training studies.

Fifth, to keep our simulation within a manageable computation time, we defined that the ranking of unseen abstracts was recalculated after every 10 screened abstracts—this setup could have had a minor influence on the performance of the stopping rules. Moreover, our results are contingent on the default balancing and query strategy settings of ASReview.<sup>19</sup> Other balancing and query strategies are likely to impact the performance of the stopping rules. Future research could investigate the impact of these design decisions.

Sixth, the three here tested stopping rules are not exhaustive. Each can be implemented differently by altering cut-off values. Other stopping rules, implementation, or combinations might improve sensitivity while lowering cost. We strongly encourage future investigations to delve deeper into identifying implementations and combinations of stopping rules in accordance with specific prevalences of relevant literature. Such research endeavors can pave the way for identifying the most efficient and dependable stopping rules tailored to the abstract collection's characteristics and research objectives.

Finally, machine learning algorithms are susceptible to various biases (see Reference 89). Whereas some biases do not apply to AI-AS, others are applicable when using AI-AS tools<sup>8</sup> but are precluded by our simulation design—such as representation bias, which occurs when the order of abstracts is biased due to nonrandom training studies. Still, other biases might be relevant. The

behavioral bias, for instance, could apply, given that different researchers screened the abstract collections, potentially with varying degrees of freedom. Similarly, user interaction bias might occur if the data contain misclassified abstracts. Furthermore, we cannot generalize our results to other research areas, thus, making our design susceptible to population bias for machine learning algorithms (see Reference 89).

### 5.3 | Conclusions and practical recommendations

Taken together, our study provides valuable insights into the performance of three distinct stopping rules for AI-assisted abstract screening: the prevalence estimation by van Haastrecht,<sup>49</sup> the knee method,<sup>22</sup> and the data-driven heuristic.<sup>24</sup> Additionally, our novel approach to manipulating the prevalence of relevant abstracts underscores the importance of considering prevalence when selecting stopping rules and learning algorithms. Integrating our findings with previous research, we recommend a seven-step process for conducting research synthesis with AI-AS tools (see Figure 3).

Step 1 involves conducting a literature search and preparing the data. Thereby, state-of-the-art tools, such as Paperfletcher<sup>10</sup> and Citationchase,<sup>9</sup> which automate the backward and forward search to prevent missing relevant articles, could be used to identify additional relevant literature. The search results should then be downloaded, combined, deduplicated, and stored in a format compatible with the AI-AS tool.<sup>8</sup> In Step 2, following the suggestions of Boetje and van de Schoot<sup>40</sup> and van Haastrecht et al.,<sup>6</sup> we recommend randomly screening a predefined percentage of the abstracts until at least one relevant abstract is identified. The abstracts of this randomly screened subset can be used to train the learning algorithm in the AI-AS phase, preventing bias due to the similarity of abstracts from studies already known to be relevant before conducting the literature search (see Boetje and van de Schoot,<sup>40</sup> for additional information). In Step 3, we recommend using the randomly screened abstracts to estimate the prevalence, as described in Equation 1, which is adapted from van Haastrecht et al.<sup>6</sup> Alternatively, in this step users could select another prevalence estimation technique that estimates prevalence while screening, such as the one provided in the AI-AS tool SWIFT-Active Screener.<sup>23</sup>

In Step 4, an appropriate combination of the learning algorithm and stopping rule should be selected according to the estimated prevalence. Our results, summarized in Figure 3, reflect the best combinations of learning algorithms and stopping rules for different prevalence ratios. In addition, we recommend the following: Users should

avoid relying solely on the prevalence estimation technique to mitigate the risk of oversampling. Instead, combine the knee method and the data-driven heuristic, prioritizing the knee method for prevalence ratios of 1% or lower and the data-driven heuristic for prevalence ratios of 5–10%. To implement the knee method in *R*, users can adapt the code presented in our repository (<https://osf.io/7yhrq>). For an implementation in Python, users can use the code provided by van Haastrecht.<sup>49</sup> Additionally, users might want to incorporate other stopping rules, such as identifying key studies (see Boetje and van de Schoot,<sup>40</sup> for additional information) or using the time-based heuristic (<sup>11</sup>; see also Reference 26). For the latter, an appropriate cut-off value could be reflected by the screening cost of the prevalence estimation method reported in Figure 3 (see our OSF repository at <https://osf.io/7yhrq> for information on the performance of additional learning algorithms).

In Step 5, we outline how to set up the AI-AS when using ASReview.<sup>19</sup> For instance, a query and balancing strategy needs to be selected (see Supporting Information). As most evaluation studies and our results are based on the default strategies in ASReview (see Reference 36), we recommend users adhere to this setting when using this tool. Furthermore, while we selected only one relevant and one irrelevant abstract to simulate AI-AS, users might want to use the abstracts screened in Step 2 to train the algorithm. While we cannot rule out that this difference in training data impacts the performance of stopping rules, results from Oude Wolcherink et al.<sup>56</sup> suggested that the median performance of the data-driven heuristic remains similar or increases with more training studies. However, after the setup, the AI-AS can be started and stopped after the predefined stopping rules are met.

To enhance the quality of the AI-AS, users might also want to consider integrating two additional screening phases, following the suggestions of Boetje and van de Schoot<sup>40</sup> (i.e., Steps 6 and 7 in Figure 3). In Step 6, users should continue their AI-AS employing a different learning algorithm to identify additional relevant articles (see Reference 16). In this phase, all previously labeled abstracts serve as the training set, and screening can be concluded after identifying 50 consecutive irrelevant abstracts. In Step 7, all previously labeled irrelevant abstracts should be rescreened to control for misclassification. In this phase, the authors recommend using all previously labeled relevant abstracts and one randomly chosen irrelevant abstract to train the algorithm. Screening can be stopped after 50 consecutive irrelevant abstracts. However, we did not examine the impact of Steps 6 and 7. Therefore, we cannot conclude that these methods will increase sensitivity in our data.

Finally, we strongly encourage users of AI-AS tools to adhere to reporting standards.<sup>42,43</sup> We hope our findings will aid researchers in enhancing and simplifying their AI-AS procedures in the future, ultimately catalyzing scientific progress and knowledge gain.

## AUTHOR CONTRIBUTIONS

**Lars König:** Conceptualization; methodology; software; formal analysis; investigation; resources; data curation; writing – original draft; writing – review and editing; visualization; project administration; validation. **Steffen Zitzmann:** Conceptualization; writing – review and editing; supervision; project administration. **Tim Fütterer:** Conceptualization; writing – review and editing. **Diego G. Campos:** Conceptualization; writing – review and editing. **Ronny Scherer:** Conceptualization; writing – review and editing. **Martin Hecht:** Conceptualization; resources; writing – review and editing; supervision; project administration.

## ACKNOWLEDGMENTS

For correspondence regarding this article, please contact Lars König at the Helmut-Schmidt-Universität Hamburg, Department of Psychology, located at Am Stadtrand 50, 22043 Hamburg. Study data, code, and preregistration can be found on the Open Science Framework (preregistration: <https://osf.io/ucz8d>, data and code: <https://osf.io/7yhrq>). Computational resources (HPC-cluster HSUPER) were provided by the hpc.bw project, funded by dtcc.bw (Digitalization and Technology Research Center of the Bundeswehr), and further supported by the European Union—NextGenerationEU. This work is part of the Centers of Excellence scheme, funded by the Research Council of Norway, project number 331640. Diego G. Campos and Ronny Scherer were supported by the Centres of Excellence scheme, funded by the Research Council of Norway, project number 331640, and the project “Mapping of Longitudinal data of Inequalities in Education (MapIE)”, funded by the European Commission under the Horizon Europe scheme, project number 101132474.

## CONFLICT OF INTEREST STATEMENT


The authors declare that there is no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in OSF at <https://osf.io/7yhrq>.

## ORCID

Lars König  <https://orcid.org/0009-0000-3525-771X>

Steffen Zitzmann  <https://orcid.org/0000-0002-7595-4736>

Tim Fütterer  <https://orcid.org/0000-0001-5399-9557>

Diego G. Campos  <https://orcid.org/0000-0002-8820-5881>

Ronny Scherer  <https://orcid.org/0000-0003-3630-0710>

Martin Hecht  <https://orcid.org/0000-0002-5168-4911>

## REFERENCES

- Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction—the why, what, when, and how. *J Clin Epidemiol*. 2017;91:23-30. doi:10.1016/j.jclinepi.2017.08.010
- Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2):e012545. doi:10.1136/bmjopen-2016-012545
- Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8(1):163. doi:10.1186/s13643-019-1074-9
- Pham B, Bagheri E, Rios P, et al. Improving the conduct of systematic reviews: a process mining perspective. *J Clin Epidemiol*. 2018;103:101-111. doi:10.1016/j.jclinepi.2018.06.011
- Cooper HM, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed. Russell Sage Foundation; 2009.
- van Haastrecht M, Sarhan I, Yigit Ozkan B, Brinkhuis M, Spruit M. SYMBALS: a systematic review methodology blending active learning and snowballing. *Front Res Metr Anal*. 2021; 6:685591. doi:10.3389/frma.2021.685591
- Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, 1–10. 2014. doi:10.1145/2601248.2601268
- Zhang Q, Neitzel A. Choosing the right tool for the job: screening tools for systematic reviews in education. *J Res Educ Effect*. 2023;1–27:513-539. doi:10.1080/19345747.2023.2209079
- Haddaway NR, Grainger MJ, Gray CT. Citationchaser: a tool for transparent and efficient forward and backward citation chasing in systematic searching. *Res Synth Methods*. 2022;13(4): 533-545. doi:10.1002/jrsm.1563
- Pallath A, Zhang Q. *PAPERFETCHER*: a tool to automate hand-searching and citation searching for systematic reviews. *Res Synth Methods*. 2023;14(2):323-335. doi:10.1002/jrsm.1604
- Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinform*. 2010;11(1):55. doi:10.1186/1471-2105-11-55
- Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc*. 2006;13(2):206-219. doi:10.1197/jamia.M1929
- Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Syst Rev*. 2016;5(1):140. doi:10.1186/s13643-016-0315-4
- Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Syst Rev*. 2014;3(1):74. doi:10.1186/2046-4053-3-74
- Burgard T, Bittermann A. Reducing literature screening workload with machine learning: a systematic review of tools and their performance. *Z Psychol*. 2023;231(1):3-15. doi:10.1027/2151-2604/a000509
- Teijema JJ, Hofstee L, Brouwer M, et al. Active learning-based systematic reviewing using switching classification models: the case of the onset, maintenance, and relapse of depressive disorders. *Front Res Metr Anal*. 2023;8:1178181. doi:10.3389/frma.2023.1178181
- Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. *PLoS One*. 2020;15(1):e0227742. doi:10.1371/journal.pone.0227742
- Gates A, Guitard S, Pillay J, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. *Syst Rev*. 2019;8(1):278. doi:10.1186/s13643-019-1222-2
- van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell*. 2021a;3(2):125-133. doi:10.1038/s42256-020-00287-7
- Bron MP, van der Heijden PGM, Feelders AJ, Siebes APJM. Using Chao's estimator as a stopping criterion for technology-assisted review (version 1). arXiv. 2024. doi:10.48550/ARXIV.2404.01176
- Callaghan MW, Müller-Hansen F. Statistical stopping criteria for automated screening in systematic reviews. *Syst Rev*. 2020; 9(1):273. doi:10.1186/s13643-020-01521-4
- Cormack GV, Grossman MR. Engineering quality and reliability in technology-assisted review. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 75–84. 2016. doi:10.1145/2911451.2911510
- Howard BE, Phillips J, Tandon A, et al. SWIFT-active screener: accelerated document screening through active learning and integrated recall estimation. *Environ Int*. 2020;138:105623. doi:10.1016/j.envint.2020.105623
- Ros R, Bjarnason E, Runeson P. A machine learning approach for semi-automated search and selection in literature studies. Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, 118–127. 2017. doi:10.1145/3084226.3084243
- Yu Z, Menzies T. FAST2: an intelligent assistant for finding relevant papers. *Expert Syst Appl*. 2019;120:57-71. doi:10.1016/j.eswa.2018.11.021
- 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. *Educ Psychol Rev*. 2024;36(1): 19. doi:10.1007/s10648-024-09862-5
- Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag*. 1988;24(5):513-523. doi:10.1016/0306-4573(88)90021-0
- Le QV, Mikolov T. Distributed representations of sentences and documents. 2014. doi:10.48550/ARXIV.1405.4053
- Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990. 2019. doi:10.18653/v1/D19-1410
- Breiman L. Random forests. *Mach Lear*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

31. Vapnik VN. *The Nature of Statistical Learning Theory*. Springer; 1995. doi:10.1007/978-1-4757-2440-0
32. Aggarwal CC. *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing; 2018. doi:10.1007/978-3-319-94463-0
33. Cox DR. The regression analysis of binary sequences. *J R Stat Soc Series B Stat Methodology*. 1958;20(2):215-232. doi:10.1111/j.2517-6161.1958.tb00292.x
34. Gomes SR, Saroar SG, Mosfaiul M, et al. A comparative approach to email classification using Naive Bayes classifier and hidden Markov model. 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), 482–487. 2017. doi:10.1109/ICAEE.2017.8255404
35. Settles B. Active Learning Literature Survey. 2009.
36. ASReview LAB Developers. ASReview LAB software documentation. version 1.1. 2022. doi:10.5281/ZENODO.7319090
37. Ferdinands G. AI-assisted systematic reviewing: selecting studies to compare Bayesian versus frequentist SEM for small sample sizes. *Multivar Behav Res*. 2021;56(1):153-154. doi:10.1080/00273171.2020.1853501
38. Ferdinands G, Schram R, de Bruin J, et al. *Active Learning for Screening Prioritization in Systematic Reviews—A Simulation Study [Preprint]*. Open Science Framework; 2020. doi:10.31219/osf.io/w6qbg
39. Harmsen W, de Groot J, Harkema A, et al. Artificial intelligence supports literature screening in medical guideline development: towards up-to-date medical guidelines. [Preprint]. 2021. doi:10.5281/ZENODO.5031907
40. Boetje J, van de Schoot R. *The SAFE Procedure: A Practical Stopping Heuristic for Active Learning-Based Screening in Systematic Reviews and Meta-Analyses [Preprint]*. Research Square; 2023. doi:10.21203/rs.3.rs-2856011/v1
41. Neeleman R, Leenaars CHC, Oud M, Weijdemans F, van de Schoot R. Addressing the challenges of reconstructing systematic reviews datasets: a case study and a noisy label filter procedure. *Syst Rev*. 2024;13(1):69. doi:10.1186/s13643-024-02472-w
42. Cacciamani GE, Chu TN, Sanford DI, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. *Nat Med*. 2023;29(1):14-15. doi:10.1038/s41591-022-02139-w
43. Lombaers P, de Bruin J, van de Schoot R. Reproducibility and data storage for active learning-aided systematic reviews. *Appl Sci*. 2024;14(9):3842. doi:10.3390/app14093842
44. Bourke M, Haddara A, Loh A, Carson V, Breau B, Tucker P. Adherence to the World Health Organization's physical activity recommendation in preschool-aged children: a systematic review and meta-analysis of accelerometer studies. *Int J Behav Nutr Phys Act*. 2023;20(1):52. doi:10.1186/s12966-023-01450-0
45. Guan X, Feng X, Islam AYMA. The dilemma and countermeasures of educational data ethics in the age of intelligence. *Humanit Social Sci Commun*. 2023;10(1):138. doi:10.1057/s41599-023-01633-x
46. Marsili F, Pellegrini M. The relation between nominations and traditional measures in the gifted identification process: a meta-analysis. *School Psychol Int*. 2022;43(4):321-338. doi:10.1177/01430343221105398
47. Tian X, Xia Z, Xie J, Zhang C, Liu Y, Xu M. A meta-analytical review of intervention experiments to reduce food waste. *Environ Res Lett*. 2022;17(6):064041. doi:10.1088/1748-9326/ac72b6
48. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a “Knee” in a haystack: detecting knee points in system behavior. 2011 31st International Conference on Distributed Computing Systems Workshops, 166–171. 2011. doi:10.1109/ICDCSW.2011.20
49. van Haastrecht M. ASReview knee method. GitHub, GitHub Repository. 2022 <https://github.com/MaxvanHaastrecht/ASReview-Knee-Method/blob/main/Cormack%20Grossman%20Knee%20Algorithm.ipynb>
50. Hou Z, Tipton E. Enhancing recall in automated record screening: a resampling algorithm. *Res Synth Methods*. 2024;15(3):372–383. doi:10.1002/jrsm.1690
51. Kastner M, Straus SE, McKibbin KA, Goldsmith CH. The capture–mark–recapture technique can be used as a stopping rule when searching in systematic reviews. *J Clin Epidemiol*. 2009;62(2):149-157. doi:10.1016/j.jclinepi.2008.06.001
52. ASReview Lab Developers. ASReview—discussions. GitHub, as review. 2023 <https://github.com/asreview/asreview/discussions>
53. Yang E, Lewis DD, Frieder O. Heuristic stopping rules for technology-assisted review. *Proceedings of the 21st ACM Symposium on Document Engineering*, 1–10. 2021. doi:10.1145/3469096.3469873
54. Scherhag J, Burgard T. Performance of semi-automated screening using Rayyan and ASReview: a retrospective analysis of potential work reduction and different stopping rules. ZPID (Leibniz Institute for Psychology). 2023. doi:10.23668/PSYCHARCHIVES.12843
55. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—A web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. doi:10.1186/s13643-016-0384-4
56. Oude Wolcherink MJ, Pouwels XGLV, van Dijk SHB, Doggen CJM, Koffijberg H. Can artificial intelligence separate the wheat from the chaff in systematic reviews of health economic articles? *Expert Rev Pharmacoecon Outcomes Res*. 2023;23(9):1049-1056. doi:10.1080/14737167.2023.2234639
57. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-reviewer. *Syst Rev*. 2020;9(1):73. doi:10.1186/s13643-020-01324-7
58. Westgate MJ. revtools: an R package to support article screening for evidence synthesis. *Res Synth Methods*. 2019;10(4):606-614. doi:10.1002/jrsm.1374
59. Vermillet A, Tølbøll K, Litsis Mizan S, Skewes, J C, Parsons CE. Crying in the first 12 months of life: a systematic review and meta-analysis of cross-country parent-reported data and modeling of the “cry curve”. *Child Dev*. 2022;93(4):1201-1222. doi:10.1111/cdev.13760
60. Bottema-Beutel K, Crowley S, Sandbank M, Woynaroski TG. Research review: conflicts of interest (COIs) in autism early intervention research—a meta-analysis of COI influences on intervention effects. *J Child Psychol Psychiatry*. 2021;62(1):5-15. doi:10.1111/jcpp.13249
61. Khazanov GK, Morris PE, Beed A, et al. Do financial incentives increase mental health treatment engagement? A meta-analysis. *J Consult Clin Psychol*. 2022;90(6):528-544. doi:10.1037/ccp0000737
62. Reimer NK, Sengupta NK. Meta-analysis of the “ironic” effects of intergroup contact. *J Pers Soc Psychol*. 2023;124(2):362-380. doi:10.1037/pspi0000404

63. Hall C, Dahl-Leonard K, Cho E, et al. Forty years of reading intervention research for elementary students with or at risk for dyslexia: a systematic review and meta-analysis. *Reading Res Quart.* 2023;58(2):285-312. doi:10.1002/rrq.477
64. Simonsmeier BA, Flaig M, Deiglmayr A, Schalk L, Schneider M. Domain-specific prior knowledge and learning: a meta-analysis. *Educ Psychol.* 2022;57(1):31-54. doi:10.1080/00461520.2021.1939700
65. Hsieh W, Faulkner N, Wickes R. What reduces prejudice in the real world? A meta-analysis of prejudice reduction field experiments. *Br J Soc Psychol.* 2022;61(3):689-710. doi:10.1111/bjso.12509
66. Alden LE, Matthews LR, Wagner S, et al. Systematic literature review of psychological interventions for first responders. *Work Stress.* 2021;35(2):193-215. doi:10.1080/02678373.2020.1758833
67. Liu RT, Steele SJ, Hamilton JL, et al. Sleep and suicide: a systematic review and meta-analysis of longitudinal studies. *Clin Psychol Rev.* 2020;81:101895. doi:10.1016/j.cpr.2020.101895
68. Tang X, Renninger KA, Hidi SE, Murayama K, Lavonen J, Salmela-Aro K. The differences and similarities between curiosity and interest: meta-analysis and network analyses. *Learn Instr.* 2022;80:101628. doi:10.1016/j.learninstruc.2022.101628
69. Ober TM, Brooks PJ, Homer BD, Rindskopf D. Executive functions and decoding in children and adolescents: a meta-analytic investigation. *Educ Psychol Rev.* 2020;32(3):735-763. doi:10.1007/s10648-020-09526-0
70. Castro-Alonso JC, Wong RM, Adesope OO, Paas F. Effectiveness of multimedia pedagogical agents predicted by diverse theories: a meta-analysis. *Educ Psychol Rev.* 2021;33(3):989-1015. doi:10.1007/s10648-020-09587-1
71. Bourke M, Patten RK, Dash S, et al. The effect of interventions that target multiple modifiable health behaviors on symptoms of anxiety and depression in young people: a meta-analysis of randomized controlled trials. *J Adolesc Health.* 2022;70(2):208-219. doi:10.1016/j.jadohealth.2021.08.005
72. Karabinski T, Haun VC, Nübold A, Wendsche J, Wegge J. Interventions for improving psychological detachment from work: a meta-analysis. *J Occup Health Psychol.* 2021;26(3):224-242. doi:10.1037/ocp0000280
73. Estevez Cores S, Sayed AA, Tracy DK, Kempton MJ. Individual-focused occupational health interventions: a meta-analysis of randomized controlled trials. *J Occup Health Psychol.* 2021;26(3):189-203. doi:10.1037/ocp0000249
74. Schindler S, Hilgard J, Fritsche I, Burke B, Pfattheicher S. Do salient social norms moderate mortality salience effects? A (challenging) meta-analysis of terror management studies. *Pers Soc Psychol Rev.* 2023;27(2):195-225. doi:10.1177/10888683221107267
75. Endendijk JJ, van Baar AL, Deković M. He is a stud, she is a slut! A meta-analysis on the continued existence of sexual double standards. *Pers Soc Psychol Rev.* 2020;24(2):163-190. doi:10.1177/1088868319891310
76. Dailey S, Bergelson E. Language input to infants of different socioeconomic statuses: a quantitative meta-analysis. *Dev Sci.* 2022;25(3):e13192. doi:10.1111/desc.13192
77. Woods S, Dunne S, Gallagher P, McNicholl A. A systematic review of the factors associated with athlete burnout in team sports. *Int Rev Sport Exerc Psychol.* 2022;1-41. <https://www.tandfonline.com/doi/full/10.1080/1750984X.2022.2148225?scroll=top&needAccess=true#abstract>
78. Leijten P, Weisz JR, Gardner F. Research strategies to discern active psychological therapy components: a scoping review. *Clin Psychol Sci.* 2021;9(3):307-322. doi:10.1177/2167702620978615
79. Zaneva M, Guzman-Holst C, Reeves A, Bowes L. The impact of monetary poverty alleviation programs on children's and adolescents' mental health: a systematic review and meta-analysis across low-, middle-, and high-income countries. *J Adolesc Health.* 2022;71(2):147-156. doi:10.1016/j.jadohealth.2022.02.011
80. R Core Team. R: a language and environment for statistical computing. Version 4.1.3. 2023 <https://www.R-project.org/>
81. Ushey K, Allaire J, Tang Y. reticulate: interface to "Python." R package (version 1.27). 2023 <https://CRAN.R-project.org/package=reticulate>
82. Mersmann O, Trautmann H, Steuer D, Bornkamp B. truncnorm: truncated normal distribution. R package (version 1.0-8). 2018 <https://CRAN.R-project.org/package=truncnorm>
83. Singmann H, Bolker B, Westfall J, Aust F, Mattan SB-S. afex: analysis of factorial experiments. R package (version 1.2-1). 2023 <https://CRAN.R-project.org/package=afex>
84. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* Routledge; 2013. doi:10.4324/9780203771587
85. Kassambara A. rstatix: pipe-friendly framework for basic statistical tests. R package (version 0.7.2). 2023b <https://rpkgs.datanovia.com/rstatix/>
86. Kassambara A. ggpubr: "ggplot2" based publication ready plots. R-package (version 0.6.0). 2023a <https://rpkgs.datanovia.com/ggpubr/>
87. Greenhouse SW, Geisser S. On methods in the analysis of profile data. *Psychometrika.* 1959;24(2):95-112. doi:10.1007/BF02289823
88. Schielzeth H, Dingemanse NJ, Nakagawa S, et al. Robustness of linear mixed-effects models to violations of distributional assumptions. *Meth Ecol Evol.* 2020;11(9):1141-1152. doi:10.1111/2041-210X.13434
89. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning (Version 3). arXiv. 2019. doi:10.48550/ARXIV.1908.09635

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** 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. *Res Syn Meth.* 2024;15(6): 1120-1146. doi:10.1002/jrsm.1762

## APPENDIX A

TABLE A1 Means and CIs pertaining the main effects of the design factors.

Main effects	Groups	Sensitivity	Specificity	Cost
Stopping rule	Knee method	65.71 [65.64, 65.78]	76.40 [76.35, 76.46]	22.99 [22.94, 23.04]
	Heuristic method	72.27 [72.21, 72.32]	76.08 [76.05, 76.12]	22.76 [22.73, 22.80]
	Prevalence estimation	81.87 [81.82, 81.93]	41.69 [41.61, 41.78]	56.10 [56.01, 56.17]
Prevalence ratio	0.5%	68.77 [68.69, 68.85]	60.91 [60.81, 61.01]	38.92 [38.83, 39.02]
	1%	65.40 [65.32, 65.48]	67.24 [67.16, 67.32]	32.45 [32.37, 32.53]
	5%	77.21 [77.15, 77.28]	66.63 [66.56, 66.71]	31.79 [31.72, 31.87]
	10%	81.75 [81.69, 81.81]	64.13 [64.06, 64.20]	32.63 [32.57, 32.70]
Learning algorithm	LR + doc2vec	74.94 [74.83, 75.05]	62.61 [62.48, 62.72]	36.01 [35.89, 36.12]
	LR + SBERT	74.70 [74.60, 74.81]	66.94 [66.82, 67.06]	31.81 [31.69, 31.93]
	LR + TFIDF	73.04 [72.93, 73.15]	66.41 [66.29, 66.53]	32.32 [32.19, 32.44]
	NB + TFIDF	72.82 [72.72, 72.93]	66.90 [66.77, 67.02]	31.82 [31.70, 31.95]
	nn2layer + doc2vec	73.04 [72.93, 73.15]	62.64 [62.51, 62.76]	35.98 [35.86, 36.10]
	nn2layer + SBERT	72.03 [71.92, 72.14]	66.08 [65.96, 66.21]	32.66 [32.54, 32.77]
	RF + doc2vec	73.34 [73.24, 73.44]	63.50 [63.38, 63.62]	35.16 [35.03, 35.29]
	RF + TFIDF	71.35 [71.24, 71.45]	62.64 [62.51, 62.76]	35.94 [35.82, 36.07]
	SVM + TFIDF	74.28 [74.18, 74.39]	64.83 [64.70, 64.96]	33.84 [33.70, 33.96]

Note: Means and 95% CIs for all groups within the manipulated design factors stopping rule, learning algorithm, and prevalence ratio separate for sensitivity, specificity, and screening cost. All means and CIs are bootstrapped based on 1000 iterations. The design factors stopping rule, prevalence ratio, and learning algorithm comprise 756,000; 567,000; and 252,000 data points, respectively.

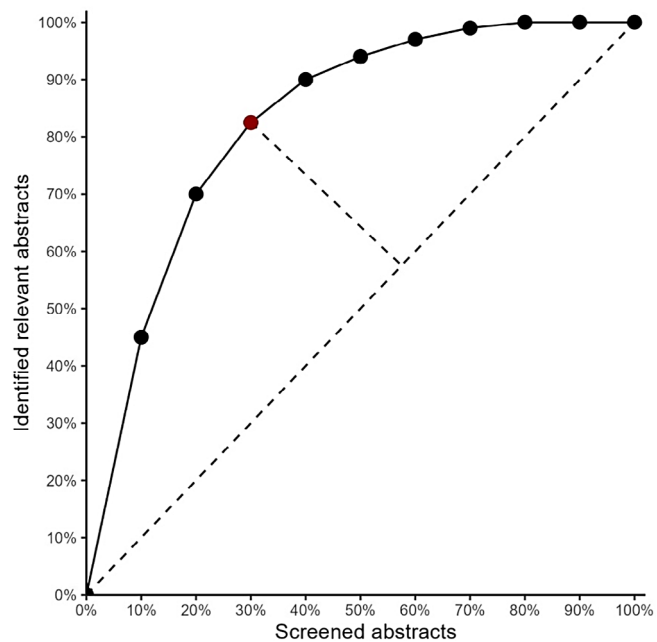
TABLE A2 Means and CIs pertaining the interaction of stopping rule and prevalence ratio.

Prevalence ratio	Stopping rule	Sensitivity	Specificity	Cost
0.5%	KM	71.75 [71.63, 71.88]	60.77 [60.62, 60.92]	39.07 [38.92, 39.23]
	HM	57.87 [57.75, 57.99]	81.12 [81.04, 81.19]	18.80 [18.73, 18.88]
	PE	76.69 [76.55, 76.81]	40.84 [40.66, 40.99]	58.90 [58.73, 59.06]
1%	KM	63.35 [63.20, 63.50]	73.97 [73.85, 74.09]	25.79 [25.67, 25.90]
	HM	56.71 [56.59, 56.83]	84.98 [84.95, 85.02]	14.87 [14.84, 14.91]
	PE	76.14 [76.01, 76.27]	42.77 [42.59, 42.94]	56.69 [56.51, 56.86]
5%	KM	60.91 [60.78, 61.03]	85.94 [85.88, 86.00]	13.40 [13.34, 13.46]
	HM	84.42 [84.34, 84.50]	72.20 [72.14, 72.27]	26.48 [26.42, 26.54]
	PE	86.31 [86.24, 86.39]	41.75 [41.60, 41.91]	55.50 [55.34, 55.66]
10%	KM	66.83 [66.72, 66.95]	84.94 [84.88, 84.99]	13.70 [13.66, 13.75]
	HM	90.06 [90.00, 90.13]	66.03 [65.96, 66.11]	30.89 [30.82, 30.97]
	PE	88.35 [88.29, 88.42]	41.41 [41.25, 41.58]	53.29 [53.15, 53.44]

Note: Means and 95% CIs regarding the interaction of Stopping rule and Prevalence ratio separate for sensitivity, cost, and specificity. All means and CIs are bootstrapped based on 1000 iterations. Each group consists of 189,000 observations. KM = Knee method, HM = heuristic stopping rule, PE = prevalence estimation.



## APPENDIX B



**FIGURE B1** Gain curve of a learning algorithm in AI-aided screening. The x-axis shows the percentage of screened abstracts. The y-axis shows the percentage of identified relevant abstracts. The dashed diagonal line represents the gain curve with random screening. The solid line shows a potential gain curve when active learning is applied. The points on the solid line represent potential knees, with the red point indicating the knee. The figure is adapted from Cormack and Grossman.<sup>22</sup>

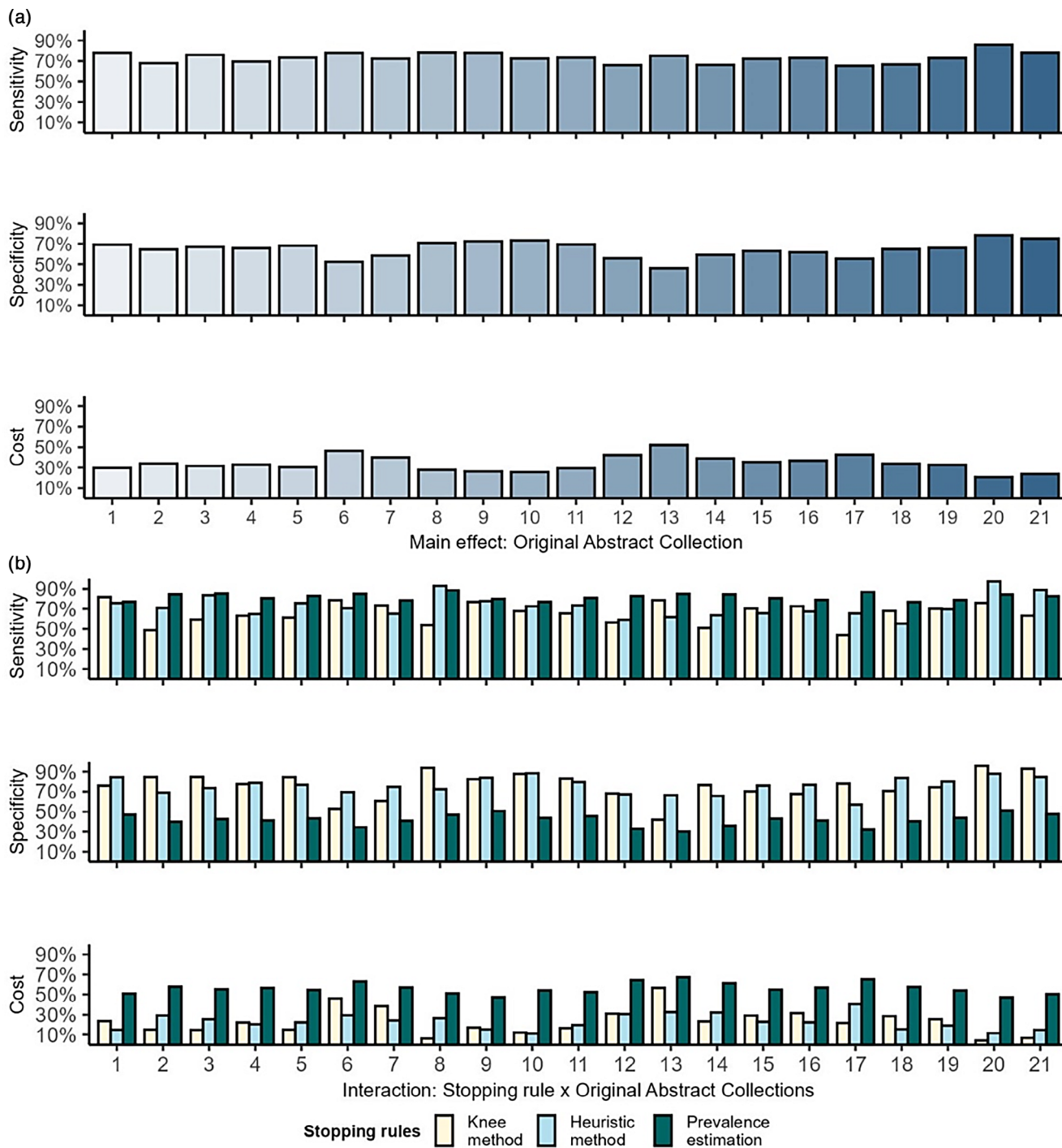
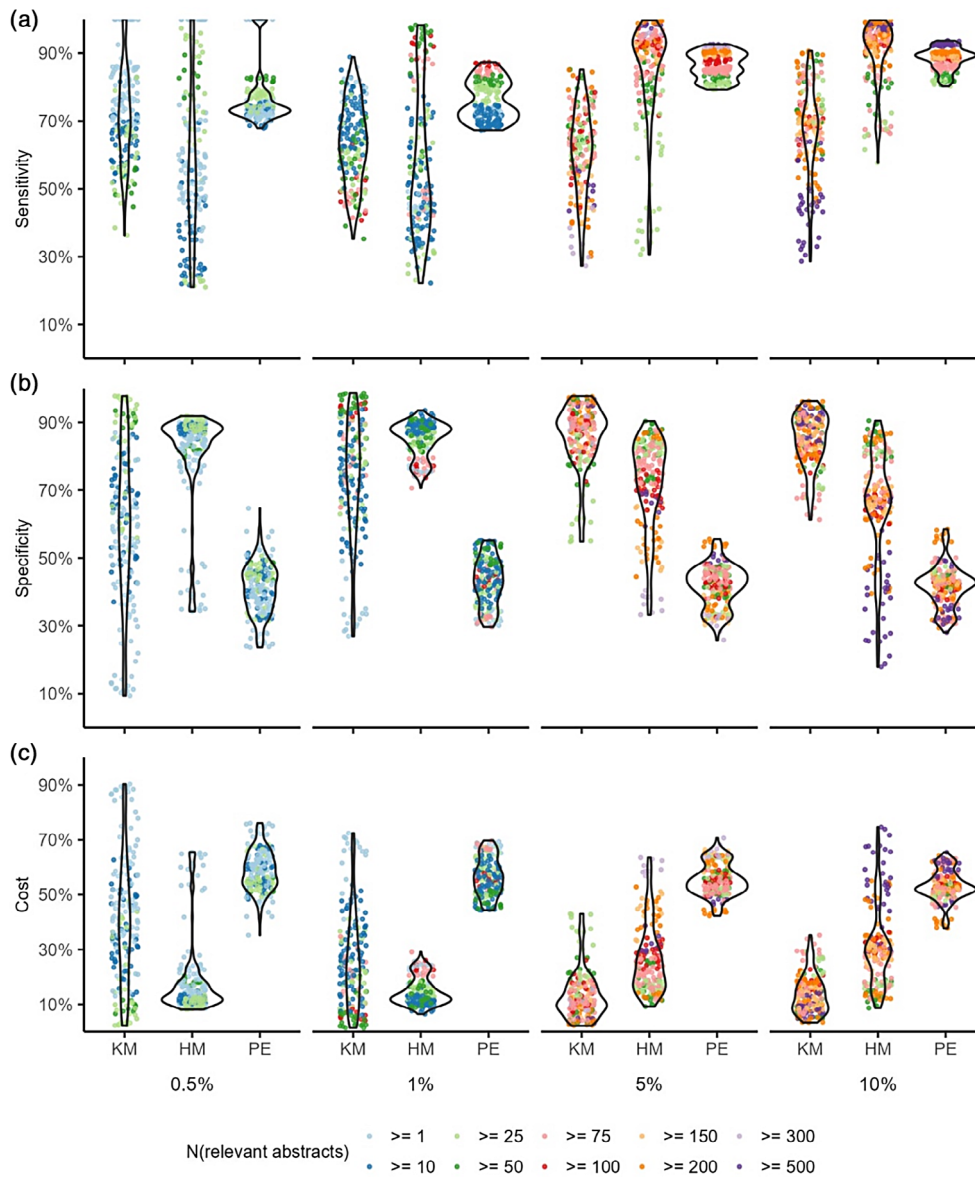


FIGURE B2 Bar-plot pertaining to the main effect of OAC and interaction of Stopping rule and OAC. The main effect of OAC (a) and the interaction of Stopping rule and OAC (b). As the CIs were too small to be visible, we did not include them in the figure. The OACs 1 to 21 are received from Alden et al.,<sup>66</sup> Bottema-Beutel et al.,<sup>60</sup> Bourke et al.,<sup>71</sup> Castro-Alonso et al.,<sup>70</sup> Dailey and Bergelson,<sup>76</sup> Endendijk et al.,<sup>75</sup> Estevez Cores et al.,<sup>73</sup> Hall et al.,<sup>63</sup> Hsieh et al.,<sup>65</sup> Karabinski et al.,<sup>72</sup> Khazanov et al.,<sup>61</sup> Leijten et al.,<sup>78</sup> Liu et al.,<sup>67</sup> Ober et al.,<sup>69</sup> Reimer and Sengupta,<sup>62</sup> Schindler et al.,<sup>74</sup> Simonsmeier et al.,<sup>64</sup> Tang et al.,<sup>68</sup> Vermillet et al.,<sup>59</sup> Woods et al.,<sup>77</sup> and Zaneva et al.,<sup>79</sup> respectively.



**FIGURE B3** Violin-plot pertaining to the interaction of stopping rule and prevalence ratio. Violin density plots pertaining to the interaction of Stopping rule and Prevalence ratio. Each dot represents a combination of learning algorithm and OAC and is the average of 21,000 data points. The colors represent groups matched by the number of relevant abstracts. KM = Knee method, HM = Heuristic method, PE = Prevalence estimation.