Original Paper

Performance of Two AI Approaches in ASReview Compared With Manual Screening for Dementia Care Literature Screening: Comparative Analysis

Dirk Steijger[1,2,3], MSc; Stella Thissen[4], MSc; Sil Aarts[2,3], Dr; Hilde Verbeek[2,3], Prof Dr; Marjolein E de Vugt[1], Prof Dr; Hannah Christie[5], Dr

[1]Department of Psychiatry and Neuropsychology, Mental Health and Neuroscience Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[2]Department of Health Service Research, CAPHRI Care and Public Health Research Institute, Faculty of Health Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

[3]The Living Lab in Ageing & Long-Term Care, Maastricht, The Netherlands

[4]Department of Primary and Community care, Research Institute for Medical Innovation, Radboud university medical center, Nijmegen, The Netherlands

[5]School of Population Health, Royal College of Surgeons, Dublin, Ireland

Corresponding Author:

Dirk Steijger, MSc
Department of Psychiatry and Neuropsychology
Mental Health and Neuroscience Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University
Dr Tanslaan, 12
Maastricht 6229ET
The Netherlands
Phone: 31 612365497
Email: dirk.steijger@maastrichtuniversity.nl

Abstract

Background: Literature reviews rely on rigorous title and abstract screening by researchers, which is time-consuming. AI-assisted literature screening tools have been proposed to improve efficiency by prioritizing titles and abstracts with the highest likelihood of meeting the inclusion criteria, thereby reducing the need to screen all records.

Objective: This study aims to evaluate the performance of two AI-assisted screening approaches in ASReview (version 1.3; Department of Methodology and Statistics, Utrecht University) compared with manual title and abstract screening in a previously completed and published scoping review on how AI can support the quality of life in people with dementia.

Methods: This study used a dataset of 4690 titles and abstracts from a published scoping review. The manual screening decisions of the scoping review served as the reference standard. Both ASReview approaches were applied by the same author who conducted the majority of the original manual title and abstract screening. Approach A used a simpler model with minimal prior input, whereas approach B used a more advanced model with a larger training set. Both ASReview approaches applied predefined stopping rules: (1) more than 10% of the dataset to be screened; and (2) 50 consecutive irrelevant titles and abstracts. Performance was evaluated in terms of sensitivity, specificity, precision, accuracy, and screening time. 95% CIs were calculated for sensitivity, specificity, precision, and accuracy. Agreement between manual and ASReview approaches was assessed using the Cohen κ, and differences in how manual and both ASReview approaches classified titles and abstracts were examined using the McNemar test. Performance and agreement were calculated at two levels: (1) after title and abstract screening and (2) after full-text inclusion.

Results: Manual screening identified 283 titles and abstracts for full-text review and resulted in 30 final included studies, requiring 19 hours. Of the 4690 titles and abstracts, approach A screened 830 (17.7%) in 4.3 hours and retrieved 16 of the 30 (sensitivity 0.53, 95% CI 0.36-0.70) final included studies, whereas approach B screened 798 (17.0%) in 5.5 hours and retrieved 21 of the 30 (sensitivity 0.70, 95% CI 0.52-0.83) final included studies. Although both ASReview approaches showed high specificity and accuracy, these metrics should be interpreted cautiously because the dataset was highly imbalanced and contained relatively few relevant titles and abstracts. McNemar tests showed significant directional imbalance

(P<.001): ASReview missed more manually selected titles and abstracts at level 1, whereas at level 2, ASReview more often labeled titles and abstracts not included in the final review as relevant.

Conclusions: ASReview can support workload reduction in title and abstract screening, but the evaluated ASReview approaches did not retrieve all final included studies from the original dementia care scoping review. These findings suggest that the evaluated ASReview configurations may be insufficient for reviews in which near-complete retrieval of relevant evidence is required.

JMIR Form Res 2026;10:e84989; doi: 10.2196/84989

Keywords: ASReview; dementia; literature screening; AI; information retrieval

Introduction

In health research, evidence synthesis is used to integrate and summarize existing literature and to inform future research, guidelines, policy, and decision-making [1]. Title and abstract screening is a tedious but vital step in this process [2]. The purpose of title and abstract screening is to identify those that are potentially eligible for full-text assessment and to exclude those that clearly do not meet the eligibility criteria [2]. Although title and abstract screening is only 1 component of the evidence synthesis process, it is a key step in determining which records proceed to full-text review. Overlooking relevant titles and abstracts during screening can introduce bias and affect the completeness of the evidence base [1]. Because the title and abstract screening phase often involves a large number of titles and abstracts, it is time-consuming and cognitively demanding [3]. Therefore, efficient and innovative approaches to support the title and abstract screening process are needed.

To streamline the title and abstract screening process, various AI-assisted screening tools have emerged [4-8]. These AI-assisted screening tools may use techniques such as text mining, machine learning, natural language processing, and deep learning to identify and actively learn patterns in titles and abstracts and prioritize titles and abstracts that are more likely to be relevant. AI-assisted screening can be implemented in specialized screening tools, such as ASReview (version 1.3; Department of Methodology and Statistics, Utrecht University) [6], and Abstrackr [9] or as AI-supported functions within broader review management platforms, such as DistillerSR [10], Covidence [11], and Rayyan [12]. These AI-assisted screening tools are intended to support reviewer decision-making by prioritizing titles and abstracts based on predicted relevance [5]. By identifying likely relevant studies early, these AI-assisted screening tools could potentially allow reviewers to screen only a subset of the dataset, thereby reducing workload.

Previous studies evaluating AI-assisted screening tools in health research have shown that these tools can potentially reduce screening time while maintaining the ability to identify relevant titles and abstracts [9,13-17]. Focusing specifically on ASReview, a widely used open-source AI-assisted screening tool [18], previous evaluations have reported promising performance in different review contexts [14,19-23]. The performance of AI-assisted screening tools depends on the effectiveness of the underlying AI techniques and algorithms, the quality and quantity of data used for training, and the degree of human involvement [7,24,25]. Therefore, further evaluations are needed to understand how ASReview performs in title and abstract screening when different model configurations are applied across different review contexts.

Given this dependence on context, the intersection of dementia care, quality of life, and AI provides a relevant context for further evaluation. This intersection provides a useful starting point for examining how ASReview performs in a review area that combines dementia care research with technology-oriented literature. To our knowledge, ASReview has not yet been evaluated in such a dementia care literature review context. Therefore, this study aimed to evaluate the performance of two AI-assisted screening approaches within ASReview compared with manual title and abstract screening in a previously completed and published scoping review on how AI can support quality of life in people with dementia [26]. Specifically, this study examined the extent to which both ASReview approaches reduced screening workload and identified the same titles and abstracts as included via manual screening in the scoping review. By comparing title and abstract screening in ASReview with a manual screening process, this study adds empirical evidence on the practical use of AI-assisted screening in a real-world dementia care review context.

Methods

Study Design

This study was designed as a comparative methodological evaluation of two AI-assisted screening approaches within ASReview compared with the manual screening approach used in a previously published scoping review on how AI can support quality of life in people with dementia [26]. Reporting of the present study was guided by the PRISMA-trAIce (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Transparent Reporting of AI-Assisted Evidence Synthesis) guidelines for the transparent reporting of AI-assisted evidence synthesis, where applicable [27].

Unit of Analysis, Data Source, and Study Size

The unit of analysis was the individual title and abstract record. The data were derived from the title and abstract dataset of the original scoping review, including both the manual screening decisions and the final full-text inclusion decisions. Manual screening and inclusion decisions from

the original review served as the reference standard against which the two ASReview approaches were evaluated. The study size corresponded to that of the full screening dataset from the original scoping review. After deduplication and preparation for screening, the dataset consisted of 4690 titles and abstracts, of which 283 were selected for full-text review and 30 were included in the final review.

Compared Approaches

Manual Approach

The manual screening approach was performed as part of a previously published scoping review [26]. In brief, the original review searched PubMed, Scopus, the Association for Computing Machinery Digital Library, and Google Scholar for studies published between 2010 and January 2024 on AI-based approaches that could support the quality of life of people with dementia. Eligible studies involved people with at least suspected dementia, evaluated an AI-based approach to the daily living of people with dementia, and reported original empirical research. Studies were excluded if they focused on diagnostic or acute care applications, if they were not available in English or Dutch, or if the full text was unavailable.

After deduplication, records were imported into Rayyan and screened, without the use of the AI function in Rayyan, against the eligibility criteria. Initially, 2 reviewers (DS with 5 years of research experience and the other reviewer HC with 10 years of research experience) independently screened 10% of the titles and abstracts from the scientific databases. As the interrater agreement exceeded the predefined threshold (80%), 1 reviewer (DS) continued screening the remaining titles and abstracts. The same procedure was applied to full-text screening. For the Google Scholar gray literature search, 1 reviewer (DS) assessed the first 300 results. However, no Google Scholar records were included in the final review or in the dataset used for the present ASReview evaluation. A detailed description of the search strategy, eligibility criteria, and screening process is reported in the original scoping review [26]. The resulting manual title and abstract screening decisions and final full-text inclusion decisions were used as the reference standards for the present methodological evaluation.

ASReview

ASReview is an open-source AI-assisted screening tool that uses active learning to prioritize titles and abstracts according to their predicted relevance [6]. Before the title and abstract screening starts, ASReview requires prior knowledge consisting of titles and abstracts labeled as relevant and irrelevant by the reviewer. Based on this prior knowledge, the tool ranks the remaining titles and abstracts from highest to lowest predicted relevance. The reviewer then screens the highest-ranked title and abstract and labels it as relevant or irrelevant according to the eligibility criteria. After each reviewer's decision, ASReview updates the ranking and presents the next title and abstract with the highest predicted relevance. This researcher-in-the-loop process continues until a predefined stopping rule is reached. ASReview allows reviewers to preselect different model configurations, including feature extraction methods (ie, Doc2Vec, embedding inverse document frequency, term frequency–inverse document frequency [TF-IDF], and Sentence-Bidirectional Encoder Representations from Transformers [sBERT]) and classifiers (Naïve Bayes, support vector machines, deep neural network, logistic regression, long short-term memory, and random forest) [6]. In the present study, ASReview was applied to the same dataset and eligibility criteria as the manual screening approach. This allowed a direct comparison between the original manual screening decisions and the screening decisions generated through two AI-assisted ASReview configurations. The two ASReview configurations were based on recommendations from ASReview's model selection guide and the SAFE (Screen, Apply, Find, Evaluate) procedure [24,25]: a default configuration and a custom configuration.

Approach A: Default Configuration

Approach A was included as a baseline AI-assisted screening approach because it reflects ASReview's default configuration and requires minimal prior-knowledge input from the reviewer [28,29]. It used term TF-IDF for feature extraction, combined with a Naïve Bayes classifier. The query strategy was set to maximum, and the balance strategy was set to dynamic resampling (double). TF-IDF represents titles and abstracts based on word frequencies, while Naïve Bayes uses these patterns to estimate the likelihood of relevance. Four titles and abstracts, consisting of 1 relevant and 3 irrelevant examples, were used as prior knowledge. The prior-knowledge titles and abstracts were included in the analysis as classified titles and abstracts by ASReview.

Approach B: Custom Configuration

Approach B was included to evaluate whether a more semantically rich model configuration, combined with a larger prior knowledge set, would improve the retrieval of relevant titles and abstracts compared with the default configuration. Approach B used sBERT for feature extraction, combined with a fully connected neural network classifier. In ASReview version 1.3, sBERT-based feature extraction required an extension package. The exact extension package name and version used in the present analysis were not retained. The query strategy was set to maximum, and the balance strategy used dynamic resampling (double). sBERT represents titles and abstracts as dense semantic vectors that capture contextual meaning beyond individual word frequencies, while the neural network classifier uses these representations to estimate the likelihood of relevance. Prior knowledge consisted of 47 titles and abstracts, corresponding to approximately 1% of the original dataset. This set included 2 relevant titles and abstracts randomly selected from the 30 final included titles and abstracts and 45 irrelevant titles and abstracts randomly selected from the titles and abstracts that were excluded for full-text screening in the original review. This ensured that the prior knowledge set contained both relevant and irrelevant titles and abstracts, consistent with the SAFE procedure's recommendation to start with a labeled training set of approximately 1% of the dataset containing at

least 1 relevant and 1 irrelevant title and abstract [25]. Prior knowledge titles and abstracts were included in the analysis as classified titles and abstracts by ASReview.

Screening Procedure

The screening procedure and stopping rule were informed by selected elements of the SAFE procedure [6,24,30]. The SAFE procedure provides practical guidance for active learning–based screening, including the use of prior knowledge and stopping heuristics that combine a minimum screening proportion with a run of consecutive irrelevant titles and abstracts. Because the present study was a retrospective methodological evaluation rather than a live review selection process, the full SAFE procedure was not implemented. Instead, selected SAFE elements were used to define the prior-knowledge strategy and stopping rule for the ASReview screening approaches.

For the present study, 2 separate ASReview projects were created in the ASReview environment, 1 for approach A and 1 for approach B. For each project, the complete title and abstract dataset from the original scoping review were exported from Rayyan and uploaded to ASReview. Approach A and approach B were conducted as independent comparative screening projects. The screening decisions from approach A were not used as training data for approach B.

To ensure consistency in the application of the eligibility criteria, the same reviewer who conducted the majority of the title and abstract screening in the original scoping review (DS) also screened the titles and abstracts presented by ASReview for each approach. The same eligibility criteria as in the original scoping review were applied. Titles and abstracts were labeled as relevant or irrelevant independently for approach A and approach B.

Screening continued until the predefined stopping rule was met. This stopping rule required that 2 conditions were fulfilled: first, at least 10% of the complete dataset had to be screened; second, after this minimum threshold had been reached, screening continued until 50 consecutive titles and abstracts had been labeled as irrelevant by the reviewer. The threshold of 50 consecutive irrelevant titles and abstracts was selected as a pragmatic stopping criterion. The authors acknowledge that more conservative thresholds, such as 100 consecutive irrelevant titles and abstracts, may increase sensitivity by extending the screening process. The same stopping rule was applied to both ASReview approaches. Because each ASReview configuration generated a different ranking of titles and abstracts, the stopping rule was reached after a different number of screened titles and abstracts in each approach. After stopping, the titles and abstracts labeled as relevant by each ASReview approach were compared with the manual title and abstract screening decisions and final inclusion decisions from the original scoping review.

Performance Outcomes

To assess performance, both ASReview approaches were compared with manual screening using sensitivity, specificity, precision, accuracy, screening time, and agreement. Sensitivity was used to assess the proportion of manually selected titles and abstracts that were also identified by each ASReview approach. Specificity was used to assess the proportion of manually excluded titles and abstracts that were also excluded by ASReview. Precision was used to assess the proportion of titles and abstracts labeled as relevant by each ASReview approach that were also relevant according to the manual approach. Accuracy was used to assess the overall proportion of titles and abstracts that were correctly classified by each ASReview approach, including both records considered relevant and records considered irrelevant according to the manual approach. Agreement between each ASReview approach and the manual approach was measured using the Cohen κ and observed agreement (P_o) [31], and the McNemar test was used to assess directional imbalance in discordant classifications [32]. Total screening time was estimated using the time-tracking information displayed within Rayyan for the manual approach and within ASReview for both AI-assisted approaches. Screening time did not include full-text screening or data extraction. False negatives were examined descriptively to explore potential screening-related selection bias by assessing whether missed titles and abstracts shared common characteristics, such as dementia care context, AI methods, or failure to report AI performance. The evaluation level at which each metric was calculated is described in the Data Analysis section.

Statistical Analysis

Performance was assessed at 2 levels. At level 1, manual title and abstract screening decisions were used as the reference standard. Thus, titles and abstracts selected for full-text review during manual screening were considered relevant, whereas titles and abstracts excluded during manual title and abstract screening were considered irrelevant. At this level, sensitivity, Cohen κ, observed agreement, and the McNemar test were calculated.

At level 2, the final full-text inclusion decisions from the original scoping review were used as the reference standard: titles and abstracts included in the final review were considered relevant, whereas all other records were considered irrelevant. At this level, sensitivity, specificity, precision, accuracy, Cohen κ, observed agreement, and the McNemar test were calculated. This approach aligns with prior research, suggesting that although ASReview retrieves fewer abstracts, a higher proportion may be included in the final inclusion set [25,33]. For binary classification metrics, 95% CIs were calculated using Wilson score intervals [34]. At both levels, prior knowledge titles and abstracts were included in the statistical analysis of the performance metrics.

Screening time referred only to the title and abstract screening process and did not include full-text screening (during the manual approach). For all approaches, total screening time was recorded in minutes at the end of the screening process.

Results

Figure 1 summarizes the flow of titles and abstracts through the manual screening approach and the two ASReview approaches. The manual screening approach screened a highly imbalanced dataset of 4690 abstracts, identified 283 abstracts for full-text review, and included 30 final studies, requiring 19 hours. Approach A achieved 50 consecutive irrelevant abstracts and thus stopped screening at 830 (17.7%) abstracts in 4.3 hours, with a sensitivity of 0.16 at level 1 and 0.53 at level 2 (16 abstracts), specificity of 0.99, precision of 0.21, and accuracy of 0.98 at level 2. Approach B achieved 50 consecutive irrelevant abstracts and thus stopped screening at 798 (17.0%) abstracts in 5.5 hours, achieving higher sensitivity at both level 1: 0.23 and level 2: 0.70 (21 abstracts), specificity of 0.99, precision of 0.24, and accuracy of 0.98 at level 2. Both ASReview approaches reduced screening time compared with manual screening. At the final full-text inclusion decisions from the original scoping review stage, sensitivity was higher for approach B than for approach A (0.70 vs 0.53; an absolute difference of 0.17). The performance metrics are shown in Table 1. The 95% CIs indicate uncertainty around the sensitivity and precision estimates, particularly at level 2. For approach B, level 2 sensitivity was 0.70 (95% CI 0.52-0.83) and level 2 precision was 0.24 (95% CI 0.16-0.34).

Figure 1. Flow diagram of titles and abstracts through two AI-assisted literature screening approaches within ASReview compared with manual title and abstract screening from a dementia care scoping review. The figure summarizes how the dataset from the original dementia care scoping review was screened manually and with two AI-assisted approaches within ASReview. After deduplication and preparation, the dataset consisted of 4690 titles and abstracts. Manual screening assessed all 4690 titles and abstracts, selected 283 for full-text review, and resulted in 30 final included studies. Approach A screened 830 titles and abstracts until the predefined stopping rule was reached, labeled 78 titles and abstracts as relevant, and retrieved 16 of the 30 final included studies. Approach B screened 798 titles and abstracts until the predefined stopping rule was reached, labeled 87 titles and abstracts as relevant, and retrieved 21 of the 30 final included studies.

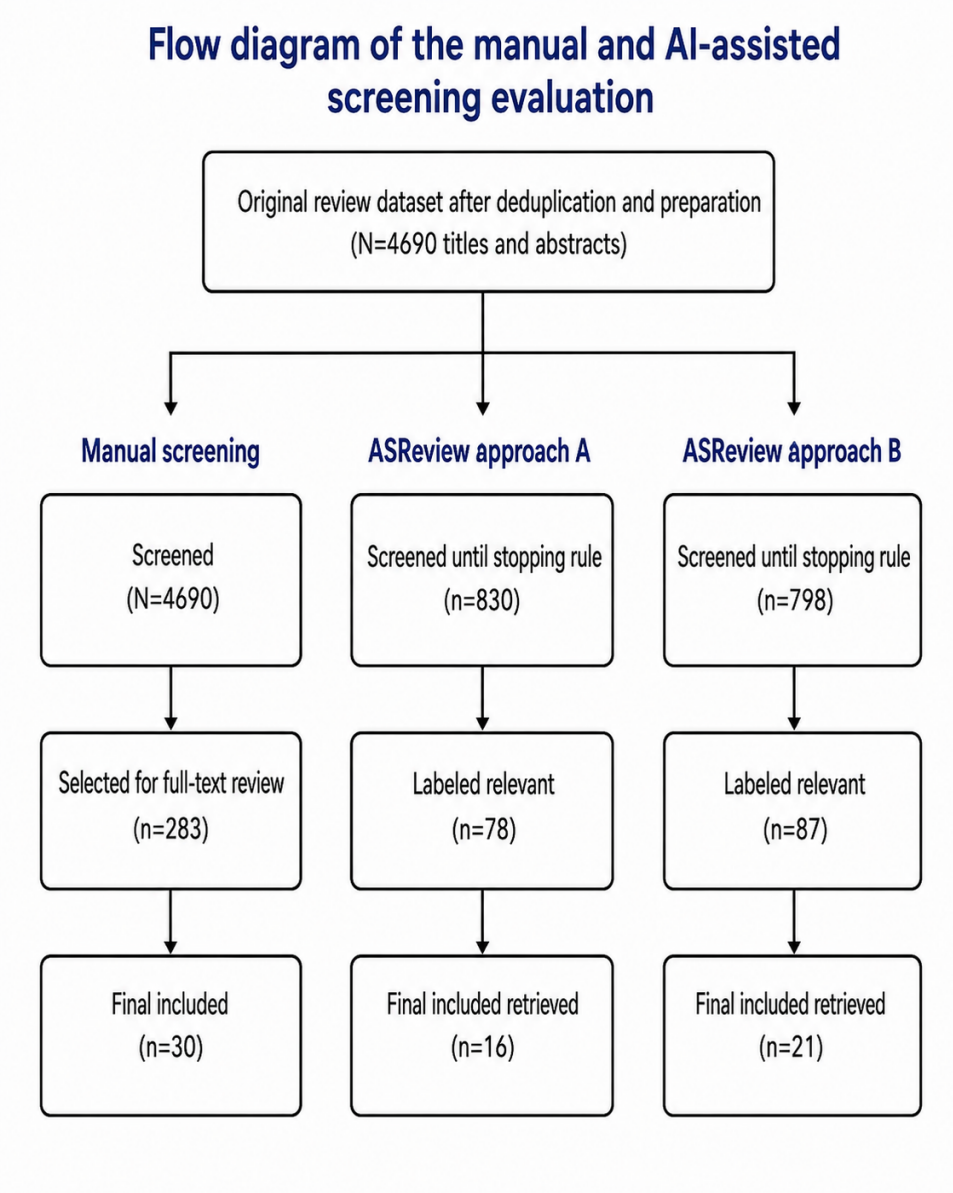

Table 1. Performance of two AI-assisted literature screening approaches within ASReview compared with manual title and abstract screening from a dementia care scoping review.[a]

Approach	Sensitivity (level 1, 95% CI)[b,c]	Sensitivity (level 2, 95% CI)[b,d]	Specificity (level 2, 95% CI)[b,d]	Precision (level 2, 95% CI)[b,d]	Accuracy (level 2, 95% CI)[b,d]	Screened titles and abstract, n (%)	Screening time (h)[e]
Manual[f]	1.00	1.00	1.00	1.00	1.00	4690 (100)	19
A	0.16 (0.12-0.21)	0.53 (0.36-0.70)	0.99 (0.98-0.99)	0.21 (0.13-0.31)	0.98 (0.98-0.99)	830 (17.7)	4.3
B	0.23 (0.18-0.28)	0.70 (0.52-0.83)	0.99 (0.98-0.99)	0.24 (0.16-0.34)	0.98 (0.98-0.99)	798 (17.0)	5.5

[a]The dataset consisted of 4690 titles and abstracts from a previously published scoping review on how AI can support quality of life in people with dementia, covering studies published between 2010 and January 2024.
[b]The CIs were calculated using Wilson score intervals.
[c]Level 1 sensitivity was calculated using the 283 titles and abstracts selected for full-text review during manual screening as the reference standard.
[d]Level 2 sensitivity, specificity, precision, and accuracy were calculated using the 30 titles and abstracts included in the final review as the reference standard.
[e]Screening time refers only to title and abstract screening and does not include full-text screening.
[f]Manual screening decisions from the original scoping review served as the reference standard; therefore, values of 1.00 for the manual approach are fixed by definition and should not be interpreted as evidence of perfect manual screening performance.

At level 1, agreement between manual screening and ASReview was moderate for both ASReview approaches (κ=0.24, P_o=0.94 for approach A; κ=0.24, P_o=0.95 for approach B). At level 2, Cohen κ values increased slightly (κ=0.28 for approach A; κ=0.35 for approach B) with high observed agreement (P_o=0.98 for both). McNemar tests showed a significant but level-dependent directional imbalance (P<.001). At level 1, ASReview missed more manually selected titles and abstracts than it additionally labeled as relevant: 237 vs 32 for approach A and 219 vs 23 for approach B. At level 2, the direction was reversed: ASReview labeled more titles and abstracts not included in the final review as relevant than it missed final included titles and abstracts, with 62 vs 14 for approach A and 66 vs 9 for approach B. Agreement statistics are presented in Table 2, with the corresponding 2 × 2 contingency tables provided in Multimedia Appendix 1. Analysis of the false negatives revealed no dominant thematic pattern; false-negative abstracts covered both AI and dementia, with no specific topic being consistently missed.

Table 2. Agreement and directional imbalance between two ASReview approaches and manual title and abstract screening from a dementia care scoping review.[a]

Level[b]	Comparison[c]	Cohen κ[d]	Observed agreement (P_o)[d]	b[e]	c[e]	McNemar test (P value)[f]
1	Manual vs A	0.24	0.94	237	32	<.001
1	Manual vs B	0.24	0.95	219	23	<.001
2	Manual vs A	0.28	0.98	14	62	<.001
2	Manual vs B	0.35	0.98	9	66	<.001

[a]The dataset consisted of 4690 titles and abstracts from a previously published scoping review on AI-based approaches to support quality of life in people with dementia, covering studies published between 2010 and January 2024.
[b]At level 1, titles and abstracts selected for full-text review during manual screening were considered relevant. At level 2, titles and abstracts included in the final scoping review were considered relevant.
[c]Manual screening decisions from the original review served as the reference standard.
[d]Agreement between each ASReview approach and manual screening was assessed using Cohen κ and observed agreement (P_o).
[e]The values b and c represent discordant classifications between manual screening and ASReview.
[f]The McNemar test was used to assess directional imbalance in discordant classifications.

Discussion

Principal Findings

This study evaluated the performance of two AI-assisted screening approaches within ASReview compared with manual title and abstract screening in a previously completed dementia care scoping review. Both ASReview approaches substantially reduced screening time and the number of titles and abstracts that required screening, but neither approach retrieved all studies included in the final scoping review. Approach B (custom configuration), which used a semantically richer feature extraction method and a larger prior knowledge set, performed better than approach A (default configuration). However, approach B still missed several final included studies. These findings illustrate that ASReview can support workload reduction in the title and abstract screening process. However, the use of ASReview may be insufficient for reviews in which the aim is to identify and capture as much evidence as possible.

The finding that ASReview can reduce screening workload is in line with previous evaluations of active learning-based screening tools for title and abstract screening [13,14,17,22,35,36]. However, there is no solid overview of how well active learning-based screening tools perform in the title and abstract screening process [23], and previous evaluations show that performance varies across review topics, dataset size, model configurations, prior knowledge strategies,

and stopping rules [7,19,37]. This means that ASReview performance cannot be inferred from previous evaluations alone but should be assessed within the specific review context in which it is used. For active learning-based screening workflows, this has implications: a fixed minimum screening proportion, such as 10% or 50% of the total dataset, cannot by itself ensure full identification of relevant titles and abstracts [19,38]. Sensitivity also depends on the model configuration and the composition of the prior knowledge set [19,36]. All these implications may be particularly important in review topics with heterogeneous terminology, where relevant titles and abstracts may be described indirectly or inconsistently [19].

The high specificity and accuracy observed for both ASReview approaches should be interpreted in light of the highly imbalanced dataset. In title and abstract screening, irrelevant titles and abstracts typically form the large majority of the dataset [39]. As a result, high specificity and accuracy can mainly reflect the correct exclusion of irrelevant titles and abstracts rather than the successful identification of relevant titles and abstracts [40]. Therefore, sensitivity, false negatives, and the number of missed final included studies are particularly important for assessing whether an AI-assisted screening workflow is suitable for evidence synthesis.

The 70% (21/30) sensitivity observed for the best-performing ASReview approach represents a critical limitation for using the evaluated workflow as a screening strategy. Approach B retrieved 21 of the 30 final included studies from the original review, meaning that 9 studies would have been missed if this approach had been used as the primary screening approach during the original review. This level of sensitivity would be insufficient for reviews in which near-complete capturing of evidence is expected [41]. By comparison, dual-reviewer screening of titles and abstracts misses around 2.5% of the relevant titles and abstracts [42]. However, dual-reviewer screening is not always feasible because it requires substantial resources [5,13,15]. This trade-off could be acceptable for rapid reviews, where reviewers are willing to give up some degree of certainty regarding the comprehensiveness of the included evidence [42]. However, it remains unclear what level of reduced sensitivity is acceptable when AI-assisted screening tools are used. Further research is therefore needed to determine what level of reduced sensitivity is acceptable when using AI-assisted screening tools in different review contexts.

Agreement analysis highlighted the effect of class imbalance in title and abstract screening. Observed agreement between ASReview and manual screening was high; this was largely driven by shared exclusions of titles and abstracts. In contrast, Cohen κ values ranged from 0.24 to 0.35, which indicate poor agreement beyond chance [43]. This suggests that, despite high observed agreement, ASReview showed limited agreement with manual screening on which titles and abstracts should be selected as relevant. This is expected in title and abstract screening, where relevant titles and abstracts are typically a small proportion of the dataset, but the low kappa values should not be interpreted as acceptable agreement [39,43,44]. The McNemar test provided complementary information by showing that the direction of discordance differed between evaluation levels [45]. At level 1, ASReview missed more titles and abstracts selected for full-text review by manual screening than it additionally labeled as relevant. At level 2, the direction was reversed: ASReview labeled more titles and abstracts that were not included in the final review as relevant than it missed among the final included titles and abstracts. Together, these findings show that high observed agreement could mask practically important disagreement. Missed final included titles and abstracts may reduce the completeness of the evidence base, whereas level 2 false-positive titles and abstracts indicate that workload reduction was accompanied by limited precision among ASReview-selected titles and abstracts.

Current study findings should be interpreted in the context of a broader movement toward responsible use of AI in evidence synthesis. The RAISE (Responsible Use of AI in Evidence Synthesis) recommendations, developed by Cochrane, among others, provide a framework for ensuring the responsible use of AI across all roles within the evidence synthesis process [46]. Similarly, PRISMA-trAIce frameworks provide a reporting framework to improve transparency when AI is used as a methodological tool in evidence synthesis [27]. These developments may support broader adoption of AI-assisted screening tools in the title and abstract screening process but also increase the need for transparent reporting and validation of how such tools are used in specific review contexts [37]. Retrospective validation studies may help identify context-specific best practices for using active learning during title and abstract screening [47]. However, a lack of dataset availability can hinder reproducibility [23,33]. The present study contributes to this evidence by evaluating ASReview's performance in the context of a dementia care scoping review and by adopting an open science approach, making the full dataset and reviewer decisions publicly available so that the current study is reproducible.

Strengths and Limitations

To our knowledge, this is the first study to evaluate the performance of ASReview in the dementia care literature research. By making the full title and abstract dataset and all screening decisions available, this study also contributes to transparency, reproducibility, and future validation of AI-assisted screening in this review context. Another strength is that the same researcher conducted both the manual and AI-assisted screening, which helped ensure consistent adherence to the eligibility criteria across all screening approaches. However, the current results need to be viewed in light of some possible limitations. This study focused on a single scoping review, which limits generalizability to other review topics. Second, the manual screening and final inclusion decisions from the original scoping review were used as the reference standard for evaluating both ASReview approaches. Although the original review included an initial dual-screening validation procedure, the final screening decisions should not be interpreted as an error-free gold standard. Third, the ASReview screening

approaches were conducted retrospectively by a reviewer who had been involved in the original scoping review title and abstract screening process. Therefore, the reviewer may have been more familiar with the review topic, eligibility criteria, and terminology during the ASReview screening. This familiarity may have influenced labeling decisions and may also have reduced the time needed to assess titles and abstracts compared with the original manual screening process, in which the reviewer still had to become familiar with the literature and screening criteria. Finally, this evaluation was conducted using ASReview version 1.3. The exact extension package name and version used to enable sBERT-based feature extraction in approach B were not retained. Because ASReview and its extensions are under active development, findings may differ with newer releases, plugin versions, alternative model configurations, or updated workflow options.

Conclusions

This study evaluated two AI-assisted screening approaches within ASReview against manual title and abstract screening in a completed dementia care scoping review. Both ASReview approaches reduced screening time and the number of titles and abstracts requiring manual screening, but neither approach retrieved all titles and abstracts included in the final review. These findings suggest that AI-assisted screening tools, in particular ASReview, can support workload reduction in title and abstract screening but that both evaluated ASReview approaches may be insufficient for reviews in which near-complete capturing of evidence is required. Further research is needed to determine how AI-assisted screening tools perform in review contexts and to determine what level of reduced sensitivity is acceptable when using AI-assisted screening tools in different review contexts.

Acknowledgments

During the preparation and revision of this manuscript, the corresponding author (DS) used ChatGPT (OpenAI) to support language editing and to assist with drafting responses to reviewer comments. The tool was not used for data analysis, interpretation of the results, generation of references, or autonomous scientific decision-making. All AI-assisted text was critically reviewed, edited, and approved by the authors. The authors take full responsibility for the content of this manuscript.

Funding

This research was funded by the Dutch Research Council (NWO) through the QoLEAD (Quality of Life by use of Enabling AI in Dementia) project (project: KICH1.GZ02.20.008). Additional support from Alzheimer Nederland is gratefully acknowledged.

Data Availability

All data generated or analyzed during this study are included in the supplementary information files. The original title and abstract dataset from the published scoping review, including the manual screening decisions and final inclusion decisions, is provided in Multimedia Appendix 1. The screening decisions generated by both ASReview approaches, as well as the titles and abstracts used as prior knowledge for each ASReview approach, are also provided in Multimedia Appendix 1.

Authors' Contributions

Conceptualization: DS, ST, SA, HV, MEdV, HC
Formal analysis: DS
Investigation: DS, ST
Methodology: DS, ST, SA, HV, MEdV, HC
Supervision: HV, MEdV, HC
Validation: SA
Visualization: DS, SA
Writing – original draft: DS, ST
Writing – review & editing: DS, ST, SA, HV, MEdV, HC

Conflicts of Interest

None declared.

Multimedia Appendix 1

Source datasets and screening decision files for the comparative evaluation of 2 AI-assisted screening approaches within ASReview compared with manual screening.
[ZIP File (ZIP archive File), 12375 KB-Multimedia Appendix 1]

References

1. Higgins JPT, Thomas J, Chandler J, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. Wiley; 2019. URL: https://dariososafoula.wordpress.com/wp-content/uploads/2017/01/cochrane-handbook-for-systematic-reviews-of-interventions-2019-1.pdf [Accessed 2026-07-22]

2. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Res Synth Methods. Sep 2019;10(3):330-342. [doi: 10.1002/jrsm.1354]
3. Nussbaumer-Streit B, Ellen M, Klerings I, et al. Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol. Nov 2021;139:287-296. [doi: 10.1016/j.jclinepi.2021.05.019] [Medline: 34091021]
4. Blaizot A, Veettil SK, Saidoung P, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. Res Synth Methods. May 2022;13(3):353-362. [doi: 10.1002/jrsm.1553] [Medline: 35174972]
5. Mogoale PD, Pretorius AB, Mogase RC, Segooa MA. Evaluating the efficacy of AI tools in systematic literature reviews: a comprehensive analysis. J Inf Syst Inform. 2025;7(1):870-888. [doi: 10.51519/journalisi.v7i1.1035]
6. van de Schoot R, de Bruin J, Schram R, et al. Open source software for efficient and transparent reviews. Preprint posted online on Jun 22, 2020. [doi: 10.48550/arXiv.2006.12166]
7. Ge L, Agrawal R, Singer M, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. Syst Rev. Oct 25, 2024;13(1):269. [doi: 10.1186/s13643-024-02682-2] [Medline: 39456077]
8. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res. Jan 12, 2024;26:e48996. [doi: 10.2196/48996] [Medline: 38214966]
9. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev. Jun 15, 2015;4(1):80. [doi: 10.1186/s13643-015-0067-6] [Medline: 26073974]
10. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening - impact on reviewer-relevant outcomes. BMC Med Res Methodol. Oct 15, 2020;20(1):256. [doi: 10.1186/s12874-020-01129-1] [Medline: 33059590]
11. Babineau J. Product review: covidence (systematic review software). J Can Health Libr Assoc. 2014;35(2):68-71. [doi: 10.5596/c14-016]
12. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. Dec 5, 2016;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
13. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. Jul 7, 2023;13(7):e072254. [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
14. van der Pol JA, Huizinga TW, Bergstra SA. Is AI-assisted active learning software able to reliably speed-up systematic literature reviews in rheumatology? A real-time comparison of AI-assisted and manual abstract selection. RMD Open. Dec 4, 2024;10(4):e005024. [doi: 10.1136/rmdopen-2024-005024] [Medline: 39632096]
15. Delgado-Chaves FM, Jennings MJ, Atalaia A, et al. Transforming literature screening: the emerging role of large language models in systematic reviews. Proc Natl Acad Sci U S A. Jan 14, 2025;122(2):e2411962122. [doi: 10.1073/pnas.2411962122] [Medline: 39761403]
16. Akaraci S, Jones SM, Tate C, et al. Evaluating the use of artificial intelligence (AI) in systematic review abstract screening: a comparative study of AI-aided tools. Research Square. Preprint posted online on Jan 21, 2026. [doi: 10.21203/rs.3.rs-7915337/v1]
17. Gauthier Mongeon J, Ouadfel S, Thullier F, Gaboury S, Arsenault-Lapierre G. Manual versus AI-assisted document screening (ASReview): a comparative analysis within a rapid systematized review in the social sciences. Int J Soc Res Methodol. 2026:1-19. [doi: 10.1080/13645579.2026.2668099]
18. Kataoka Y, Banno M, Kyo M, et al. TiAb review plugin: a browser-based tool for AI-assisted title and abstract screening. arXiv. Preprint posted online on Apr 8, 2026. [doi: 10.48550/arXiv.2604.08602]
19. Boesen K, Hemkens L, Janiaud P, Hirt J. Machine-learning assisted screening for evidence synthesis: a case study of using the asreview tool. SSRN. Preprint posted online on Feb 6, 2025. [doi: 10.2139/ssrn.5125681]
20. 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] [Medline: 37573521]
21. Chan YT, Abad JE, Dibart S, Kernitsky JR. Assessing the article screening efficiency of artificial intelligence for systematic reviews. J Dent. Oct 2024;149:105259. [doi: 10.1016/j.jdent.2024.105259] [Medline: 39067652]
22. Scherhag J, Burgard T. Performance of semi-automated screening using rayyan and asreview: a retrospective analysis of potential work reduction and different stopping rules. PsychArchives. Preprint posted online on May 3, 2023. [doi: 10.23668/psycharchives.12843]

23. Teijema JJ, Ribeiro G, Seuren S, Anadria D, Bagheri A, van de Schoot R. Simulation-based active learning for systematic reviews: a scoping review of literature. J Inf Sci. 2025:01655515251379058. [doi: 10.1177/01655515251379058]
24. Navigating the maze of models in ASReview. ASReview. 2025. URL: https://asreview.nl/blog/asreview-model-selection-guide/ [Accessed 2026-07-22]
25. Boetje J, van de Schoot R. The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. Syst Rev. Mar 1, 2024;13(1):81. [doi: 10.1186/s13643-024-02502-7] [Medline: 38429798]
26. Steijger D, Christie H, Aarts S, IJselsteijn W, Verbeek H, de Vugt M. Use of artificial intelligence to support quality of life of people with dementia: a scoping review. Ageing Res Rev. Jun 2025;108:102741. [doi: 10.1016/j.arr.2025.102741] [Medline: 40188991]
27. Holst D, Moenck K, Koch J, Schmedemann O, Schüppstuhl T. Transparent reporting of AI in systematic literature reviews: development of the PRISMA-trAIce checklist. JMIR AI. Dec 10, 2025;4:e80247. [doi: 10.2196/80247] [Medline: 41370833]
28. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag. Jan 1988;24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]
29. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. Presented at: AAAI-98 Workshop on Learning for Text Categorization; Jul 27, 1998; Madison, Wisconsin, USA. URL: https://cdn.aaai.org/Workshops/1998/WS-98-05/WS98-05-007.pdf [Accessed 2026-07-22]
30. 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. 2021;3(2):125-133. [doi: 10.1038/s42256-020-00287-7]
31. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. Apr 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
32. McNEMAR Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. Jun 1947;12(2):153-157. [doi: 10.1007/BF02295996] [Medline: 20254758]
33. 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. Mar 2024;36(1):19. [doi: 10.1007/s10648-024-09862-5]
34. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statist Sci. 2001;16(2):101-133. [doi: 10.1214/ss/1009213286]
35. Rai K, Tabata K, Sasaki Y, et al. Expanding the feasibility of systematic reviews with AI support: a practical case using ASReview. Igaku Toshokan. 2025;72(3):136-141. [doi: 10.7142/igakutoshokan.72.3_136]
36. Yao X, Kumar MV, Su E, Flores Miranda A, Saha A, Sussman J. Evaluating the efficacy of artificial intelligence tools for the automation of systematic reviews in cancer research: a systematic review. Cancer Epidemiol. Feb 2024;88:102511. [doi: 10.1016/j.canep.2023.102511] [Medline: 38071872]
37. Teijema JJ, de Bruin J, Bagheri A, van de Schoot R. Large-scale simulation study of active learning models for systematic reviews. Int J Data Sci Anal. Nov 2025;20(6):5435-5456. [doi: 10.1007/s41060-025-00777-0]
38. Kempny C, Annac K, Wahidie D, Yilmaz-Aslan Y, Brzoska P. When to stop reviewing: validation of stop criteria in ASReview. BMC Med Res Methodol. May 9, 2026;26(1):109. [doi: 10.1186/s12874-026-02866-5] [Medline: 42106619]
39. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. Jan 14, 2015;4(1):5. [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
40. Sanghera R, Thirunavukarasu AJ, El Khoury M, et al. High-performance automated abstract screening with large language model ensembles. J Am Med Inform Assoc. May 1, 2025;32(5):893-904. [doi: 10.1093/jamia/ocaf050] [Medline: 40119675]
41. Cumpston M, Li T, Page MJ, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev. Oct 3, 2019;10(10):ED000142. [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
42. Gartlehner G, Affengruber L, Titscher V, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol. May 2020;121:20-28. [doi: 10.1016/j.jclinepi.2020.01.005] [Medline: 31972274]
43. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. Jun 1977;33(2):363-374. [Medline: 884196]
44. Zec S, Soriani N, Comoretto R, Baldi I. High agreement and high prevalence: the paradox of Cohen's kappa. Open Nurs J. 2017;11:211-218. [doi: 10.2174/1874434601711010211] [Medline: 29238424]

45. Agresti A. Categorical Data Analysis. John Wiley & Sons; 2013. ISBN: 978-0-470-46363-5
46. Flemyng E, Noel-Storr A, Macura B, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025. Environ Evid. 2025;14(1). [doi: 10.1186/s13750-025-00374-5]
47. Ferdinands G, Schram R, de Bruin J, et al. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the average time to discover relevant records. Syst Rev. Jun 20, 2023;12(1):100. [doi: 10.1186/s13643-023-02257-7] [Medline: 37340494]

Abbreviations

ACM: Association for Computing Machinery
PRISMA-trAIce: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Transparent Reporting of AI-Assisted Evidence Synthesis
RAISE: Responsible Use of AI in Evidence Synthesis
SAFE: Screen, Apply, Find, Evaluate
SBERT: Sentence-Bidirectional Encoder Representations from Transformers
TF-IDF: Term Frequency–Inverse Document Frequency

Edited by Javad Sarvestan; peer-reviewed by Mark Scheper, Yunguo Yu; submitted 29.Sep.2025; final revised version received 09.Jul.2026; accepted 10.Jul.2026; published 20.Aug.2026

Please cite as:
Steijger D, Thissen S, Aarts S, Verbeek H, de Vugt ME, Christie H
Performance of Two AI Approaches in ASReview Compared With Manual Screening for Dementia Care Literature Screening: Comparative Analysis
JMIR Form Res 2026;10:e84989
URL: https://formative.jmir.org/2026/1/e84989
doi: 10.2196/84989