Usability and Accuracy of the SWIFT-ActiveScreener:

Preliminary evaluation for use in clinical research

Jenny J.W. Liu

Natalie Ein

Julia Gervasio

Bethany Easterbrook

Maede S. Nouri

Anthony Nazarov

J. Don Richardson

Abstract

Systematic reviews (SRs) employ standardized methodological processes for synthesizing empirical evidence to answer specific research questions. These processes include rigorous screening phases to determine eligibility of articles against strict inclusion and exclusion criteria. Despite these processes, SRs are a significant undertaking, and this type of research often necessitates extensive human resource requirements, especially when the scope of the review is large. Given the substantial resources and time commitment required, we investigated a way in which the screening process might be accelerated while maintaining high fidelity and adherence to SR processes. More recently, researchers have increasingly turned to artificial intelligence-based (AI) software to expedite the screening process. This paper evaluated the accuracy and usabiity of a novel, machine learning program, Sciome SWIFT-ActiveScreener (ActiveScreener) in a large SR of mental health outcomes following treatment for PTSD. ActiveScreener exceeded the expected 95% accuracy of the program to predict inclusion or exclusion of relevant articles, and was reported to be user friendly by both novice and seasoned screeners. Our results showed that ActiveScreener, when used appropriately, may save considerable time and human resources when performing SR.

Introduction

Systematic reviews (SRs) are the current standard to collate and synthesize empirical evidence and evaluate trends across a specific body of literature in response to specific research questions. SRs involve strict structured and formal methodological processes (Gough et al., 2020; White et al., 2012). Standardized protocols, such as the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) offer researchers a roadmap to conducting SRs with rigour and fidelity (Page et al., 2021). In addition, formal guides established by Cochrane provide additional evaluation criteria in order to provide appropriate context for the interpretation of study data in various research settings (Higgins et al., 2021). Despite these protocols, SRs continue to be a huge undertaking due to extensive resource requirements. Depending on the scope of review and precision of search terms used, researchers may review tens of thousands of articles during various stages of screening. Therefore, given the substantial resource and time commitment required to complete the screening phases for SRs it is crucial to investigate opportunities which may accelerate the screening process. In this paper, we evaluated ActiveScreener in terms of its accuracy and usability in a large SR of mental health outcomes following treatment for PTSD. ActiveScreener was selected namely for its departure from programs that uses AI to identify records, and instead, uses machine learning to build a predictive algorithm to reduce time spent in screening phases of SRs.

The screening phases of a SR includes de-duplicating search outputs across multiple database, and screening title and abstracts and full text (Page et al., 2021). During these steps, researchers examine each article against strict inclusion and exclusion criteria in order to determine its eligibility for inclusion in the SR. To ensure standards of quality, more than one individual must screen the same article independently at each screening stage, with the reliability

between screeners calculated and reported as part of the SR (Belur et al., 2021). Altogether, screening phases can take hundreds of hours for each individual reviewer involved.

Artificial intelligence-based (AI) softwares such as COVIDENCE (Veritas Health Innovation, n.d.), CUREDATIS (Research Solutions, 2023), and Sciome SWIFT-ActiveScreener (Howard et al., 2020) have been developed help expedite SR screening. For the purpose of this paper, AI programs refer to programs that are enabled to perform tasks that normally require human intelligence during in the context of conducting a SR. While they do not eliminate human involvement in the screening process, each program may reduce time and resources spent using various proprietary solutions. For example, COVIDENCE aids clinical research reviews with its ability to distinguish between articles which are randomized controlled trials (RCTs) versus non-RCTs.

ActiveScreener is a novel, machine learning and web-based AI software for SRs. ActiveScreener uses a pretrained machine learning program to identify and prioritize articles for screening (Howard et al., 2020). Based on user feedback via patterns of screening, ActiveScreener uses its pretrained algorithm to build a model estimating articles for inclusions versus exclusions, trimming screening time and effort by nearly 70% (Howard et al., 2020). Past research utilizing ActiveScreener have found the algorithm to work well in reviews involving the physical health literature (Elmore et al., 2020; Lam et al., 2019).

Aims

Despite indications of past use in health reviews, there is little evidence for how ActiveScreener may perform in evaluations of mental health and treatment outcomes. Further, the precision of the estimation model remains unclear. In this paper, we set out to evaluate the precision and usability of ActiveScreener in conducting screening for a mental health treatment

SR (Liu et al., 2021). Specifically, in Part 1, we formally evaluated the accuracy of its predictive model relative to the actual outcomes of screening conducted by individual human screeners, and in Part 2, we collected informal feedback regarding the usability of ActiveScreener amongst a cohort of screeners.

Part 1 – Accuracy

Methods

Procedure

Eighteen screeners were trained to identify articles for inclusion and exclusion and on the use of ActiveScreener for a meta-analysis and systematic review (for more details on this project see Liu et al., 2021). A total of 10002 references required review at the title and abstract stage. ActiveScreener accuracy statistics were set at 95% resulting in 5390 of these references to be reviewed by screeners. Once screening reached 95% accuracy, all screeners stopped. At this stage, data consisting of the screening results for the 5390 references reviewed by screeners and the remaining 4612 references reviewed by the ActiveScreener AI were exported. Accuracy statistics were then reset to 100% prompting the screeners to continue screening the remining 4612 references. Data was once again exported. Screening results for the 4612 references from ActiveScreener and the screeners were then compared.

Data Analysis

A confusion matrix and statistics were generated and used to evaluate the predictive accuracy of ActiveScreener across three classes. The three classes were Included (represents references identified as meeting inclusion criteria), Excluded (representing references identified as meeting exclusion criteria), and Conflicted (representing disagreement on whether the

reference should be included or excluded). Analyses were performed in R-Studio, using the tidyverse (Wickham et al., 2023), stringr (Wickham, 2022), and caret (Kuhn, 2008) packages. Results are reported for only the title and abstract screening stage.

Results

The multiclass confusion matrix for 4612 references is presented in Table 1. As shown, both the screeners and the ActiveScreener AI accurately identified 1365 included references, 2528 excluded references, and 622 conflicted references. For 97 references, the screeners identified these references as included, while the ActiveScreener AI identified these references as conflicted.

Table 1. Confusion Matrix (n = 4612)

		Actual		
		Conflicted	Excluded	Included
Predicted	Conflicted	622	0	97
	Excluded	0	2528	0
	Included	0	0	1365

Notes. Actual = screeners; Predicted = ActiveScreener

Overall, accuracy was 97.9%, 95% CI [0.97, 0.98], $p < .001$. Interrater reliability was reported with Kappa [Fleiss and Conger; 0.96]). Sensitivity for the three classes were: Included (0.93), Excluded (1.00), and Conflicted (1.00). Specificity for the three classes were: Included (1.00), Excluded (1.00), and Conflicted (0.98).

Part 2 - Usability

Methods

Respondents

Eighteen screeners completed a survey around the usability of ActiveScreener. All respondents were paid employees or unpaid volunteers of the MacDonald Franklin Operational Stress Injury Research Centre (MFOSIRC).

Measures

Demographic. Demographic information included: (1) the respondents' role within MFOSIRC, (2) whether the respondents conducted or assisted on a systematic review or meta-analysis prior to their placement at the MFOSIRC, (3) respondents' level of experience with systematic reviews or meta-analyses (e.g., intermediate), and (4) types of software used by respondents for screening for systematic reviews or meta-analyses.

ActiveScreener User Experience Survey. This survey was created by authors (J.J.W.L & A.N) to capture respondents experiences using ActiveScreener. The survey consisted of 12 items (statements or questions) related to usability of ActiveScreener for screening (e.g., "SWIFT Active Screener is easy to use", "SWIFT Active Screener software was easy to learn"). Nine statements were quantitative, and three questions were qualitative. Of the quantitative items, eight statements were rated on a 5-point Likert scale, ranging from strongly agree to strongly disagree, and one question was rated on a 5-point Likert scale, ranging from very confident to not at all confident. The qualitative items included three open-ended questions capturing information related to features of ActiveScreener the respondents enjoyed, any challenges experienced while using ActiveScreener, and any suggestions the respondents had to improve ActiveScreener.

Procedure

All respondents received an email with the link directing them to the online survey. Respondents were asked to complete both the demographic information and the ActiveScreener User Experiences Survey online via Google Forms. Data was collected in April 2022. No direct compensation was given for participating in this study. However, many of the respondents were paid employees of the MFOSIRC and completed the survey during working hours thereby receiving nominal monetary compensation for the time spent participating. For unpaid volunteers, the time spent completing this survey was included in their volunteer hours for which they are provided a letter of recognition.

Data Analysis Plan

Both quantitative and qualitative data was used to provide descriptive information related to the respondents' experiences using ActiveScreener. For qualitative data, common themes were extracted from responses provided regarding enjoyable features of the software, challenges with ActiveScreener and suggested improvements.

Results

Quantitative Data

All 18 respondents completed all nine quantitative items. All respondents (100%) either agreed or strongly agreed that: their training needs were met; ActiveScreener was easy to learn; they felt confident using ActiveScreener; and they would recommend ActiveScreener for use in other reviews. Nearly all respondents reported either agreeing or strongly agreeing that: ActiveScreener was easy to use (94.4%); and ActiveScreener had a user-friendly interface (94.5%). The majority of respondents (88.9%) also reported that they either agreed or strongly agreed that ActiveScreener had all the features needed for adequate screening. Of the eight respondents who had prior experience with other screening programs or tools, seven respondents

(87.5%) rated that they either agreed or strongly agreed that they preferred ActiveScreener over other programs. With regards to the experience of technical or system-related glitches, respondents varied in their perspectives, with 44.5% of respondents indicating that they experienced no technical or system-related glitches (either agreed or strongly agreed), while 22.2% indicated experiencing technical or system-related glitches (disagreed). Results for each survey items are reported in Table 2.

Table 2. Respondent Data Across Measures (N = 18)

	n	%
Demographic Information		
What is your role within MacDonald Franklin OSI Research Centre?		
Project Lead/Co-Lead	5	27.8
Volunteer	8	44.4
Research Assistant (paid, full time)	3	16.7
Research Assistant (paid, part time)	1	5.6
Research Associate	1	5.6
Have you conducted/assisted in a systematic review/meta-analysis prior to your placement with us?		
Yes	9	50.0
No	9	50.0
Level of experience in systematic review/meta-analyses.		
Beginner (assisted in 3 or less)	12	66.7
Intermediate (lead one or engaged in 5 or less)	3	16.7
Advanced (lead multiple/engaged in 5 or more)	3	16.7
What softwares have you used for strictly screening in reviews?[a]		
Swift ActiveScreener	18	100
Microsoft Excel (offline, via 365, or as google doc)	7	33.9
Smartsheets	11	61.1
Covidence	6	33.3
SysREV	1	5.6

EPPI-Reviewer	0	0
Distiller SR	1	5.6
SUMARI	0	0
Reference Management Softwares (e.g., Mendeley, Endnotes, etc.)	0	0
Other	1	5.6
ActiveScreener User Experience Survey (Quantitative Items Only)		
SWIFT ActiveScreener is easy to use.		
Strongly Agree	6	33.3
Agree	11	61.1
Neutral	1	5.6
Disagree	0	0.0
Strongly Disagree	0	0.0
Training to use the SWIFT ActiveScreener met my needs.		
Strongly Agree	11	61.1
Agree	7	38.9
Neutral	0	0.0
Disagree	0	0.0
Strongly Disagree	0	0.0
SWIFT ActiveScreener was easy to learn.		
Strongly Agree	13	72.2
Agree	5	27.8
Neutral	0	0.0
Disagree	0	0.0

Strongly Disagree	0	0.0
SWIFT ActiveScreener has all the features I need for screening.		
Strongly Agree	7	38.9
Agree	9	50.0
Neutral	2	11.1
Disagree	0	0.0
Strongly Disagree	0	0.0
SWIFT ActiveScreener is user friendly.		
Strongly Agree	5	27.8
Agree	12	66.7
Neutral	1	5.6
Disagree	0	0.0
Strongly Disagree	0	0.0
SWIFT ActiveScreener does not have any technical/system glitches.		
Strongly Agree	1	5.6
Agree	7	38.9
Neutral	6	33.3
Disagree	4	22.2
Strongly Disagree	0	0.0
I would recommend SWIFT ActiveScreener for use in screening with other reviews.		
Strongly Agree	10	55.6
Agree	8	44.4

Neutral	0	0.0
Disagree	0	0.0
Strongly Disagree	0	0.0
I prefer SWIFT ActiveScreener over other platforms/softwares for screening.		
Strongly Agree	3	16.7
Agree	4	22.2
Neutral	1	5.6
Disagree	0	0.0
Strongly Disagree	0	0.0
Not Applicable (have used no other software/platforms)	10	55.6
If you were to conduct another systematic review, how confident are you that you would use SWIFT ActiveScreener for citation screening?		
Very confident - will absolutely use ActiveScreener	8	44.4
Confident - most likely will use ActiveScreener	10	55.6
Neutral – no preference	0	0.0
Not Confident – may use other software	0	0.0
Not at all Confident – will definitely use other software	0	0.0

Notes. [a] indicates respondents could choose more than one answer.

Qualitative Data

Features Enjoyed. For the question capturing the features of ActiveScreener enjoyed most by respondents, three primary themes emerged from the data (see Table 3 for quotes).

AI Predictability. Respondents noted that ActiveScreener accelerates the screening process through predictive capabilities. Specifically, ActiveScreener reorders references based on individual patterns of inclusion and exclusion such that likely included articles are pushed to the top of the screening list.

Screening Process. Respondents noted that ActiveScreener makes the screening process easier and faster. Specifically, all the information required for screening is available on one page including the article title, abstract, full text, and inclusion and exclusion criteria. This allows the screener to evaluate the article quickly.

User-friendly Interface. Respondents noted that ActiveScreener has a user-friendly interface. For example, respondents noted ease of use and ability to access ActiveScreener from any device as a positive feature of this software.

Challenges. For the question capturing any challenges experienced by respondents, two primary themes emerged from the data (see Table 3 for quotes).

Technical Issues. Respondents noted that they encountered some technical difficulties and glitches while using ActiveScreener. For example, connection loss specific to the ActiveScreener website or processing or loading speeds were commonly described.

Article Uploading. Respondents noted that uploading articles individually to each reference is time consuming and could result in errors such as a mismatch of articles to references.

Suggested Improvements. For the question capturing suggested improvements or additions to the program, three primary themes emerged from the data (see Table 3 for quotes).

Data Extraction. Respondents noted that they would have liked the ability to either extract data directly within ActiveScreener or be able to export the included references with attached articles to other formats (e.g., SmartSheets).

Bulk Upload. Respondents noted that they would like the ability to upload articles to references in bulk as opposed to one at a time.

Interface Improvements. Respondents noted potential improvements to the user interface. For example, navigation opportunities, keeping a session counter of screened articles, and ability to flag references with incorrect articles attached.

Table 3. ActiveScreener User Experience Survey Qualitative Feedback

Survey Questions	Themes Identified	Examples of Respondent Quotes
What was the ActiveScreener feature you enjoyed the most?	AI Predictability	"It reorders studies based on screening patterns."
	Screening Process	"Having all the information on one page (title/abstract/full text) to decide whether to include or exclude. Love highlighting keywords."
	User-friendly Interface	"Simplicity of user interface."
What are some of the challenges you experienced	Technical Issues	"Random software glitches where we had to reach out to the

with ActiveScreener?		ActiveScreener team to find out what was happening."
	Article Uploading	"Uploading full text articles to the individual record."
What are some features you wish ActiveScreener would improve or add?	Data Extraction	"Making data extraction possible or easy to transfer all data to smartsheets with articles attached."
	Bulk Upload	"Bulk upload."
	Interface Improvements	"Being able to skip an abstract for the duration of a session (e.g., when a paper was attached to an incorrect abstract, I would skip it and go to the next abstract - but upon completion of the next abstract, the incorrectly-matched one would be next in queue.) Would be nice to be able to skip/flag/set aside without having to navigate away from it repeatedly."

Discussion

In our study, we found that ActiveScreener performed above its expected 95% accuracy in prediction and was found to be user friendly by both novice and seasoned screeners. Consistent with past evidence that the effectiveness of this program can reduce screening time and effort by nearly 50% (Howard et al., 2016), we observed similar results with a large-scale review of PTSD treatment outcomes.

Regarding its accuracy, our confusion matrix results indicated that when testing against a large-scale SR which included over 10000 articles screened in the title and abstract phase, ActiveScreener performed better than expected in its predictive algorithm. While the software was expected to reach 95% accuracy, the actual accuracy of its machine learning model in our review exceeded 95% (97.9%). Further, of the categories of accuracy examined, discrepancies between the predictive algorithm and actual human screening outcomes were minimal. Specifically, there were no discrepancies between human screeners and the ActiveScreener AI with respect to articles that should be excluded from the SR. Only a small number of discrepancies were found between human screeners that indicated articles should be included while the ActiveScreener AI predicted that the articles would be conflicted (i.e., predicted multiple human screeners would disagree on inclusion and exclusion) based on prior trends in human screening. This means that no studies that the ActiveScreener AI predicted to be included resulted in exclusions by screeners. Thus, these accuracy statistics indicate that ActiveScreener is a reliable and rigorous platform to accelerate screening at the title and abstract phase of SRs, especially when utilizing its predictive algorithm function.

In examining user feedback amongst a group of screeners, we found that ActiveScreener was endorsed as easy to learn and easy to use. However, user feedback also noted that there were

software glitches, such as the platform being unavailable from time to time, as well as glitches when uploading articles and using other features. While these challenges do not undermine its use, they provide areas of opportunity for ActiveScreener programmers to consider for future research and development. Further, to reduce human resources during screening, ActiveScreener should consider implementing new features such as bulk upload and templates for subsequent data extraction directly within the platform. Both would reduce the need for switching between programs when conducting reviews, and thereby reduce human resource requirements as well as potentials for errors.

Conclusion

In considering the merits of ActiveScreener, it should be noted that the software's machine learning algorithm is reliant on the rigour of training and the strength of screeners that it bases its user feedback on. As such, users must conduct training and screening with care. In particular, the clarity in which inclusion and exclusion criteria may be applied during the initial screening stages is of vital importance in building the accuracy of the predictive model. Thus, researchers are encouraged to spend considerable time to ensure the inclusion and exclusion criteria are clearly understood and reliably applied by all screeners during the project training stages. In addition, another time-saving feature of ActiveScreener, the deduplication function for uploading references can benefit from further development as it currently limits the deduplication to texts only, and does not extend to cover punctuation. Depending on the database, references may be exported with variable punctuations, which is not covered by the feature, resulting in many duplicate references when screening. However, it should be noted that this can easily be solved with work-arounds, such as manually combining search yields on *r* with generated codes that deduplicates references prior to uploading on ActiveScreener. Finally, it is

important to note that ActiveScreener's program to accelerate the screening stage is only currently relevant at the title/abstract stage and excludes further reviews of full-text. And thus, current study findings and the potential time and resource savings are only applicable to the initial screening phase of SRs. Taken together, ActiveScreener appears to be a user friendly and accurate platform for SRs, and when used appropriately, may save considerable time and human resources during the initial screening process.

References

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. *Sociological Methods & Research*, *50*(2), 837-865. https://doi.org/10.1177/0049124118799372

Elmore, R., Schmidt, L., Lam, J., Howard, B. E., Tandon, A., Norman, C., Phillips, J., Shah, M., Patel, S., Albert, T., Taxman, D. J., & Shah, R. R. (2020). Risk and Protective Factors in the COVID-19 Pandemic: A Rapid Evidence Map. *Frontiers in public health*, *8*, 582205. https://doi.org/10.3389/fpubh.2020.582205

Gough, D., Davies, P., Jamtvedt, G., Langlois, E., Littell, J., Lotfi, T., Masset, E., Merlin, T., Pullin, A. S., Ritskes-Hoitinga, M., Røttingen, J. A., Sena, E., Stewart, R., Tovey, D., White, H., Yost, J., Lund, H., & Grimshaw, J. (2020). Evidence Synthesis International (ESI): Position Statement. *Syst Rev*, *9*(1), 155.

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (2021, February). *Cochrane Handbook for Systematic Reviews of Interventions version 6.2*. Cochrane Handbook for Systematic Reviews of Interventions Version 6.2. Retrieved July 7, 2023, from www.training.cochrane.org/handbook

Howard, B. E., Phillips, J., Tandon, A., Maharana, A., Elmore, R., Mav, D., Sedykh, A., Thayer, K., Merrick, B. A., Walker, V., Rooney, A., & Shah, R. R. (2020). SWIFT ActiveScreener: Accelerated document screening through active learning and integrated recall estimation. *Environ Int*, *138*, 105623. https://doi.org/10.1016/j.envint.2020.105623

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lam, J., Howard, B. E., Thayer, K., & Shah, R. R. (2019). Low-calorie sweeteners and health

outcomes: A demonstration of rapid evidence mapping (rEM). *Environmentinternational*, *123*, 451–458. https://doi.org/10.1016/j.envint.2018.11.070

Liu, J. J. W., Nazarov, A., Easterbrook, B., Plouffe, R. A., Le, T., Forchuk, C., Brandwood, A., St Cyr, K., Auger, E., Balderson, K., Bilodeau, M., Burhan, A. M., Enns, M. W., Smith, P., Hosseiny, F., Dupuis, G., Roth, M., Mota, N., Lavoie, V., . . . Richardson, J. D. (2021). Four Decades of Military Posttraumatic Stress: Protocol for a Meta-analysis and Systematic Review of Treatment Approaches and Efficacy. *JMIR Res Protoc*, *10*, e33151.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71.

Research Solutions (n.d.). *Curedatis Systematic Review Engine*. Retrieved July 7, 2023, from https://www.researchsolutions.com/curedatis