This is a “preproof” accepted article for *Journal of Clinical and Translational Science*.
This version may be subject to change during the production process.
10.1017/cts.2025.10173

BRIEF REPORT

Machine-learning assisted screening for evidence synthesis: methodological case study of the ASReview tool

Kim Boesen [1], Pascal Dueblin [2], Lars G. Hemkens [1], Perrine Janiaud [1], Julian Hirt [1,3,4]

[1]Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University of Basel and University Hospital Basel, Basel, Switzerland

[2]Department of Clinical Research, University of Basel and University Hospital Basel, Basel, Switzerland

[3]Department of Health, Eastern Switzerland University of Applied Sciences, St.Gallen, Switzerland

[4]Institute of Health and Nursing Science, Medical Faculty, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

Corresponding author: Julian Hirt, Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Spitalstrasse 2, CH-4031 Basel, Switzerland, julian.hirt@usb.ch

Abstract

ASReview is a software that can potentially reduce the workload of literature screening in systematic reviews by ranking the retrieved records. We assessed the tool's feasibility, advantages and limitations, to populate a database of cancer immunotherapy trials. ASReview is easy to use, and it efficiently identified relevant records. It may save resources compared to traditional systematic reviews using two human reviewers. Predefined procedures are necessary to maintain a transparent and reproducible workflow. Limitations include that adding references to existing projects is difficult and that the algorithm learns from every decision, even when this may not be appropriate.

Keywords: Evidence synthesis, Study selection, ASReview, Review software, Artificial intelligence

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

Background

Literature screening and study selection is a one of the most resource consuming tasks during the conduct of a systematic review.[1] During comprehensive literature searches for a systematic review, reviewers usually screen hundreds or thousands of records obtained from various databases, such as PubMed, in random order to identify the (often few) relevant hits to include in the systematic review. Machine learning algorithms have been developed to reduce the workload of manual screening by reordering and reranking the obtained records from the database searches based on their relevance.[2,3]

ASReview is one such machine-learning assisted screening tool developed at the University of Utrecht, the Netherlands, in 2019.[4,5] It is open-source, free to use for non-commercial use, and it has an extensive GitHub community.[6] ASReview employs a supervised machine learning algorithm, which is updated after each decision made by the reviewer, to continuously reorder the records.[7] The reviewer may choose to screen all or only a proportion of the full sample assuming that most, or all, relevant records are shown among the first records. A quick summary of how to use ASReview can be found in Appendix 1.

ASReview's developers reported that one had to screen 8% to 33% of all records to identify 95% of relevant studies ("95% recall") based on four simulation studies.[5] A case study in health economics[8] reported to screen 8% of their sample to obtain 100% recall. Based on these case studies, one may save up to 92% of time and resources at the cost of missing 5% of relevant hits. A collection of scientific articles related to ASReview can be found elsewhere.[9] We tested ASReview to reduce the workload related to literature screening for our continuously updated database of immunotherapy trials, the Cancer Immunotherapy Evidence Living (CIEL) Library.[10,11]

Materials and methods

Our case study's objectives were to test ASReview's feasibility and to identify advantages and limitations compared to a traditional literature screening setup with two human reviewers screening all retrieved records in random order. We did not formally compare sensitivity/specificity and used resources for an ASReview assisted single-screener versus a traditional double-screener setup.

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

Information source and search strategy

We systematically searched PubMed as of May 2, 2023 (Appendix 2). We included interventional trials of tumor-infiltrating lymphocytes as the treatment for any type of cancer,[10] regardless of the trial design (e.g. single-arm or parallel group), blinding, randomization, and type of control (e.g. placebo, another cancer drug, usual care, or no treatment).

ASReview configuration

We used ASReview (version 1.2, Python 3.8.16, published April 12, 2023) and ASReview's default configuration (naïve Bayes classifier; term-frequency- inverse document frequency; certainty-based sampling; and dynamic resampling). One reviewer (KB) screened title/abstract and full-texts, and ambiguous cases were resolved with another reviewer (PJ). Records were screened based on title/abstract. When this was not enough to decide, the full text was obtained to make a final decision about inclusion or exclusion using the digital object identifier (DOI). If the DOI was not available, we searched PubMed or Google using the verbatim title.

Two-phase screening workflow

We split our screening procedure into two phases: In **phase 1 (enriched training phase),** we created an enriched 'training sample' by using the PubMed filter "Clinical study". We anticipated that many relevant records would be tagged with this filter. One reviewer (KB) screened all retrieved records and labelled them as 'relevant' or 'irrelevant'. In **phase 2 (screening phase)**, we then merged the fully labelled phase 1 dataset with the remaining unscreened pool of records. This way we trained the algorithm using the training sample from phase 1 (all records now labelled as 'relevant' or 'irrelevant'). We applied an arbitrary stopping rule of 100 consecutive irrelevant records, which has been used in another case study.[12]

Evaluation

We narratively evaluated the tool's feasibility regarding setup and interface. We noted whether there were important advantages and limitations compared to traditional screening setups of two human reviewers screening all records.

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

Results

Search results and training phase (phase 1)

The PubMed search returned 14.004 records. In phase 1 (enriched training phase), 604 records were tagged with the "Clinical Study" filter and imported into ASReview (Figure 1). We selected 5 known irrelevant and 20 known relevant records as 'prior knowledge', i.e., ineligible and eligible tumor-infiltrating lymphocyte clinical trial publications that we knew already. See recall curve for phase 1 in Figure 2.

The screening phase (phase 2)

Using ASReview's DataTools package,[13] we merged our sample of labelled records from phase 1 with the remaining unscreened sample of PubMed records (i.e. those records not filtered with 'Clinical Study') for a final sample of 13.994 records (10 records were deduplicated). This partly labelled dataset was then screened. We screened 872 further records before reaching our stopping rule of 100 consecutive irrelevant records. We included 59 further eligible records, see recall curve in Figure 2.

Screening summary

In total, we screened 1476 records (11%) of the total sample; 136 (0.97%) were relevant and 1340 were irrelevant. We stopped screening after 100 consecutive irrelevant records, at which there were 12.518 (89%) unscreened records. As we did not screen the full sample, we do not know the true number of relevant records and cannot estimate the 50%, 95% and 100% recall rates.

Evaluation

We identified several advantages and limitations compared to traditional screening tools, summarized below and highlighted in Table 1.

Setup

Compared to other commonly used screening software (e.g. Rayyan or Covidence), ASReview does not provide a 'Software as a Service' product, where the user logs into a website and uses the software without having to install it locally. Therefore, set up is required on a local computer ('local installation') or on a server/cloud provider ('ASReview Lab Server'). The installation requires no knowledge of Python, other than the ability to install Python on one's own computer. Some features to prepare and modify datasets are not part of

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

the ASReview main code base, but are offered by ASReview as separate Python packages, e.g. “ASReview Datatools”[13]. The functionalities of these Python packages are accessible via the command line and requires basic knowledge about working with Unix shells.

Mandatory decision-making

During the screening of records in ASReview, the user must always make a final decision whether to include or exclude a record before proceeding to the next record. It is not possible to ‘skip’ a record while further information is retrieved. In conventional literature screening tools, like Covidence or Rayyan, one can skip a record until there is sufficient information to make a final decision. Often immediate decisions cannot be made due to incomplete information, e.g. if the article is behind paywall; the DOI is missing and one has to manually search for the title; the record is hidden in a journal supplement, or the study authors must be contacted to retrieve additional information.[14,15] In these scenarios it is a limitation that one cannot postpone the decision for a certain record, while working in parallel to solve the uncertainty. One potential solution is to in- or exclude the record tentatively and add a note (e.g. “awaiting full text access”). These records can then later be retrieved (using the “find notes” filter) and the decisions be reverted, if necessary. Each decision influences the algorithm and it may influence ASReview’s performance if irrelevant records are tentatively included. The obligatory decision-making will lead to scenarios that require prespecified rules to ensure a standardised and reproducible workflow.

In scenarios when there are no abstracts and/or the decisive information comes from the full text (which ASReview has no access to) or from other external information, such as author correspondence, it would be desirable with a feature to include records without training the algorithm. Such inclusions may potentially obscure the algorithm’s performance.

Single-screener setup

The default single-reviewer setup may worry traditional systematic reviewers who are used to a two-reviewer setup. One may argue that a single reviewer assisted with ASReview as a secondary reviewer can substitute a traditional two-reviewer setup. Empirical testing and comparison of ASReview assisted single-screener versus double-screener setups are needed to assess the specificity, sensitivity, and used resources of each setup. It is possible to set up a double reviewer workflow having two reviewers screen the same dataset (using the same algorithm configuration and training it on the same records); the reviewers can screen independently, and once a target sample is reached (x percentage of total sample or y number

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

of consecutive irrelevant records), the datasets can be exported (e.g. as excel files). The files can then be compared and uncertainties can be resolved like in regular systematic reviews. It should also be highlighted that it is possible to work collaboratively multiple screeners on the same dataset (using ASReview Lab Server). This may be desirable when screening very large datasets or when records need to be screen within a certain timeframe. Multiple screener setups may likely introduce interrater variability, which should be acknowledged.

The interface links to the full text using the DOI only. In phase 1 (training phase), the DOI was missing for 10% of the records. The developers may consider including additional full text identifiers, such as the PubMed Identifier (PMID), which is available from imported PubMed records.

Discussion

A well-planned use of ASReview may be beneficial in many evidence synthesis projects requiring manual screening of large numbers of records. We cannot make generalizable estimations on the time saving potential based on our single methodological case study.

The potential time and resource savings must be weighed against the risk and impact of missing eligible studies if one decides to not screen all records. For topics with a high risk of missing relevant hits (e.g. a topic new to the reviewers, or a topic with heterogenous terminology), it may be difficult to prespecify a minimum of records to screen, for example 5, 10 or 50% of the total sample. Stopping rules on how to safely use active learning systems like ASReview have been published by the ASReview developers.[16] The authors propose to screen a certain proportion (e.g. 10%) of the total sample and that a certain number of consecutive irrelevant records must be passed (50 records is mentioned). These recommendations seem to be arbitrarily selected and are not based on empirical testing.

Another strategy is to use ASReview exclusively as a second reviewer and have the human reviewer screen all retrieved records. This method (theoretically) reduces the human screening hours with 50% in comparison with a standard double-screener setup.

The generation of “evidence corpuses”, i.e. labelled datasets adhering to specific population-intervention-comparison-outcome (PICO) inclusion criteria could be used and shared across different research projects. For instance, our CIEL dataset of tumor-infiltrating lymphocyte trial publications may be used to train algorithms in other systematic reviews on the same, or similar, topics. Importantly, in our use case of building a continuously updated database of

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

clinical trials, the value of using ASReview may likely increase for each update of the PubMed search as the algorithm is refined.

Limitations of our study

In phase 1 (enriched training phase), we knew many (20 of 77; 25%) of the relevant records, which may have inflated the algorithm's performance in comparison to a smaller training sample. The arbitrary stopping rule of 100 consecutive records must also be questioned. It is important to empirically assess the "sweet spot" between searching the lowest number of records with the highest recall rate. This will likely have to be decided on a case-by-case situation and it likely depends on the sample size (larger sample, smaller threshold?), field and consistency of terminology across publications, and the reviewers' prior knowledge of the literature and thus the size of the training sample.

We screened 11% of the sample and we may have missed entire clusters of eligible records if these appeared further down the ranking. We did not systematically note which records were full-text assessed, so we cannot ascertain whether there were systematic differences (e.g. year of publication or journal) between records assessed based on title/abstract or full-text. Finally, our case study pertained to one clinical topic, thus the generalisability is limited.

Conclusions

Machine learning assisted software like ASReview has the potential to revolutionise the study selection in systematic reviews. Users must be aware of several caveats while using such supervised machine learning algorithms, which are constantly updated and where every decision affects the performance of the software. If a thorough 'scenario-based' protocol is prespecified, the benefits may exceed the harms. Systematic reviewers are encouraged to test such machine learning tools and to publish their results to help establish an empirical foundation to guide best practice of transparent use of artificial assisted screening software in evidence synthesis.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

Availability of data and materials

There are no data associated with this article. The original ASReview project file can be obtained from the authors by request.

Funding

The CIEL-Library project was funded by the Krebsliga beider Basel (grant KLbB-5577-02-2022). The funder did not have any influence on the design and results of this work.

Conflicts of interest

RC2NB (Research Center for Clinical Neuroimmunology and Neuroscience Basel) is supported by Foundation Clinical Neuroimmunology and Neuroscience Basel, unrelated to this work. RC2NB has a contract with Roche for a steering committee participation of LGH, unrelated to this work.

References

1. Nussbaumer-Streit B, Ellen M, Klerings I, et al. Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol 2021;139:287–96.
2. Cierco Jimenez R, Lee T, Rosillo N, et al. Machine learning computational tools to assist the performance of systematic reviews: A mapping review. BMC Med Res Methodol 2022;22:322.
3. Rens van de schoot. Comprehensive Guide to Machine Learning Software for Text Screening. https://github.com/Rensvandeschoot/software-overview-machine-learning-for-screening-text. Accessed 8 September 2025.
4. ASReview. https://asreview.nl/. Accessed 10 June 2025.
5. Schoot RVD, Bruin JD, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence 2021; 3:125-133.
6. ASReview. https://github.com/asreview/asreview. Accessed 10 June 2025.
7. IBM. What is supervised learning? https://www.ibm.com/topics/supervised-learning. Accessed 10 June 2025.
8. Wolcherink MJO, 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; 13:1-8.

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

9. Zotero library ASReview public. https://www.zotero.org/groups/4597652/asreview_public/collections/IWU8ATMK/tags/multiple_screeners/collection. Accessed 10 June 2025.
10. Cancer Immunotherapy Evidence Living (CIEL) Library. https://ciel-library.org/. Accessed 10 June 2025.
11. Boesen K, Hirt J, Düblin P, et al. Rationale and design of the Cancer Immunotherapy Evidence Living (CIEL) Library. A continuously updated clinical trial database of cancer immunotherapies. 2023. Preprint, submitted for publication. https://doi.org/10.1101/2024.04.26.24306436
12. Van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, et al. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023;13:e072254.
13. ASReview datatools. Release v1.1.1, 18 Nov 2022. https://github.com/asreview/asreview-datatools#data-compose-experimental. Accessed 10 June 2025.
14. Mullan RJ, Flynn DN, Carlberg B, et al. Systematic reviewers commonly contact study authors but do so with limited rigor. J Clin Epidemiol 2009;62(2):138–42.
15. Meursinge Reynders R, Ladu L, Di Girolamo N. Contacting of authors modified crucial outcomes of systematic reviews but was poorly reported, not systematic, and produced conflicting results. J Clin Epidemiol 2019;115:64–76.
16. 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 2024;13:81.

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

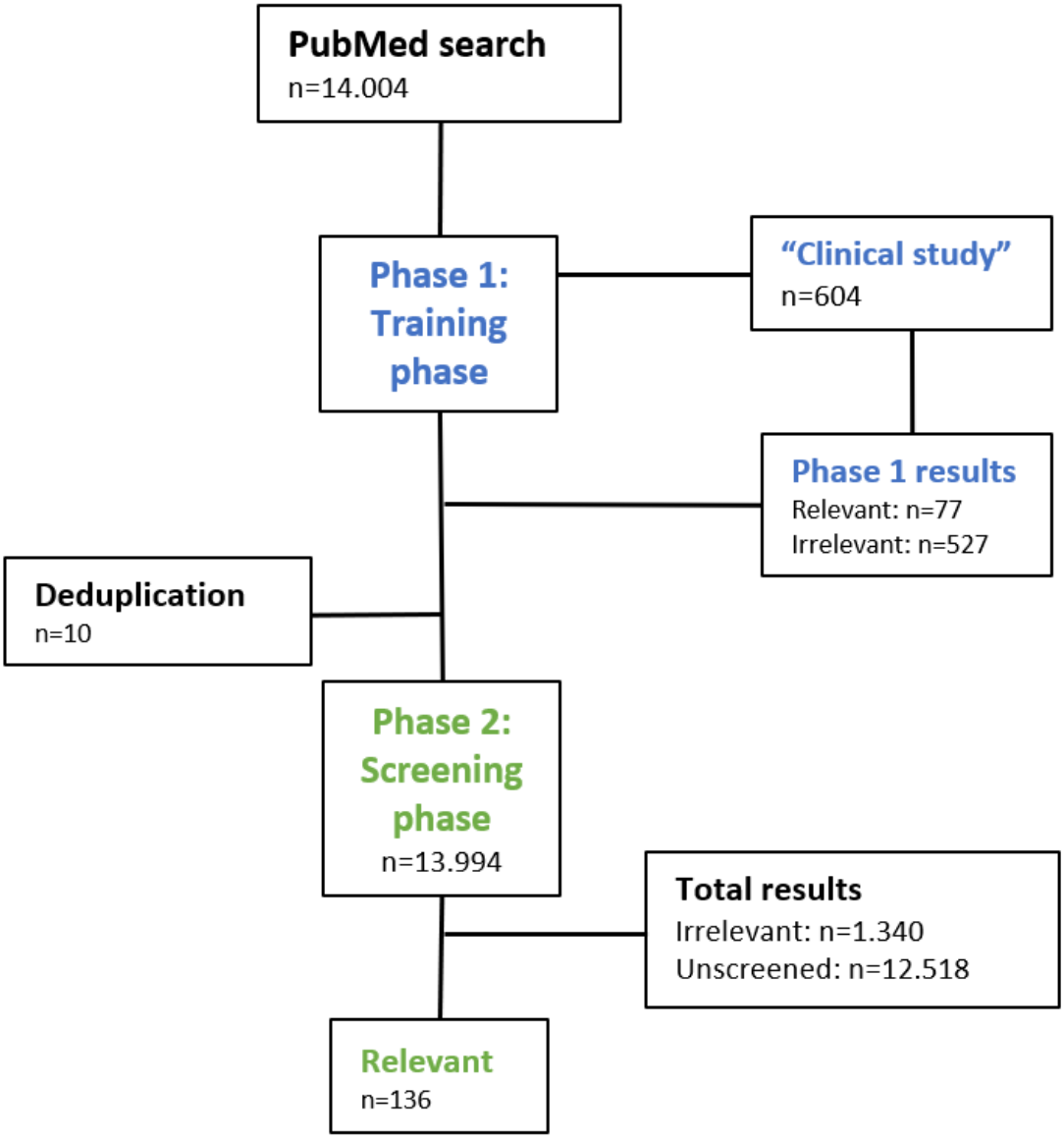

Figure 1. The two-phase screening process

Abbreviations: n = Number.

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

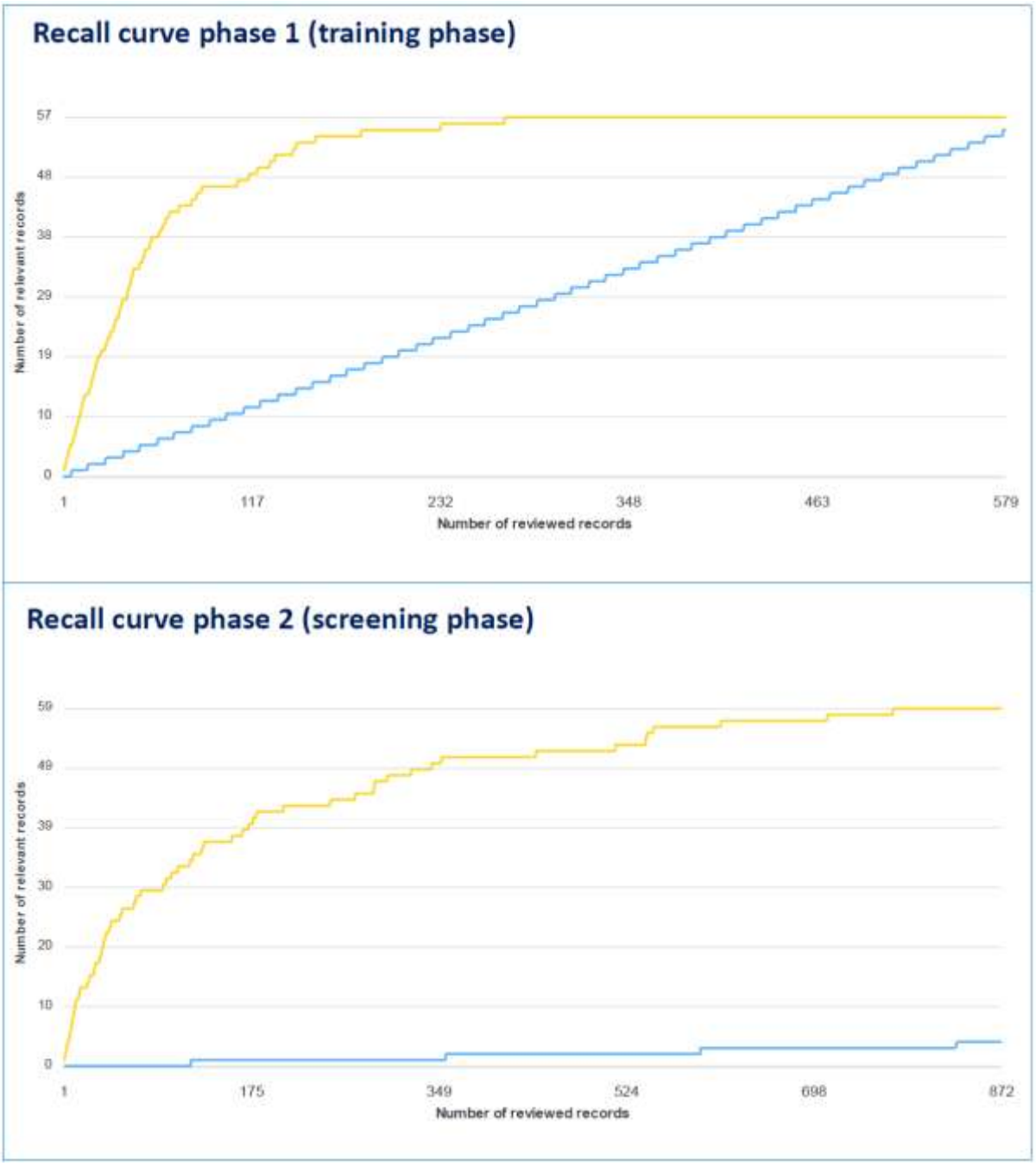

Figure 2. Recall curves

Note: The yellow lines show the identified relevant records, and the blue lines show the theoretical recall curve if the records were screened and identified randomly

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press

Table 1. Advantages and limitations of ASReview

Advantages	**Impact on research**
Open source and free to use	Easy access to advanced software
Effective machine learning algorithm	Efficient and steep recall curves
Interface	Easy to use and navigate
Reliability	Running smoothly without interruptions or excessive waiting
Limitations	**Impact on research**
Challenging setup	The program requires Python, which might challenge some users
No 'skip' function	The reviewer must make an immediate decision to include or exclude, which may not be possible
No 'include without learning' option	The algorithm learns from all decisions, also when this may not be appropriate
No PMID full-text retrieval	Bibliographic details and full text are linked through DOIs only
Default single user setup only	It is possible to set up a multiple reviewer workflow but it requires some planning
No option to easily add records to existing project	Adding records requires external Datatools, which might challenge some users

Abbreviation: DOI = Digital object identifier; PMID = PubMed ID.

https://doi.org/10.1017/cts.2025.10173 Published online by Cambridge University Press