SWAR 62: Custom machine learning classifiers for study selection in systematic reviews or maps of research: a living study within a review (living SWAR)

Objective of this SWAR
To assess the performance of custom binary machine learning classifiers for selecting eligible records to be retained for screening for potential inclusion in systematic reviews or maps of research.

Study Area: Study identification
Sample Type: Records of studies
Estimated Funding Level Needed: Very low

Background
Selecting eligible study reports for inclusion in systematic reviews or maps of research can be time-consuming. For systematic reviews, this process conventionally involves two researchers working independently to screen each title-abstract record in duplicate, followed by meeting to resolve any conflicts between their decisions. For maps of research, the process may involve one researcher screening each record (possibly after an initial phase when two researchers screen each record) but there are often more records to be screened compared with systematic reviews. If study selection could be fully or semi-automated with minimal risk of losing eligible studies, this would improve workflow efficiency and reduce the time it takes to produce evidence syntheses.

Custom binary machine learning (ML) classifiers are underpinned by ML algorithms trained to distinguish between included (positive class) and excluded (negative class) title-abstract records based on their text features. Tools that enable users to train, calibrate, evaluate and deploy custom binary ML classifiers of this kind to select records that are likely to be eligible to be retained for screening, whilst simultaneously discarding records that are unlikely to be eligible immediately prior to screening, have been available in systematic review software for more than a decade (for example, in EPPI Reviewer). However, despite this kind of ML tool being straightforward, low cost and safe to deploy, with the potential to substantively reduce title-abstract screening workload in reviews while keeping the loss of eligible records (reports) to an acceptably low level that is pre-specified by the user, they are underused in the production of reviews and maps (even taking into account that, in many reviews, it may be feasible, within available resources, to screen all of the unique records retrieved by searches).

In a new review or map, the process of training, calibrating, evaluating and (subject to the results) deploying a custom binary ML classifier can only begin when at least some metadata has accumulated from the manual screening of records. In general, the more included and excluded records available for this, the better custom binary ML classifiers are likely to perform at the task. As such, these ML classifiers are best deployed in ‘larger’ systematic reviews and in maps of research (which are typically ‘larger’ than systematic reviews in terms of numbers of includes and screening workload); as well as in those that will be periodically or continually updated in future, including living systematic reviews and living evidence maps, where they can also have the most impact in terms of screening workload reduction.

In projects with longer time horizons, it may be feasible to repeat the process of training, calibrating, evaluating and deploying a custom binary ML classifier several times, to harness the accumulating numbers of included and excluded records. Conversely, in new reviews or maps, if training, calibration and evaluation of a custom binary ML classifier is attempted too early, there may be insufficient screening metadata to identify a reliable threshold score. In

such a case, deployment may be postponed pending a repeat of the process with more conclusive results, using more metadata when this has accumulated.

Although a standard process of training, calibrating, evaluating and deploying custom binary ML classifiers can be completed by a trained researcher using the appropriate tools in less than three hours – including production of the standard study dataset, analysis and report of the results from a replication of this SWAR – the evidence base for their performance is currently fragmented. This living SWAR aims to produce cumulative evidence for the performance of custom binary ML classifiers for selecting eligible records by continually updating a prospective meta-analysis that will integrate the results of new replications of the SWAR as these become available.

Interventions and Comparators
Intervention 1: A custom binary ML classifier, trained to distinguish between eligible and ineligible title-abstract records, used to select records to be retained for screening, followed by single or duplicate manual screening of the retained records (Intervention).
Intervention 2: Single or duplicate manual screening of all records (Comparator).

Index Type: Full review

Method for Allocating to Intervention or Comparator: Not applicable.

Outcome Measures
Each replication of the SWAR
Primary Outcomes:
- Recall: percentage of eligible records retained for screening (i.e. those scoring above the identified threshold score) among evaluation set records.
- Precision: percentage of records retained for screening (i.e. those scoring above the identified threshold score) that are eligible among evaluation set records.
- Workload reduction: percentage reduction in screening workload among evaluation set records (i.e. the reduction from discarding records scoring below the identified threshold score).

Secondary Outcomes:
- True positives (TP): Number of eligible records retained for screening (i.e. those scoring above the identified threshold score) among evaluation set records.
- False positives (FP): Number of ineligible records retained for screening (i.e. those scoring above the identified threshold score) among evaluation set records.
- False negatives (FN): Number of eligible records discarded prior to screening (i.e. those scoring below the identified threshold score) among evaluation set records.
- True negatives (TN): Number of ineligible records discarded prior to screening (i.e. those scoring below the identified threshold score) among evaluation set records.

Prospective meta-analyses of multiple SWAR replications
Primary outcome:
- Recall, precision and screening workload, compared with current standard practice.

Potential effect modifiers (between-studies):-
- Use scenario: Systematic review or map of research / Calibration threshold recall: 0.95 or 0.99
- Sample size: Total number of records used in the SWAR replication
- Stratified random sampling allocation ratio: 4:1:1 or 3:1:1
- Excluded records sample: Full or random sample of excluded records
    - Size of random sample as a percentage of full sample of excludes (if applicable)

- Screening method: Single screening or double screening

Analysis Plans
This is a living SWAR and we expect to make refinements to its design and methods as we accumulate experience of implementing / replicating the design in host reviews and maps of research.

Each replication of the SWAR
All bibliographic records (title-abstract and title-only) manually screened by one ('single screening') or more ('duplicate screening) for potential inclusion in the host review or map from inception to present, on title-abstract and/or full-text – excluding trials registry records (but including all other types of publications) – will be collated into the following three sets:

1. Excluded on title-abstract[1]
2. Included on title-abstract
3. Included on full-text

Stratified random sampling will be used to randomly assign records from two of the above three sets of records – those excluded on title-abstract, and either those included on title-abstract, or those included on full-text[2] – to one of three of the following subsets, using an allocation ratio of either 3:1:1 or 4:1:1[3]: training, calibration, or evaluation.

Next, records in the training set will be used to train a custom binary ML classifier, with included records used as the positive class, and excluded records used as the negative class. Records in the calibration set will be classified using the custom binary ML classifier trained at the preceding step, and the results will be used to determine a classification threshold score at a pre-specified threshold level of recall of either 0.99 for systematic reviews, or 0.95 for maps of research. Finally, records in the evaluation set will be classified using the same custom binary ML classifier, and the classification threshold score set at the preceding step will be applied to either retain (above the threshold score) or discard records (below the threshold score), enabling recall, precision and workload reduction to be computed. Results should be reported using a standard template (an Excel sheet and a Word document containing standardised / boilerplate text to be adapted for each replication of this study).

---

[1] If the number of records in the 'excluded on title-abstract' set is very large (i.e. the set comprises at least tens of thousands of records), it may be judged reasonable to choose the option of drawing a random sample of excludes of a size that at least exceeds (but could be double, treble or more) the number of includes in the selected set (i.e. 'included on title-abstract or 'included on full-text') before the stratified random sampling step, in order to reduce the time needed to train, calibrate, evaluate and (contingent on results) deploy the custom binary ML classifier.

[2] In principle, it is more conservative to deploy a custom binary ML classifier that has been trained using 'included on title-abstract' records as the positive class before the title-abstract screening stage, compared with one trained using 'included on full-text' records only as the positive class; however, choosing the latter option may be judged reasonable when the number of 'included on full-text' records is large (i.e. the latter set comprises at least hundreds of records). The option selected at this stage (i.e. which records will be used in the positive class for training, calibration and evaluation) should reflect the potential deployment plan; however, if there is no plan to deploy, the default option should be to use 'included on title-abstract' records as the positive class (as this is more conservative).

[3] In principle, allocating larger numbers of records to the training set (i.e. choosing a 4:1:1 ratio over a 3:1:1 ratio) will result in a better performing custom binary ML classifier. However, the latter needs to be balanced against having sufficient numbers of records (this applies especially to included records) in the calibration and evaluation sets to enable a stable classification threshold to be set, and for the evaluation to be robust. As such, the 3:1:1 ratio chosen if the numbers of records in both the included and excluded sets is large (i.e. both sets comprise at least hundreds of records).

Contingent on the results among evaluation set records, the researchers undertaking a replication of this SWAR may decide to proceed to prospective deployment. The prospective deployment step is outside of the scope of the SWAR. However, in this case, we recommend considering the option of retraining a new version of the same custom binary ML classifier using all of the same included and excluded records assembled at the first step of the SWAR (i.e. 100% of the records subsequently allocated to training, calibration or evaluation sets), prospectively deploying it to classify previously unseen records, and discarding those records which score below the same classification threshold score already set in the calibration step of the SWAR.

Prospective meta-analyses of multiple SWAR replications

Meta-analysis methods primarily developed for studies of diagnostic test accuracy will be adapted and deployed to produce pooled estimates of the precision and recall of the various approaches to screening evaluated in this SWAR [1]. Prospective meta-analyses will be continually updated as new study-level data from each replication of the SWAR become available (that is, a 'living SWAR' approach).

We will conduct separate sets of meta-analyses of the results of those replications of this SWAR that used:

A: Included on title-abstract records as the positive class; or
B: Included on full-text records as the positive class

Within each set of analyses (A and B, above) we will conduct a pairwise meta-regression analysis, primarily designed to assess the performance of custom binary ML classifiers (the intervention) for selecting eligible records to be retained for screening (followed by single or duplicate manual screening of the retained records), compared with current standard practice (either single screening or duplicate screening of all records by human(s) – the comparator) in terms of their recall, precision and associated screening workload (outcome measures).

- Analysis 1 / Comparison 1: Custom binary machine learning classifier versus single or duplicate manual screening

For the intervention, study-level (replication-level) data on True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN), versus the quasi-gold standard comparator (single or duplicate manual screening), will be used to compute study-level estimates of recall, precision and screening workload, compared with the quasi-gold standard comparator, using standard formulae. Next, study-level estimates of recall, precision and screening workload will be transformed to the logit scale, and associated standard errors and 95% confidence intervals will be computed. Transformed study-level estimates will then be combined using multivariate random effects meta-analysis [2, 3] in order to account for the correlation between recall, precision and screening workload.

Following the incorporation of five replications of this SWAR, each meta-analysis described above will be extended to investigate potential sources of between-studies heterogeneity in estimates of recall and precision. When five replications have been incorporated into a meta-analysis, a final bivariate random effects meta-regression model will incorporate the first of the potential effect modifiers listed in the previous section, collected for each replication of the SWAR, as a covariate (explanatory variable). When ten replications have been incorporated into a meta-analysis, the meta-regression model will incorporate the second of the potential effect modifiers listed in the previous section as a covariate. After 15 replications, the third-listed potential effect modifier will be incorporated. After 20 replications, the fourth-listed potential effect modifier will be incorporated; and, finally, after

25 replications, the fifth-listed potential effect modifier will be incorporated. A method for selecting and/or deselecting potential effect modifiers into / from the final meta-regression models will be pre-specified in advance of the baseline meta-analyses (due after at least two replications of this SWAR) by a researcher blinded to these metadata.

It is feasible and desirable that this prospective meta-analysis will need to handle outcome data submitted from multiple replications of this SWAR that have been conducted at different points in time but within the same host review or map, when the researchers concerned decide to update their custom binary machine learning model to fully exploit further screening metadata that has accumulated since the preceding version of the model. In these cases, the sample size associated with any second replication from the same host review / map will be halved, along with the sample size from the first replication, prior to using the two estimates in the updated meta-analysis. Similar, submission of any third replication from the same host review / map will result in all three sample sizes being entered into the updated analysis as a third of their original size; a fourth replication would result in quartering the original sizes of the four samples; and so on.

Publication and co-authorship

Standard datasets and metadata from each replication of this SWAR, as well as datasets and statistical analysis code from the prospective meta-analysis of multiple replications, will be published when ready, in an open access repository on the Open Science Framework (OSF) platform (yet to be created). We also aim to publish a regularly updated, peer reviewed journal article focused on presenting continually updated results from the prospective meta-analysis component of this living SWAR in F1000 Research.

We invite review teams to consider conducting and submitting further replications of this SWAR for integration into the prospective meta-analysis, in order to help build a cumulative evidence base for the use of generative AI tools for eligibility screening for systematic reviews and maps of research.

Each contributor to a research team that replicates this SWAR and submits a complete dataset and metadata from their replication to the required standard and in the required format will thereby qualify to be added as: 1) a new contributor to the OSF repository; and 2) a new co-author of the next update of the F1000 Research article (that will feature an update of the prospective meta-analysis integrating data from the latest replications), provided they meet all four ICMJE criteria for co-authorship [4] with respect to their specific contribution. Initially, we plan to publish an updated version the F1000 article, incorporating updated meta-analytic results, every 6 months following publication of the inaugural version; however this publication frequency will be continually reviewed, and may change it contingent on how often further replications of the SWAR are submitted in practice.

Anyone considering replicating this SWAR should contact Ian Shemilt (i.shemilt@ucl.ac.uk) before starting to discuss the requirements and co-authorship policy in more detail, and to request copies of the SWAR dataset and reporting template (to be added to the OSF repository when this has been created).

Possible Problems in Implementing This SWAR

Continual updating of the prospective meta-analysis component will be reliant on the willingness and capacity of other researchers / review teams to replicate this SWAR in completed reviews, and to submit standardised datasets and metadata in the required formats in return for co-authorship of an article reporting the living SWAR (prospective meta-analysis).

References

1. Macaskill P, Takwoingi Y, Deeks JJ, Gatsonis C. Chapter 9: Understanding meta-analysis. In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (updated July 2023). Cochrane, 2023. Available from https://training.cochrane.org/handbook-diagnostic-test-accuracy/current (accessed on 24 November 2025).
2. White IR. Multivariate random-effects meta-analysis. The Stata Journal 2009; 9(1):40–56.
3. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. Statistics in Medicine 2012;31(29):3821-39.
4. https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html (accessed on 24 November 2025).

Publications or Presentations of This SWAR Design

Examples of The Implementation of This SWAR

People to show as the source of this idea: Ian Shemilt, James Thomas.
Contact email address: i.shemilt@ucl.ac.uk
Date of idea: 01/01/2025
Revisions made by: N/A
Date of revisions: