Screening Model

The Screening Model uses AI to learn from screening decisions within a specific nest, predicting inclusion (standard screening) or abstract advancement (two pass screening) probabilities based on configuration. Then it automatically re-orders studies in Screening so that the most likely to be included/advanced are presented first. This assists in identifying relevant studies early.

Robot Screener

The Screening Model can be used to power AI-assisted screening, replacing one expert in Dual Screening processes:

Video

When selecting a mode, note that in most cases, when employing Dual Two Pass Mode, **the Robot Screener should replace an expert reviewer only for the Abstract stage of screening** , as the model itself is trained on and screens based on Abstract content. Using the model in this way provides Advancement probabilities (in effect, relevancy scores) to each record.

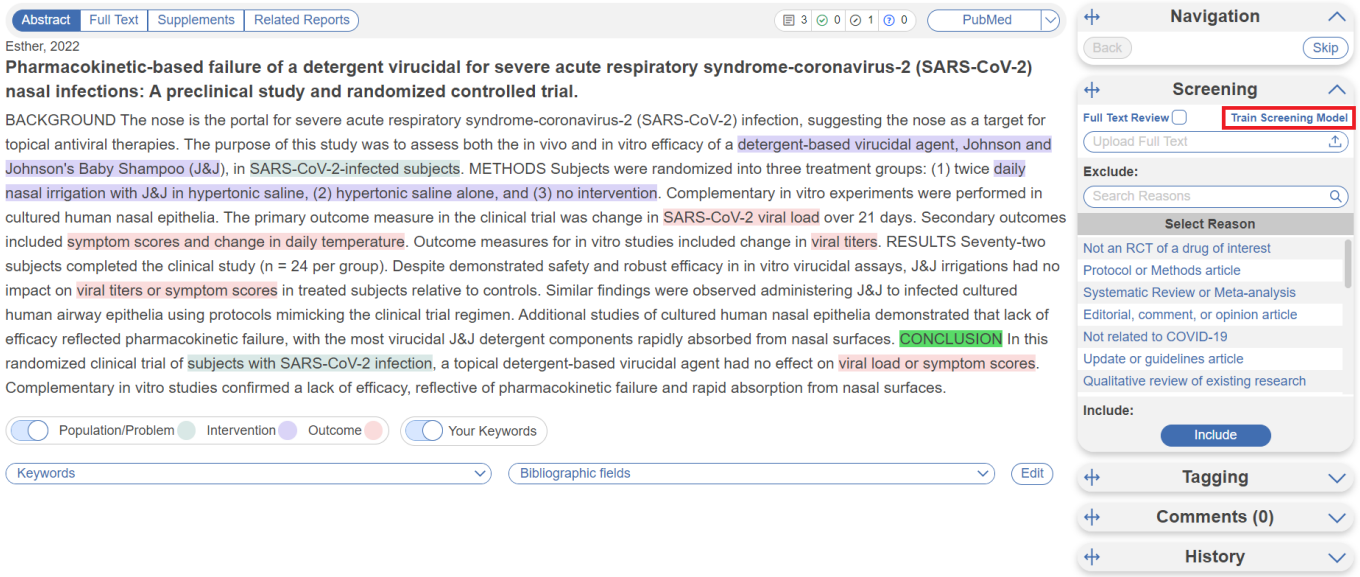
See here for full Guidance on Robot Screener in Dual Two Pass Mode, an AI alternative to a second reviewer in Dual Screening modes.

User Guide

Running the Screening Model

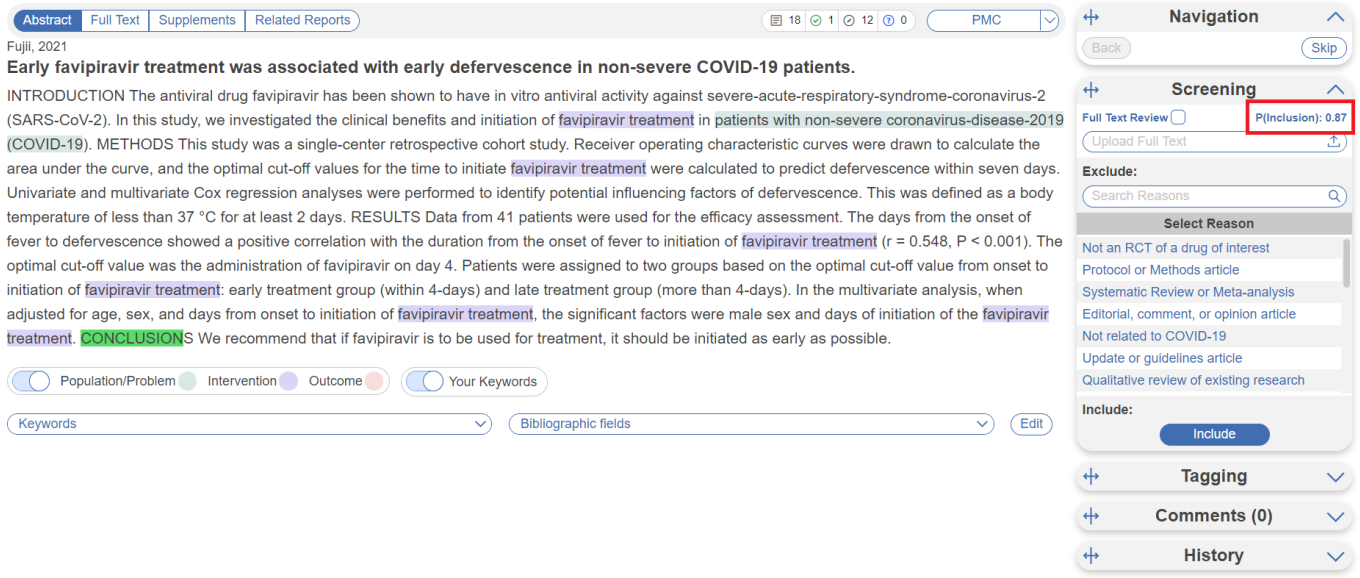
To learn about configuration settings, which enable you to toggle Manual updating vs. Automatic and Displayed vs. Hidden, see the Settings page.

In its default setting, the Screening Model must be run manually. To do so, click “Train Screening Model” on the Screening panel:



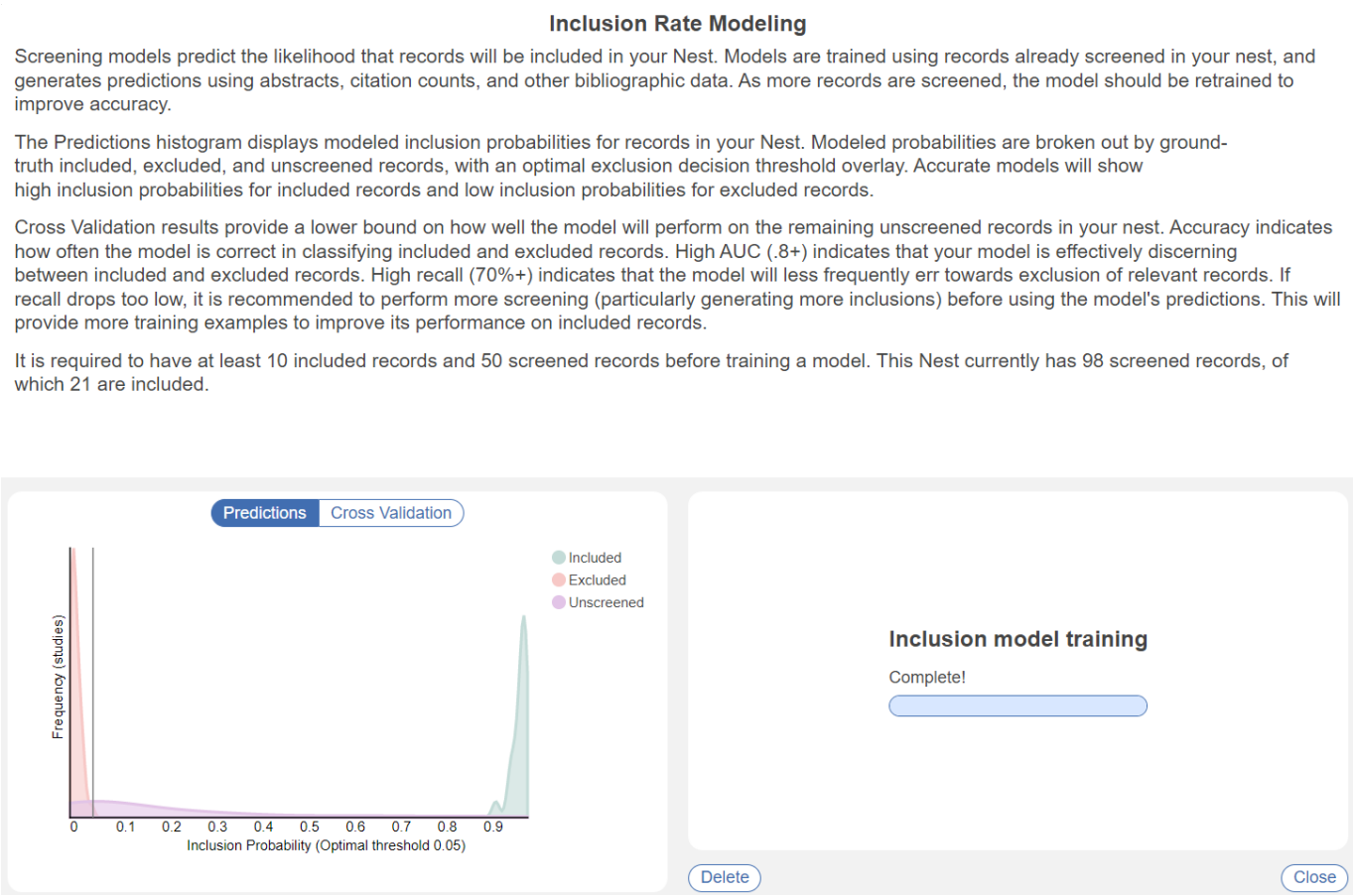
Once the modal opens, click “Train New Model.” Note: To provide the model with sufficient information to begin understanding your review, we require **50 total screens and 10 inclusions/advancements** before the model can be trained. If there is insufficient evidence to train the model, complete more screening until the “Train New Model” button becomes available.

It may take a minute to train, after which it will populate a histogram on the left. From then on, each record will show a probability of inclusion or advancement:



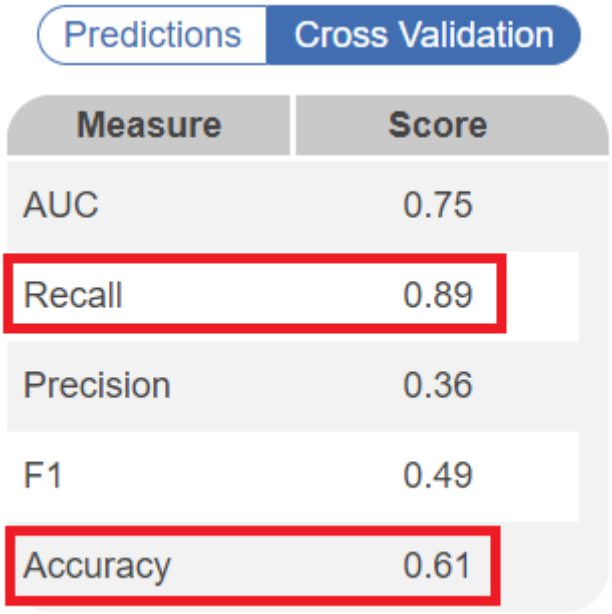
Interpreting the Model

Once the Model is trained, you should see a graph where Included or Advanced, Excluded, and Unscreened records are represented by green, red, and purple curves, respectively:



Odds of inclusion/advancement are presented on the x-axis (ranging from 0 to 1). Since the Model is trained on a nest-by-nest basis, its accuracy ranges based on how many records it can train on and how many patterns it can find in inclusion activities.

You can see the accuracy in the modal after the model is trained. In the Cross Validation tab, several statistics are shown. Scores of Recall and Accuracy can be used to interpret how the model will perform on the remaining records. High recall (0.7/70%+) indicates that the model will less frequently exclude relevant records, meaning higher performance. Similarly, accuracy indicates how correct the model's decisions are compared to already screened records, and thus how it is likely to fare on upcoming records. See below for an example of a relatively well trained model:



Implications for Screening

Inclusion Probability generated from the Screening model is also available as a filter in Inspector, which can assist with finding records based on their chance of inclusion/advancement. Bulk Actions can also be taken at your discretion, but ensure that you are careful in excluding studies if you have not reviewed their Abstracts at least!

Model Performance

Our Philosophy

Screening is a complex task that relies on human expertise. Our model may stumble due to:

- Insufficient training examples (usually included/advanced records) to learn from
- Data not available to the model (e.g. screening with a full text article, missing abstract)
- Weak signal amongst available predictors against protocol

For these reasons, we recommend using the model to augment your screening workflow, not fully automate it.

How can it augment your screening?

- Excluding clearly low-relevancy records
- Raising high relevancy records to reviewers

Our model errs towards including/advancing irrelevant records over excluding relevant records. In statistical terminology, the model aims to achieve high recall. In a review, it is far more costly to exclude a relevant study. Once excluded, reviewers are unlikely to reconsider a record. In contrast, an included/advanced study will be revisited multiple times later in the review, more readily

allowing an incorrect include/advance decision to be corrected.

Testing out the model

In an internal study, Nested Knowledge ran the model across several hundred SLR projects, finding the following cumulative accuracy statistics:

Standard Screening

- Area Under the [Receiver Operating Characteristic] Curve (AUC): 0.88
- Classification Accuracy: 0.92
- Recall: 0.76
- Precision: 0.40
- F1: 0.51

Two Pass Screening

In two pass screening, the model predicts advancement of a record from abstract screening to full text screening. Given that advancement rates are typically higher than inclusion rates, the model has more positive training examples, and demonstrates improved recall.

- AUC: 0.88
- Classification Accuracy: 0.93
- Recall: 0.81
- Precision: 0.44
- F1: 0.56

Following our philosophy, recall is relatively higher than precision: the model suggests inclusion/advancement of a larger amount of relevant records, at the cost of suggesting inclusion of some irrelevant records. Due to class imbalance, the model scores a 90%+ classification accuracy, predominantly consisting of correct exclusion suggestions.

For comparison purposes, our study found human reviewer recall (relative to the adjudicated decision) was 85% in the average nest. Our models are within 4 & 9 points of human performance on this most critical measure.

Analyzing Your Nest

When you train a new model, we generate k-fold cross validation performance measures using the same model hyperparameters the final model is trained with. These performance measures typically provide a lower bound on the performance you can expect from the model on records not yet screened in your nest. High recall (70%+) suggests that your review is less likely to be missing relevant records at the end of screening. High AUC (.8+) suggests that your model is effectively discerning between included and excluded records.

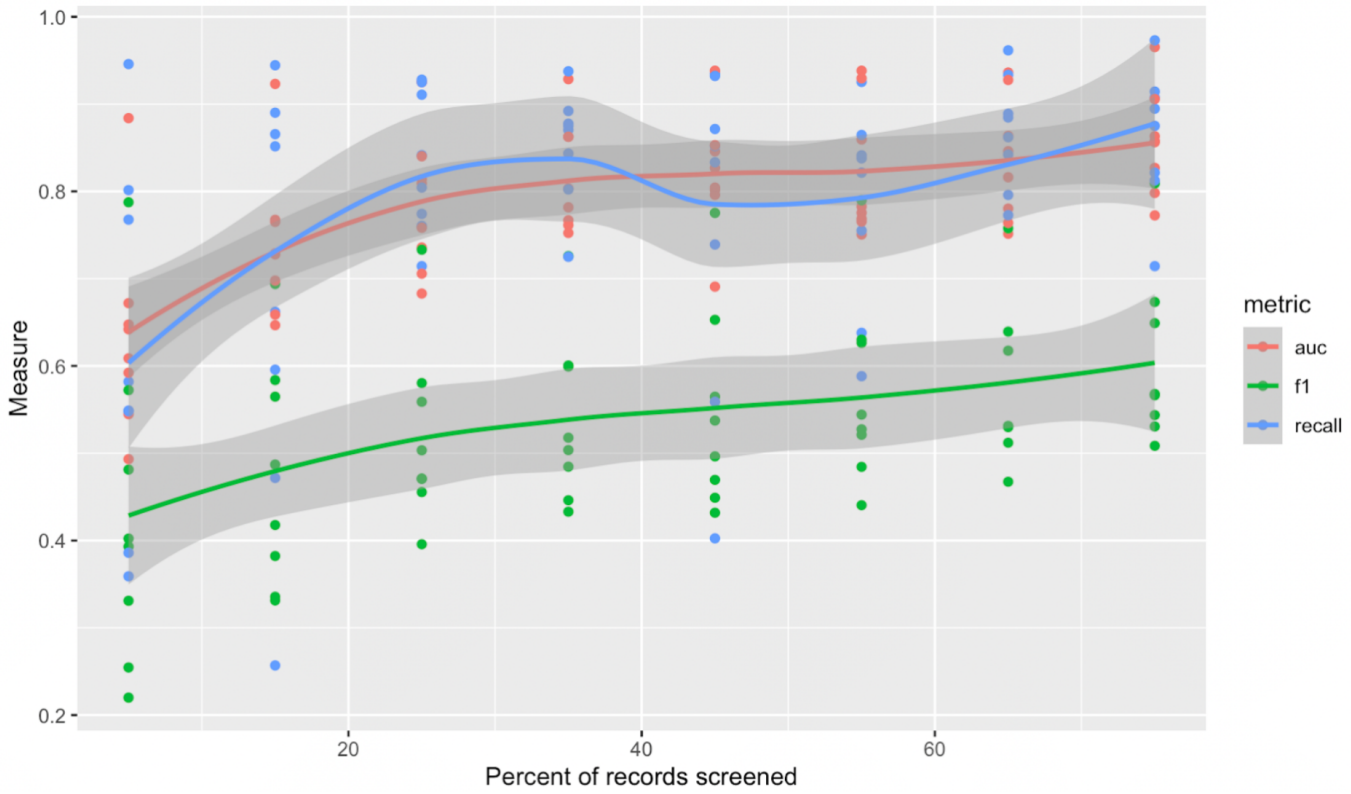
While we cannot guarantee performance improvement, below is some rough empirical data for how you might expect performance measures to improve as you screen more records in your nest.

Timing of Model Training

In general, as you screen more records, the better the model will perform. Of course, you want to use the model before you've screened every record!

To provide the model with sufficient information to begin understanding your review, we require 50 total screens and 10 inclusions/advancements. At that point, we recommend checking model performance (see above) to evaluate performance.

As the graph below shows, AUC and recall can grow on a relatively sharp curve early in your review. The curve begins to flatten around 20-30% of records screened, which is where we typically begin to recommend the use of Robot Screener in Dual screening modes.



How the Screening Model Works

At a high level, the model is a Decision Tree- a series of Yes/No questions about characteristics of records that lead to different probabilities of inclusion/advancement.

In more detail, the model is a gradient-boosted decision tree ensemble. Its hyperparameters, particularly around model complexity (number of trees, tree depth) are optimized using a cross validation grid search. The model produces posterior probabilities and is optimized on logistic loss. SMOTE oversampling is employed as a correction to highly imbalanced classes frequently seen in screening.

What data does the model use?

The model uses the following data from your records as inputs:

- Bibliographic data
  - Time since publication of the record
  - Page count
  - Keywords/Descriptors
- Abstract Content
  - N-grams
  - OpenAI text embedding (ada-002)
- Citation Counts from Scite, accessed using the DOI
  - Number of citing publications
  - Number of supporting citation statements
  - Number of contrasting citation statements

Often some of this data will be missing for records; it is imputed as if the record is approximately typical to other records in the nest.