Screening Model

How the Screening Model Works

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 are presented first. This assists in identifying relevant studies early.

It also works to power Robot Screener, an AI alternative to a second reviewer in Dual Screening modes.

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 the histogram on the left. From then on, each record will show a probability of inclusion (this is always shown if automatic training is on):

| Abstract | Full Text | Supplements | Related Reports |

☰ 18 ⊘ 1 ⊘ 12 ⊙ 0   (PMC ▾)

Fujii, 2021

Early favipiravir treatment was associated with early defervescence in non-severe COVID-19 patients.

INTRODUCTION The antiviral drug favipiravir has been shown to have in vitro antiviral activity against severe-acute-respiratory-syndrome-coronavirus-2 (SARS-CoV-2). In this study, we investigated the clinical benefits and initiation of favipiravir treatment in patients with non-severe coronavirus-disease-2019 (COVID-19). METHODS This study was a single-center retrospective cohort study. Receiver operating characteristic curves were drawn to calculate the area under the curve, and the optimal cut-off values for the time to initiate favipiravir treatment were calculated to predict defervescence within seven days. Univariate and multivariate Cox regression analyses were performed to identify potential influencing factors of defervescence. This was defined as a body temperature of less than 37 °C for at least 2 days. RESULTS Data from 41 patients were used for the efficacy assessment. The days from the onset of fever to defervescence showed a positive correlation with the duration from the onset of fever to initiation of favipiravir treatment (r = 0.548, P < 0.001). The optimal cut-off value was the administration of favipiravir on day 4. Patients were assigned to two groups based on the optimal cut-off value from onset to initiation of favipiravir treatment: early treatment group (within 4-days) and late treatment group (more than 4-days). In the multivariate analysis, when adjusted for age, sex, and days from onset to initiation of favipiravir treatment, the significant factors were male sex and days of initiation of the favipiravir treatment. CONCLUSIONS We recommend that if favipiravir is to be used for treatment, it should be initiated as early as possible.

◯ Population/Problem ◯ Intervention ◯ Outcome ◯   ◯ Your Keywords

(Keywords ▾) (Bibliographic fields ▾) (Edit)

Navigation ⌃

(Back) (Skip)

Screening ⌃

Full Text Review ☐   **P(Inclusion): 0.87**

(Upload Full Text ⬆)

Exclude:

(Search Reasons 🔍)

Select Reason

Not an RCT of a drug of interest
Protocol or Methods article
Systematic Review or Meta-analysis
Editorial, comment, or opinion article
Not related to COVID-19
Update or guidelines article
Qualitative review of existing research

Include:

(**Include**)

Tagging ⌄

Comments (0) ⌄

History ⌄

Interpreting the Model

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

Inclusion Rate Modeling

Screening models predict the likelihood that records will be included in your Nest. Models are trained using records already screened in your nest, and generates predictions using abstracts, citation counts, and other bibliographic data. As more records are screened, the model should be retrained to improve accuracy.

The Predictions histogram displays modeled inclusion probabilities for records in your Nest. Modeled probabilities are broken out by ground-truth included, excluded, and unscreened records, with an optimal exclusion decision threshold overlay. Accurate models will show high inclusion probabilities for included records and low inclusion probabilities for excluded records.

Cross Validation results provide a lower bound on how well the model will perform on the remaining unscreened records in your nest. Accuracy indicates how often the model is correct in classifying included and excluded records. High AUC (.8+) indicates that your model is effectively discerning between included and excluded records. High recall (70%+) indicates that the model will less frequently err towards exclusion of relevant records. If recall drops too low, it is recommended to perform more screening (particularly generating more inclusions) before using the model's predictions. This will provide more training examples to improve its performance on included records.

It is required to have at least 10 included records and 50 screened records before training a model. This Nest currently has 98 screened records, of which 21 are included.

Odds of inclusion 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 (see red arrow in the image above). Accuracy is presented as a Receiver Operating Characteristic Area Under the Curve (ROC-AUC).

ROC-AUC has a minimum of 0 and a maximum of 1, where 1 indicates that when the Model checks its predictions on existing inclusion decisions, it had no false positives or negatives. So, high ROC-AUC (0.85-0.99) indicates that trusting the Model may be warranted, while lower ROC-AUC means that more screening may be necessary to train it further, or the patterns in inclusion decisions are too disparate for accurate prediction.

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. 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!

Behind the Scenes: How Screening Modelling Works

What exactly is the model?

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.

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

Standard Screening

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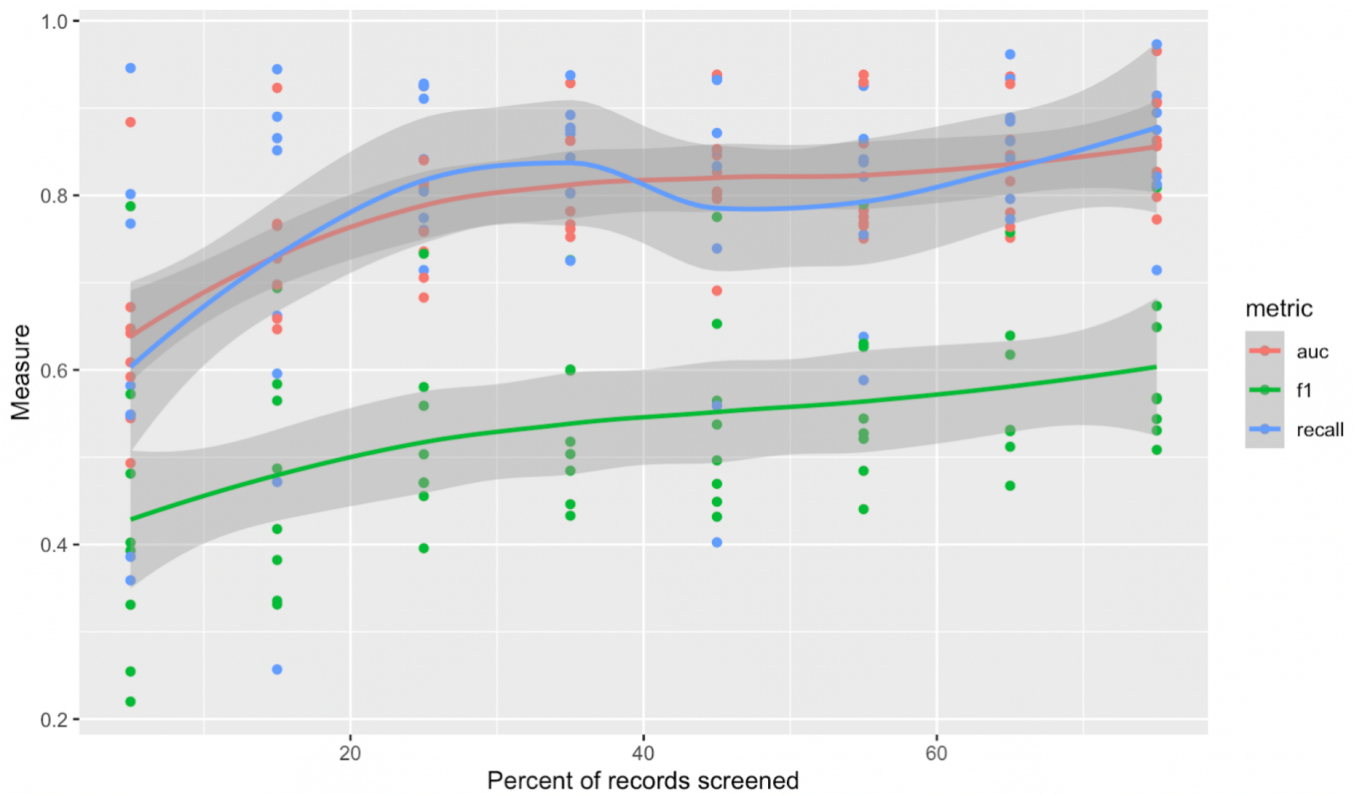
While we cannot guarantee performance improvement, below is some rough empirical data for how you might expect performance measures to improve as you screen more records in your nest.

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.