Probabilistic Performance Bounds for Evaluating Depression Models Given Noisy Self-Report Labels

Robert Różański, Elizabeth Shriberg, Yang Lu, Amir Harati, Tomasz Rutowski, Piotr Chlebek, Tulio Goulart, Ricardo Oliveira

Robert Różański

2024.03.18

Depression detection models

- Modalities
  - Acoustic
    - Acoustic characteristics (feature engineering or DNN)
    - Conversations, reading, vocalization
  - Text
    - Social media posts
    - From speech through ASR
  - Video
    - Body language (more prevalent in emotion recognition)
- Targets
  - Psychological assessment (proper diagnosis or questionnaires)
  - Self-assessment questionnaires (PHQ, BDI, IDS-RS, MADRS)
  - Manual labelling (content-based)
  - Surrogate / Proxy labels (keywords / valency / etc)

Depression detection models

- Modalities
    - **Acoustic**
        - Acoustic characteristics (feature engineering or DNN)
        - Conversations, reading, vocalization
    - **Text**
        - Social media posts
        - From speech through ASR
    - Video
        - Body language (more prevalent in emotion recognition)
- Targets
    - Psychological assessment (proper diagnosis or questionnaires)
    - **Self-assessment questionnaires** (**PHQ**, BDI, IDS-RS, MADRS)
    - Manual labelling (content-based)
    - Surrogate / Proxy labels (keywords / valency / etc)

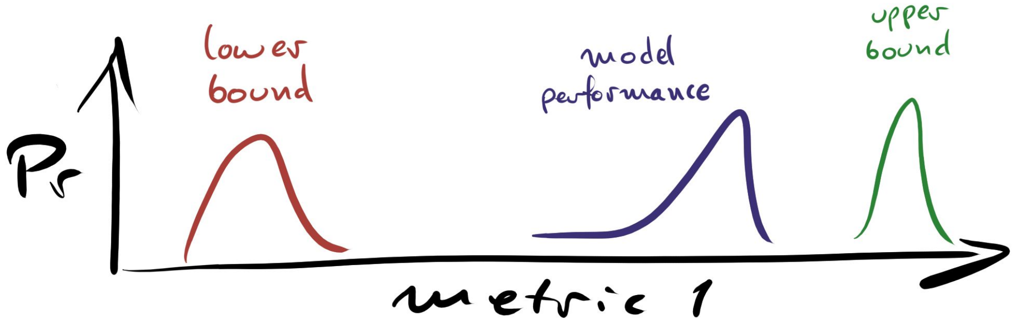
Model evaluation issues in depression detection

- Small datasets
- No specified application / non-representative test sets
  - Artificial data elicitation like reading a text passage
- Reporting of only a few performance metrics
- Point estimates (no confidence intervals)
  - No accounting for a test set size
  - No accounting for label reliability / noise
  - No accounting for label distribution (another factor limiting performance)

Model evaluation issues in depression detection

- Small datasets (somewhat)
- No specified application / non-representative test sets
    - Artificial data elicitation like reading a text passage
- **Reporting of only a few performance metrics**
- Point estimates (no confidence intervals)
    - **No accounting for a test set size**
    - **No accounting for label reliability / noise**
    - **No accounting for label distribution** (another factor limiting performance)

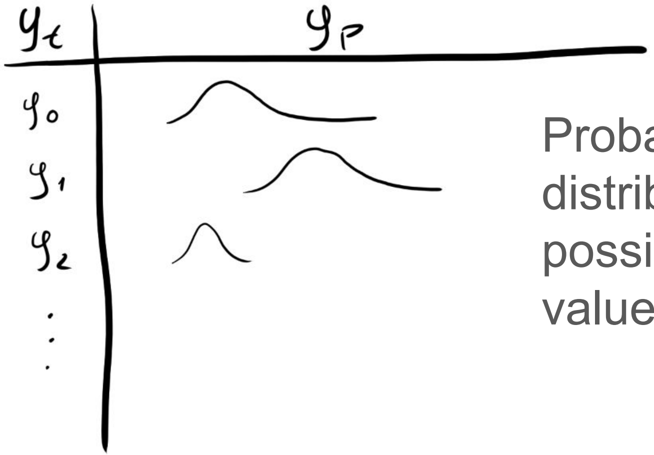
What would be nice to have



- **Lower performance bound**: predictions generated by random draws from the same distribution as the labels (test set size; label distribution)
- **Model performance**: bootstrapping model predictions (test set size; label distribution; hard/easy cases)
- **Upper performance bound**: ?? what's the "maximum" possible performance (must be compatible with variety of performance metrics)

Bayes error

Target labels



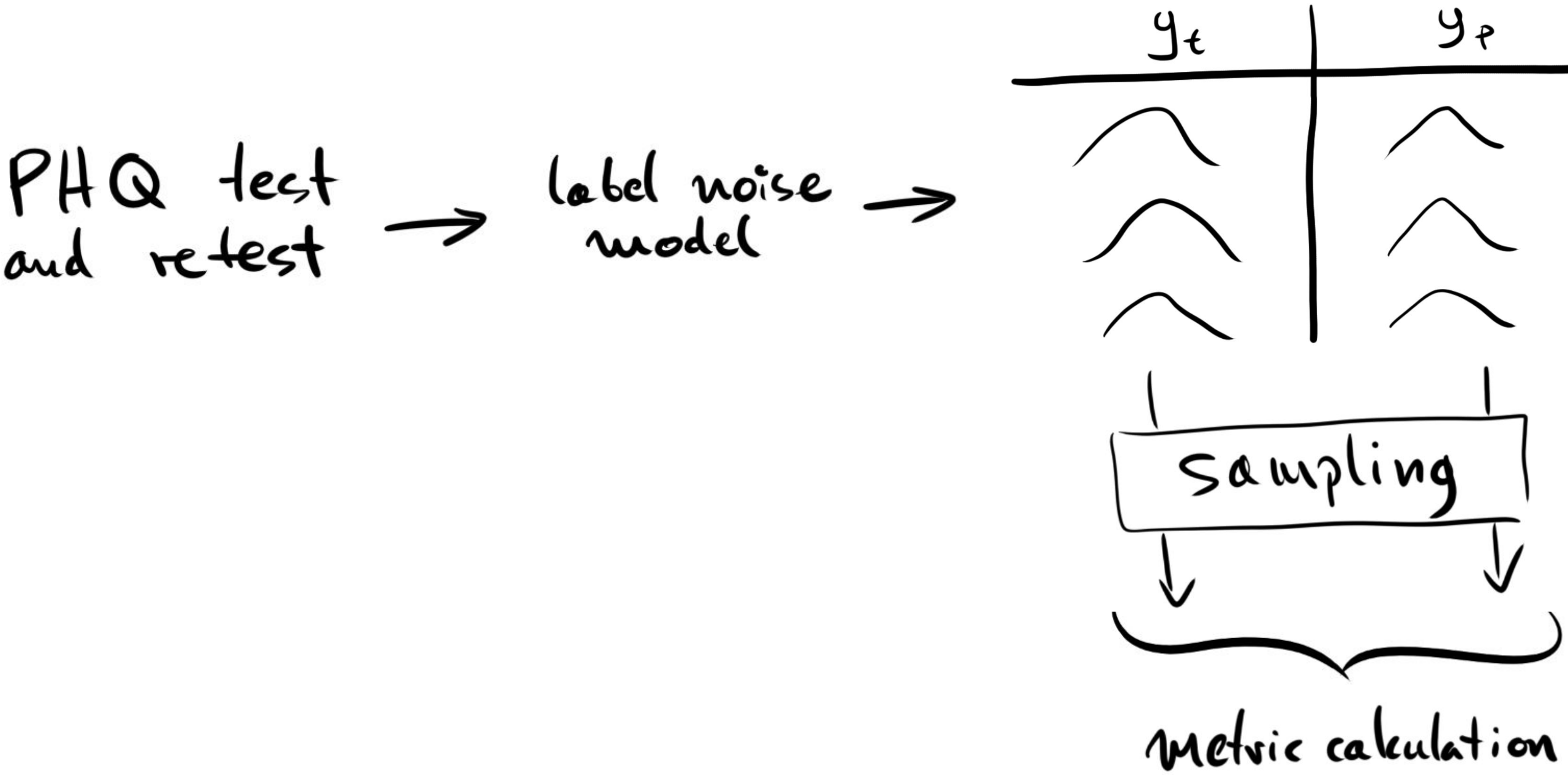
Probability distributions over possible label values

Error rate calc.

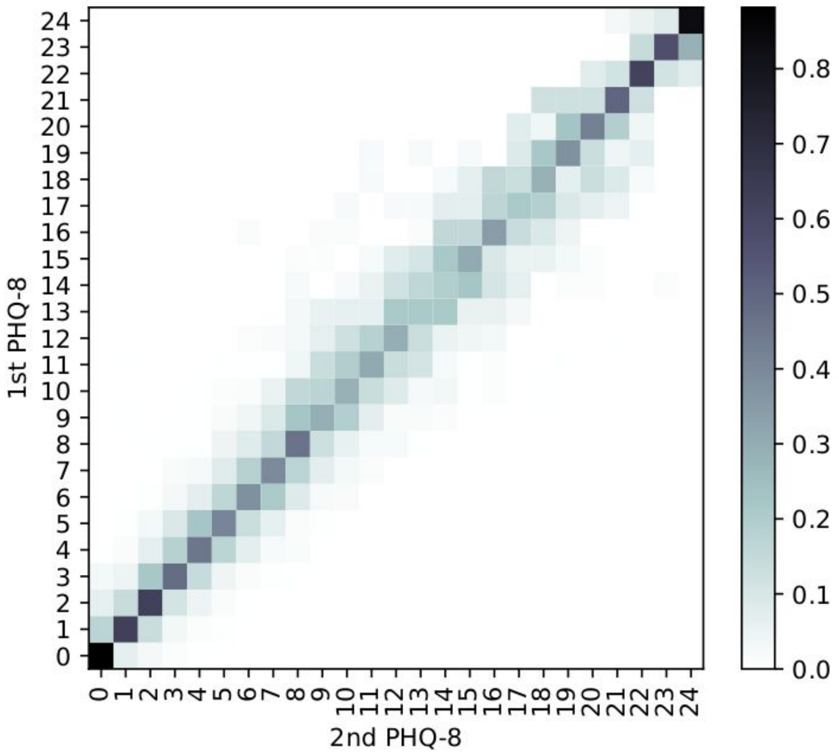
How to estimate Bayes error?

- Use a BE estimator
  - no estimators for regression; only error rate; dimensionality
- Multiple expert annotators
  - incompatible with self-assessment

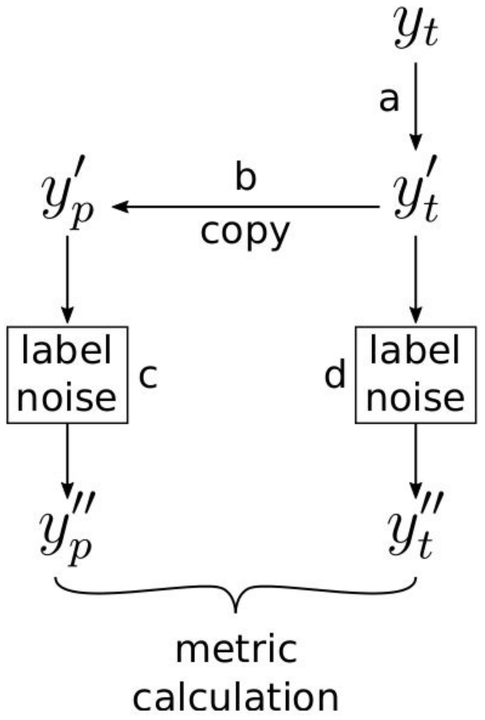
Our approach



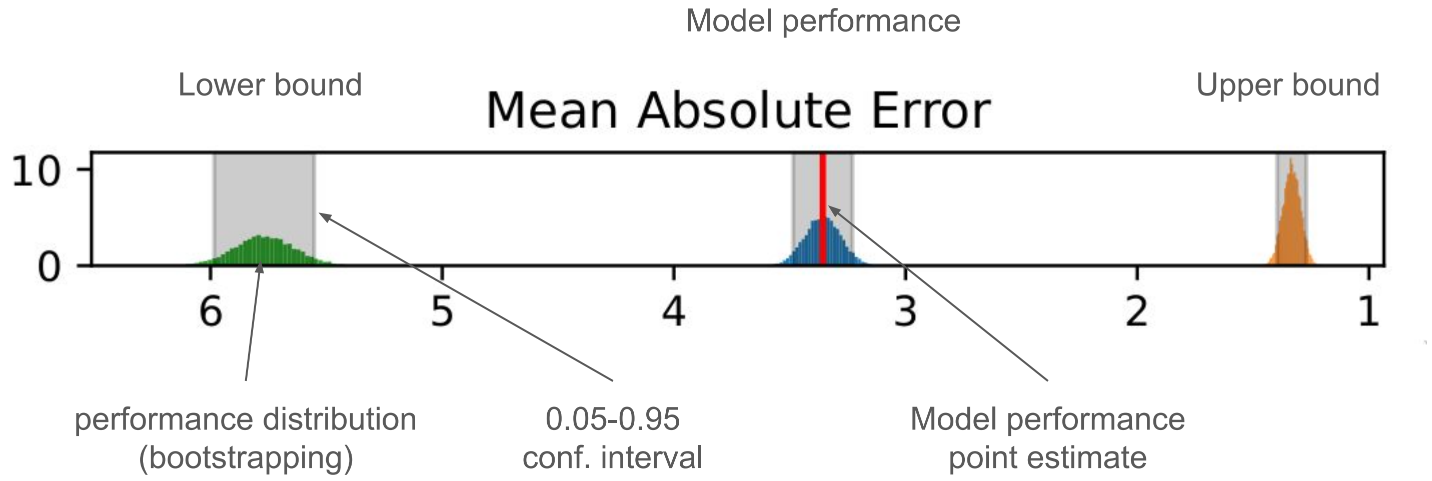
Our approach



sample (cond.) and replace each value independently

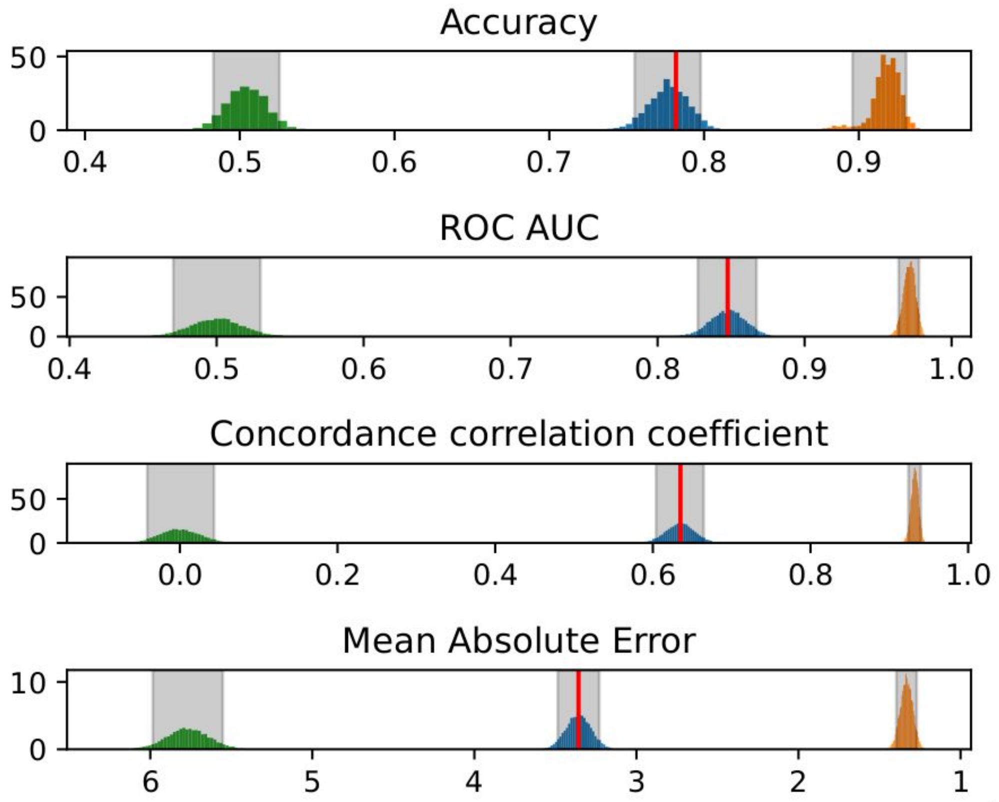


Results visualization



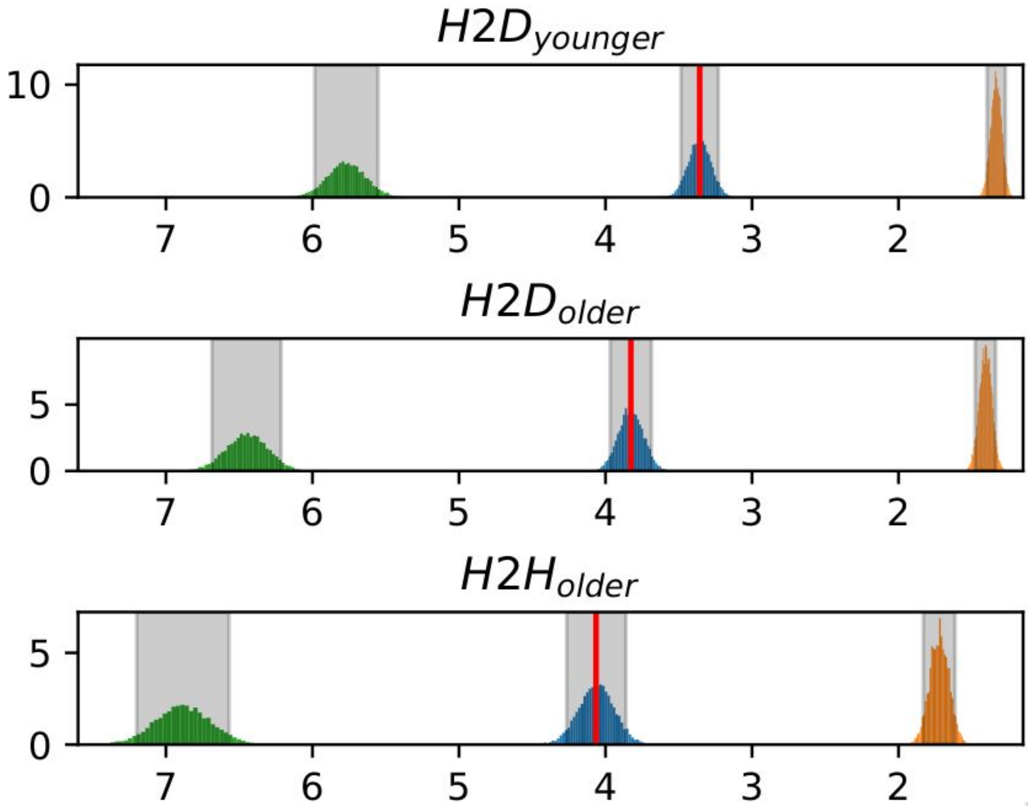
- Width of all distributions affected by:
  - Test set size (bigger datasets = lower variance)
  - Label distribution
- Upper bound affected by the noise model
- Other sources of irreducible errors omitted

Results



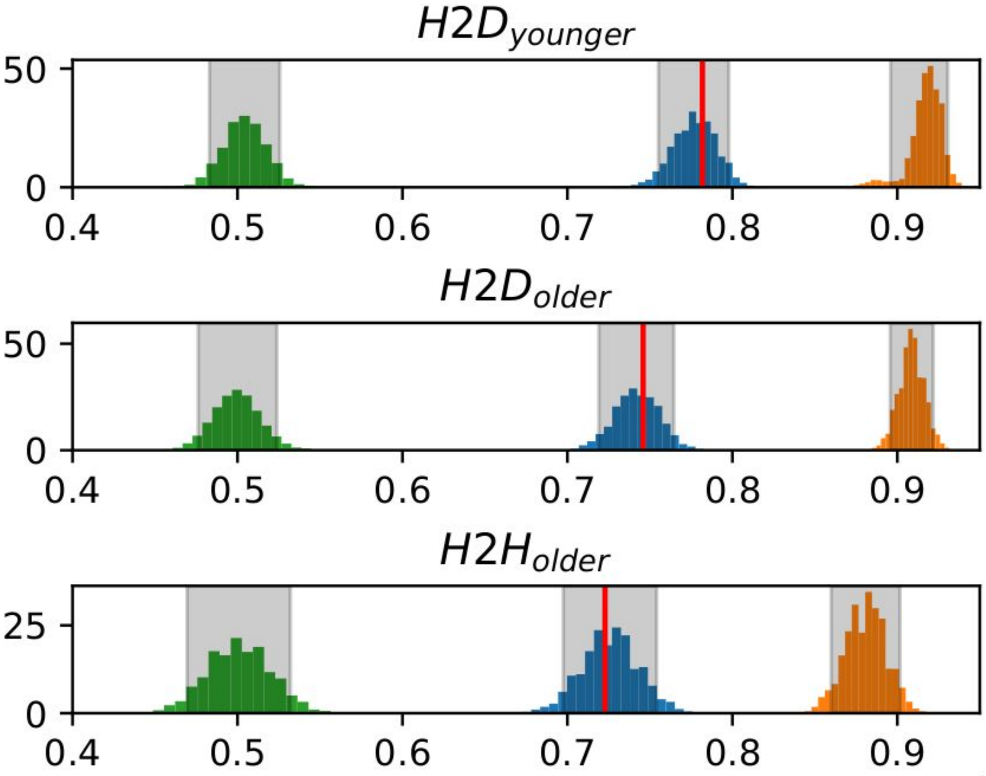
- One dataset, one model
- Metrics differ in how easy they are to satisfy (MAE vs the rest)
- Lower bound particularly useful for certain metrics (MAE, RMSE, etc.)

Results (MAE)



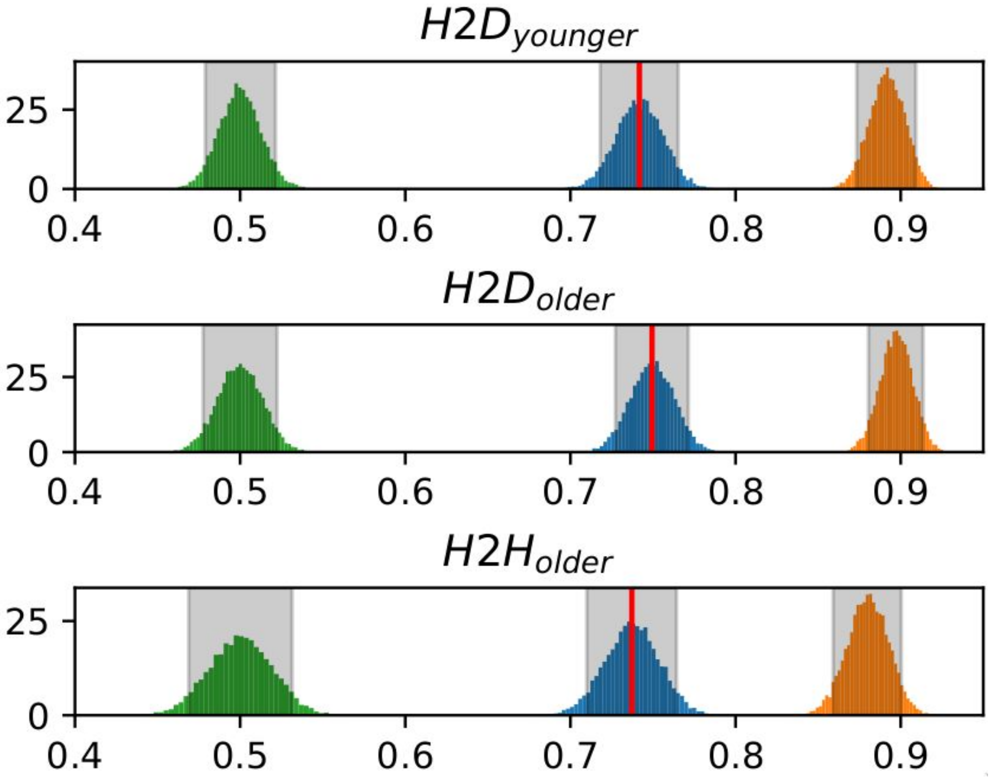
- Multiple datasets and models
- Substantial differences in locations of performance bounds
- Picture more complex than looking only at point estimates of model performance

Results (Accuracy)



- Multiple datasets and models
- What is the difference in performance between the first and third model?

Results (UAR)



- Multiple datasets and models
- Some metrics give more stable results across bounds and model performances
- Different metrics give different rankings

Summary

- We developed a method to estimate upper performance bound
- It takes into account:
  - Label noise (needs a model)
  - Test set size
  - PHQ distribution in the test set
- We show how to use it in concert with lower performance bounds, while using bootstrapping to evaluate model performance