ellipsis HEALTH

Low Versus High Label Noise for Evaluation of Depression Models Using Probabilistic Performance Bounds

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

IEEE SPMB
December 2024

Depression Symptoms

low mood

changes in appetite

sleep disturbance

agitation

fatigue

difficulty concentrating

Introduction

- Depression is a prevalent mental health condition.
  a. Major depression is one of the most common mental illnesses, affecting more than 8% (21 million) of American adults each year. 15% of youth (3.7 million) ages 12-17 are affected by major depression.
  b. In 2023, 29.0% of Americans report having been diagnosed with depression in their lifetime, while 17.8% reporting currently having depression.
- Speech technology is a promising tool for scalable screening.
- In this field, currently no large common labeled test sets available
- Many papers on depression detection from speech/language but
  a. Very small data sets
  b. Reported for specific contexts (different languages, settings, use cases)
  c. Poor labels
  d. No “blind” test / overfitting on test data (same data was used)
  e. Reports on regression or multi-class classification results are rare

Background & Objectives

- Speech biomarkers link vocal characteristics with mental health conditions.
- Machine learning models rely on noisy self-reported labels (PHQ-8) even in best case scenarios.
- Noise arises from human factors and methodological variations.
- Noise in labels affects model evaluation and fairness.
- Study explores low and high noise contexts.
- Objectives:
    a. Develop Noise Models:
        i. Focus on probabilistic performance bounds.
        ii. Simulate real-world scenarios with varying noise levels.
    b. Evaluate the impact of label noise on model performance.
    c. Evaluate Models Across Demographics:
        i. Analyze fairness in performance across age, gender, etc.

Ellipsis Vocal Biomarkers – Overview

Use Case

Digital Front Door/Portal

Telehealth

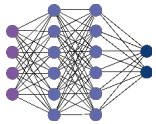
In-Clinic

Care Coordination / Case MGMT Calls

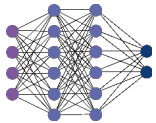
Natural Voice

Deep Learning

Acoustic Model



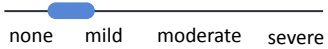
NLP Model



PHQ/GAD Equivalent Scores

Depression Score | **8**

Anxiety Score | **4**



Referral

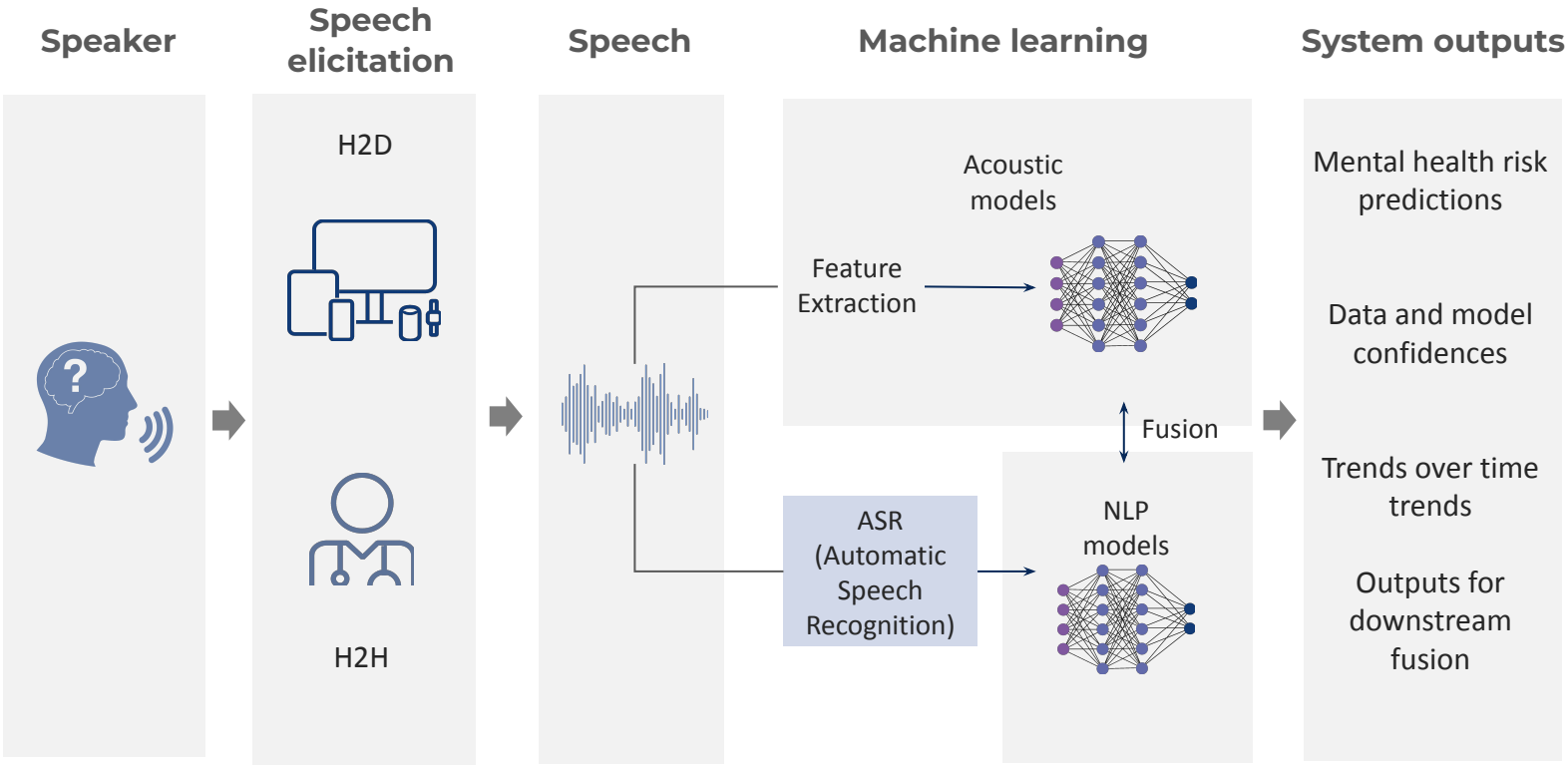
PCP

Specialist

Care Manager

Digital Intervention

Ellipsis Vocal Biomarkers – Overview



Why Use Both Acoustic And Transcript-Based Models?

Prosodic (rate, pitch, energy)

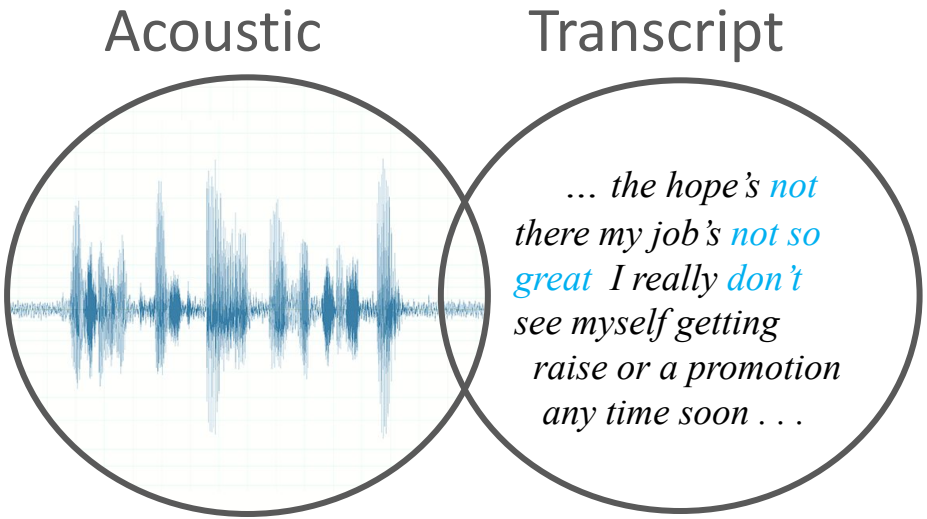
Glottal features

Formant-based

Spectral

Cepstral

. . .



Negative sentiment

Fewer positive words

Self-focus

First person pronouns

Word pattern dictionaries

Simplified syntax

. . .

Fusion of output of NLP & acoustic models

- Fusion of acoustic and language models for best performance
- Modern deep learning architectures (currently using transformers)
- Direct modeling from acoustic signal; outperforms precomputed features
- Large, diverse training (labeled for MH) and pre-training (unlabeled)
- Multiple uses of LLMs
- Topic modeling for both models (region weighting) and analytics use
- Multilingual methods for both acoustic and language models
- Method
  - Models trained & optimized individually (mostly for CCC)
  - Combined using weighted mean (found by optimizing CCC or MAE)
  - “Model” in figures reflects a weighted combination of the output of NLP and acoustic models
  - Model fusion weights are then applied during inference
- We found stable performance gain from fusion of NLP and acoustic model outputs

Metrics

Metric	+/-	C/R	Description
CCC	+	R	Concordance correlation coefficient; measures agreement between pairs of data (degree to which they fall on 45 degree line thru 0,0). Unlike a correlation, CCC penalizes for deviations from the exact value.
MAE MAE (fusion optimized for)	-	R	Mean Absolute Error. The 1st value is optimized for CCC. The 2nd is optimized for MAE.
MAE natural dist. 01 MAE (fusion optimized for)	-	R	Like above, but evaluated on data resampled to conform to a natural PHQ8 distribution.
ROC-AUC: Mean over 4 cutoffs Cutoff 5 Cutoff 10 Cutoff 15 Cutoff 20	+	C	Area under the Receiver Operating Characteristic curve; measures the separation power of a classifier. Different cutoffs are used to turn a regression problem into a classification problem.
Sensitivity/Specificity	+	C	True Positive Rate/True Negative Rate. At the point of Equal Error (FNR=FPR), they are equal as well.
Spearman’s Correlation	+	R	Measures monotonic correlation between pairs of data.
Pearson's Correlation	+	R	Measures linear correlation between pairs of data.

Metrics

Metric	+/-	C/R	Description
MSE	-	R	Mean Squared Error.
RMSE	-	R	Root Mean Squared Error.
Accuracy	+	C	True predictions divided by all predictions.
Precision	+	C	True predictions divided by the sum of True Positives and False Positives.
F1	+	C	Accuracy of classification; harmonic mean of precision and recall.
Log loss	-	C	Cross-entropy loss.
PPV	+	C	Ability to correctly label people who test positive.
NPV	+	C	Ability to correctly label people who test negative.
Pos. Likelihood	+	C	(Prob. of positive sample testing positive) / (Prob.of negative sample testing positive).
Neg. Likelihood	+	C	(Prob. of positive sample testing negative) / (Prob. of negative sample testing negative).
Cohen’s Kappa	+	C/R	Measures agreement between pairs of data, weighted to penalize the seriousness of disagreement.
Kendall’s Tau	+	C/R	Measures ordinal association between pairs of data.

Model generalization over population diversity (examples)

No model tuning or retraining

Metadata	Categories	Train set session count	Test set session count	Depression rate	Mean PHQ	Acoustic model AUC	NLP model AUC
Base performance over all test set		11 215	3080	25.7%	5.93	0.779	0.825
Gender	Male:	3125	1244	20.4%	5.74	0.769	0.819
	Female	4419	1790	35.3%	6.77	0.774	0.820
Age group	18-25	2087	847	30.0%	7.32	0.792	0.828
	26-35	3256	1382	24.8%	6.40	0.752*	0.820
	36-45	1444	513	18.7%	5.60	0.790	0.808
	46-65	766	283	34.6%	4.78	0.792	0.819
Smoking	Non-smoker	3850	813	23.2%	6.44	0.803	0.836
	Smoker	1807	397	31.3%	7.47	0.767	0.808
US States (selected)	California	924	266	26.8%	6.68	0.741	0.830
	Florida	831	253	26.2%	6.41	0.842*	0.875*
	Texas	723	232	26.0%	6.66	0.810	0.845
	New York	596	142	25.7%	6.70	0.815	0.887*
Ethnicity	Caucasian	5219	2039	24.7%	6.05	0.796	0.826
	African American	569	241	19.7%	5.63	0.777	0.812
	Hispanic	552	248	25.0%	6.73	0.676*	0.788
	Asian American	452	185	20.0%	5.61	0.789	0.841
	Mixed	364	173	31.3%	7.22	0.768	0.827
Marital	Never married	1850	188	31.5%	7.84	0.778	0.857

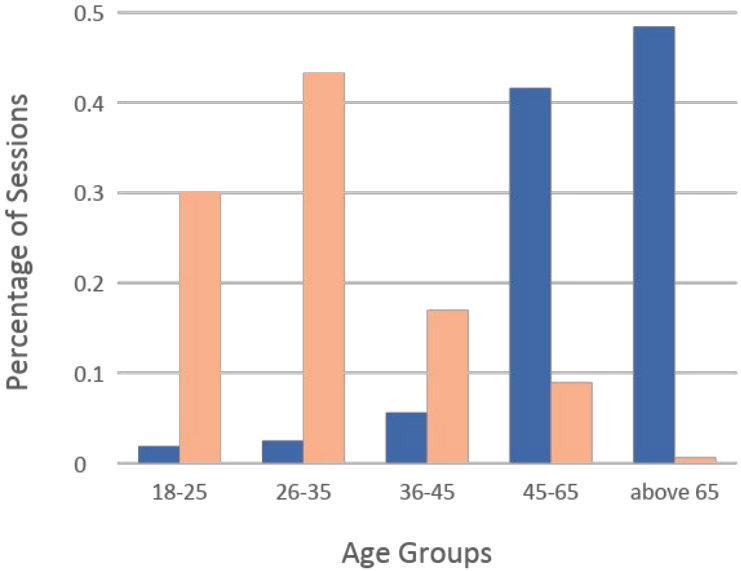
DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristics Curves: A Nonparametric Approach. Biometrics, 44, 837--845.

* Marginally significant at p<.05 in DeLong test for AUC

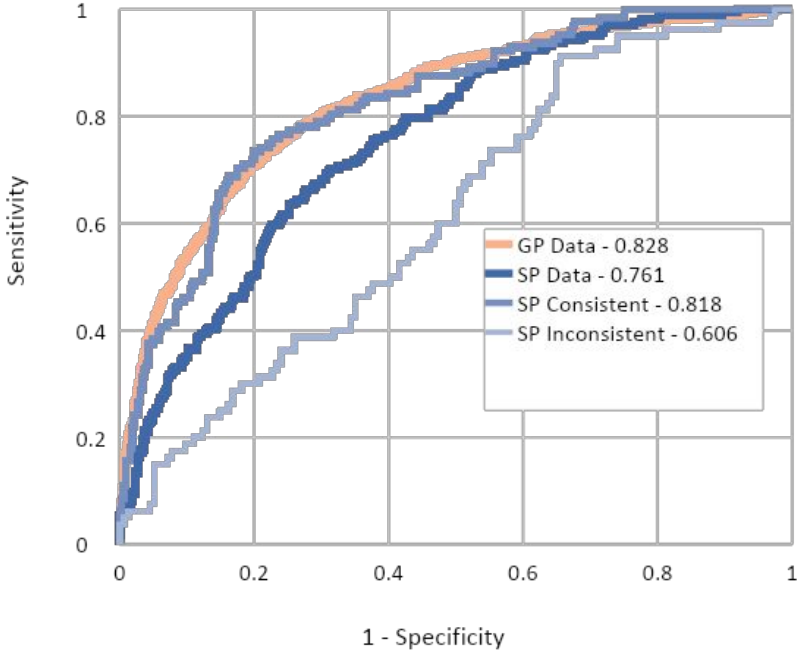
Model generalization over corpora, age

GP = General population, mean age 30, 50 states, ~10k speakers

SP = Senior population, mean age 60, 1 state, 687 speakers



NLP Model Performance. Train on GP corpus.
Test on GP vs SP corpus.
No model retraining.



Performance variation by training and test sizes, and demographic match

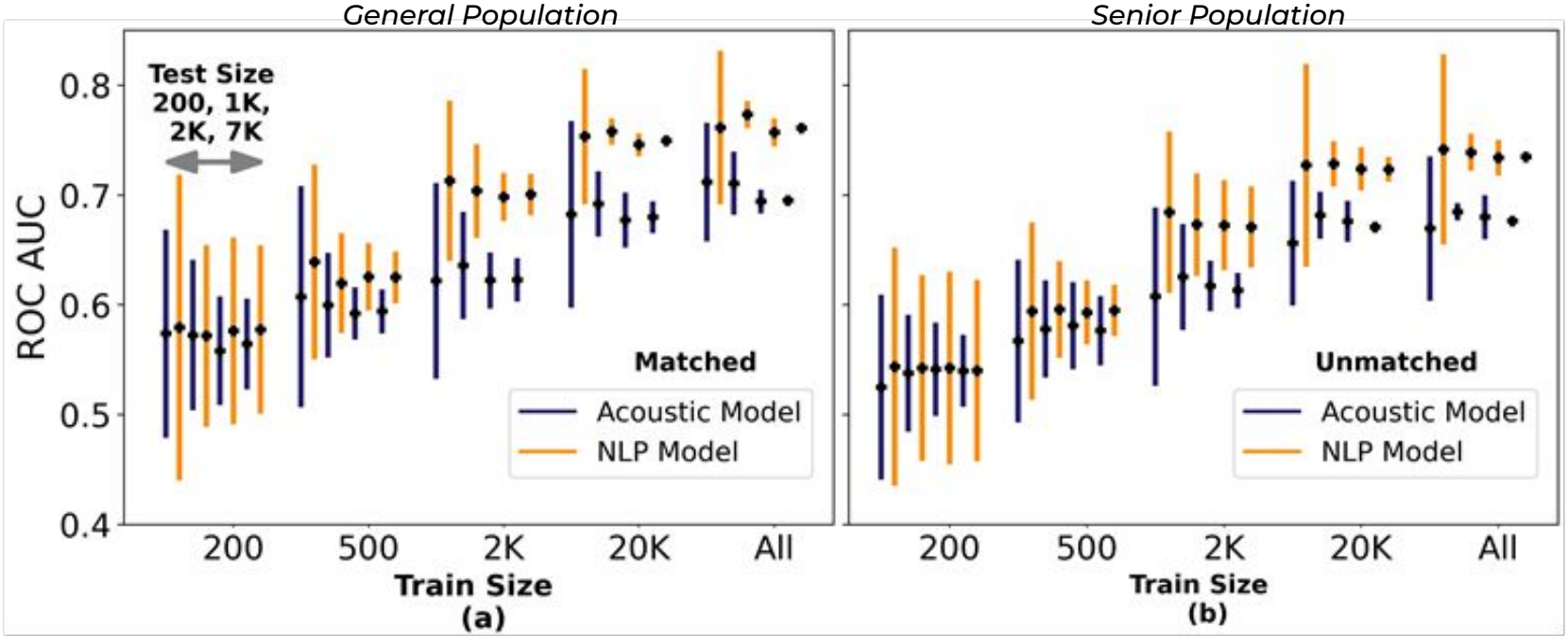


Figure 1- *(a): Variation in AUC by condition. Each vertical bar displays the mean and two standard deviations above and below the mean for a given test size. Vertical bars for test sizes of 500 and 5K were similar to those for 200 and 7K respectively and are removed for lack of space. 1- (b): Same as (a) but testing on SP data (train/test mismatch). All experiments use a fixed large development set; therefore, bar sizes are underestimates of true variation.*

Mental health diagnostic questionnaires

- PHQ (Patient Health Questionnaire)
- GAD (Generalized Anxiety Disorder) questionnaire
- Self-assessment (filled in by the patient)
- Different length variants
- PHQ-8
  - Questions about symptoms within last two weeks
  - Points per question depend on frequency (“not at all” – “nearly every day”)
  - 8 questions about:
    - Emotional state
    - Activity, appetite and sleep pattern changes
    - Interests and pleasure in doing things
    - Concentration

Dataset Sources

H2D dual (Low-Noise Dataset):

- **Purpose**: Used to develop the low-noise label models.
- **Collection Method**:
  - Participants interacted with an application to provide voice samples.
  - Completed the PHQ-8 survey twice in a single session to measure test-retest reliability.
- **Key Features**:
  - Samples: 5,625 sessions.
  - Speech Duration: 191 seconds on average (±96 seconds).
  - Demographics: Mean age = 36.3 years, Gender split = 51% female, 49% male.
  - High-quality data collected under controlled conditions, ensuring minimal noise in PHQ-8 labels.

Dataset Sources

H2D younger (Human-to-Device, Younger Demographic):

- **Purpose**: A test set representing younger participants for evaluating model performance.
- **Collection Method**:
  - Speech samples were elicited via app-based text prompts.
  - Topics covered various personal life areas to elicit conversational responses.
- **Key Features:**
  - Samples: 1,415 recordings from unique participants.
  - Speech Duration: 352 seconds on average (±155 seconds).
  - Demographics: Participants categorized into age ranges:
    - 18–25 years: 411 participants.
    - 26–35 years: 624 participants.
    - 36–45 years: 246 participants.
    - 46–65 years: 134 participants.
  - Represents the younger population with high-quality speech data for depression modeling.

Dataset Sources

H2D older (Human-to-Device, Older Demographic):

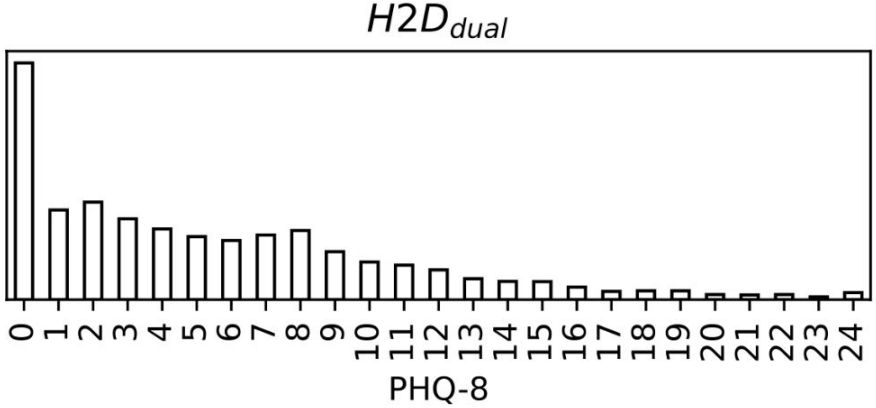
- **Purpose**: A test set representing an older demographic for model evaluation.
- **Collection Method**:
  - Similar app-based collection as H2D younger but tailored for an older audience.
  - Participants were recruited from a Southern California retirement community.
- **Key Features:**
  - **Samples**: 1,342 recordings from 270 unique participants.
  - **Speech Duration**: 239 seconds on average (±120 seconds).
  - **Demographics**:
    - Average participant age: 60.6 years.
    - Gender split: 60% female, 39% male.
  - Focus on the older population with insights into age-related differences in speech-based modeling.

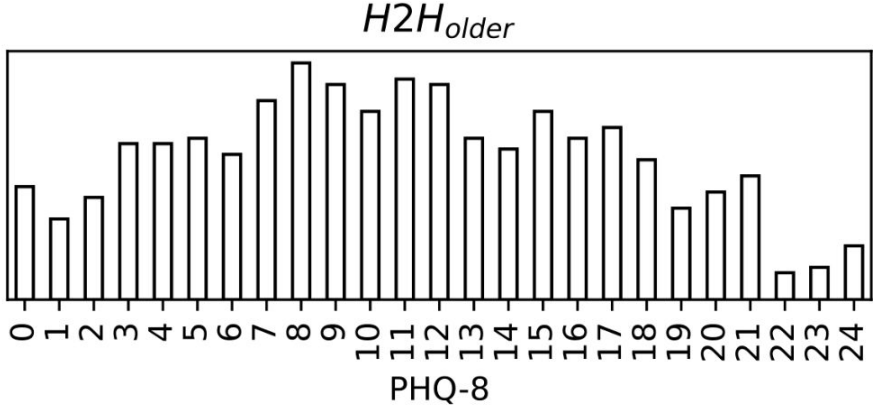
Dataset Sources

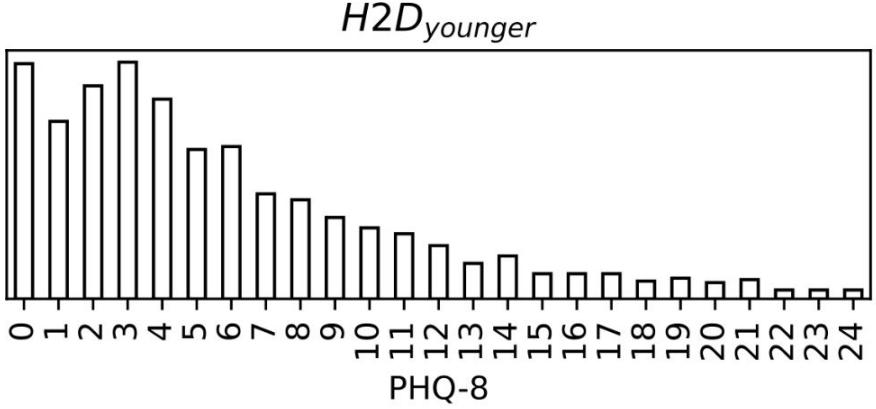
H2H older (Human-to-Human, Older Demographic):

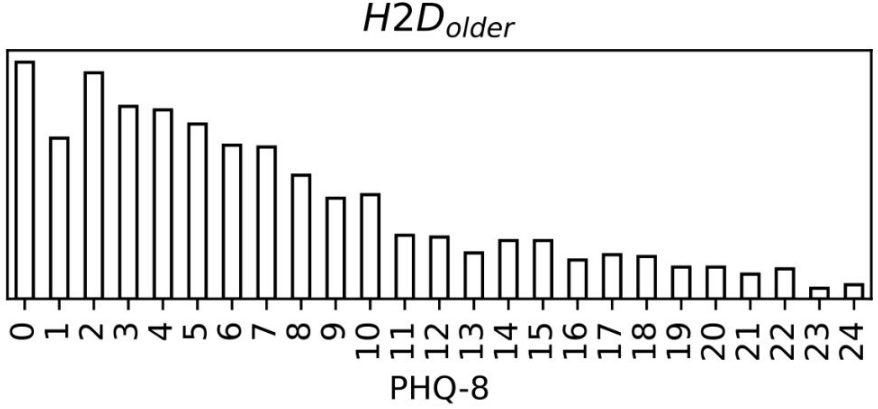
- **Purpose**: Represents real-world settings with human-to-human interactions.
- **Collection Method**:
  - Speech samples obtained from case management calls recorded by a healthcare provider.
  - PHQ-8 labels derived from verbal administration during the call.
  - Survey responses removed from recordings to ensure unbiased evaluation.
- **Statistics**:
  - **Samples**: 669 recordings, each from unique participants.
  - **Speech Duration**: Conversations lasting up to 40 minutes, covering broad topics.
  - **Demographics**:
    - Age groups:
      - ≤39 years: 182 participants.
      - 40–64 years: 343 participants.
      - 65+ years: 144 participants.
    - Gender split: 69% female, 31% male.
- **Key Feature**:
  - Passive, naturalistic speech data capturing diverse conversational topics.

Datasets PHQ Label Distribution



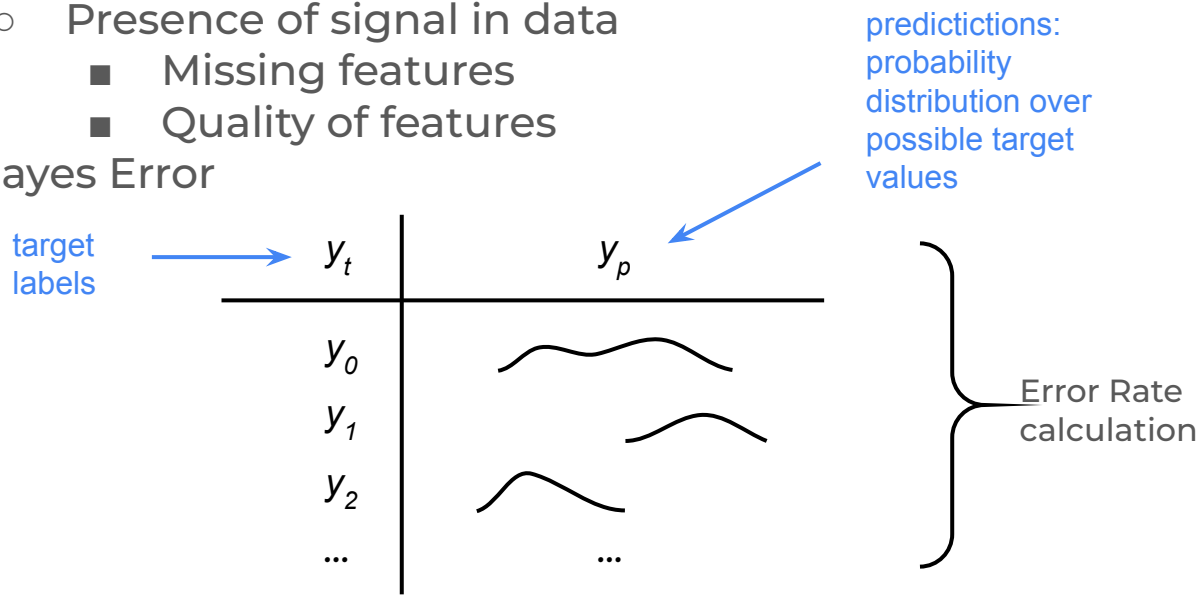






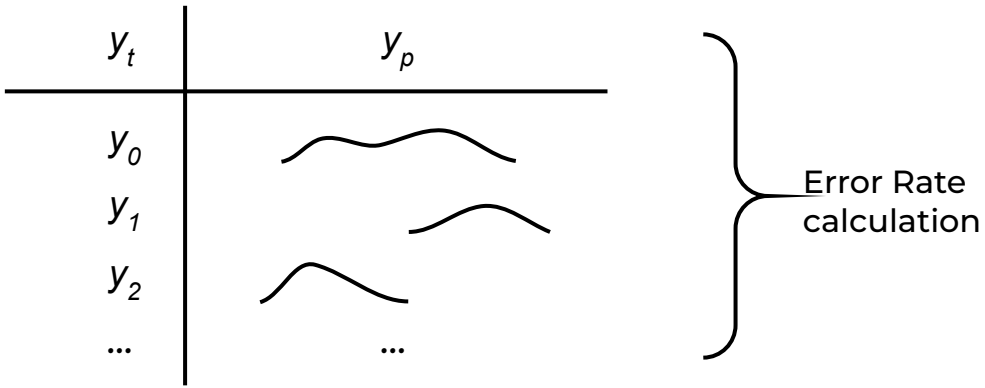
Upper Performance Bounds

- Irreducible sources of error limit attainable model performance
  - Noise in labels
  - Natural variability of labels
  - Presence of signal in data
    - Missing features
    - Quality of features
- Bayes Error

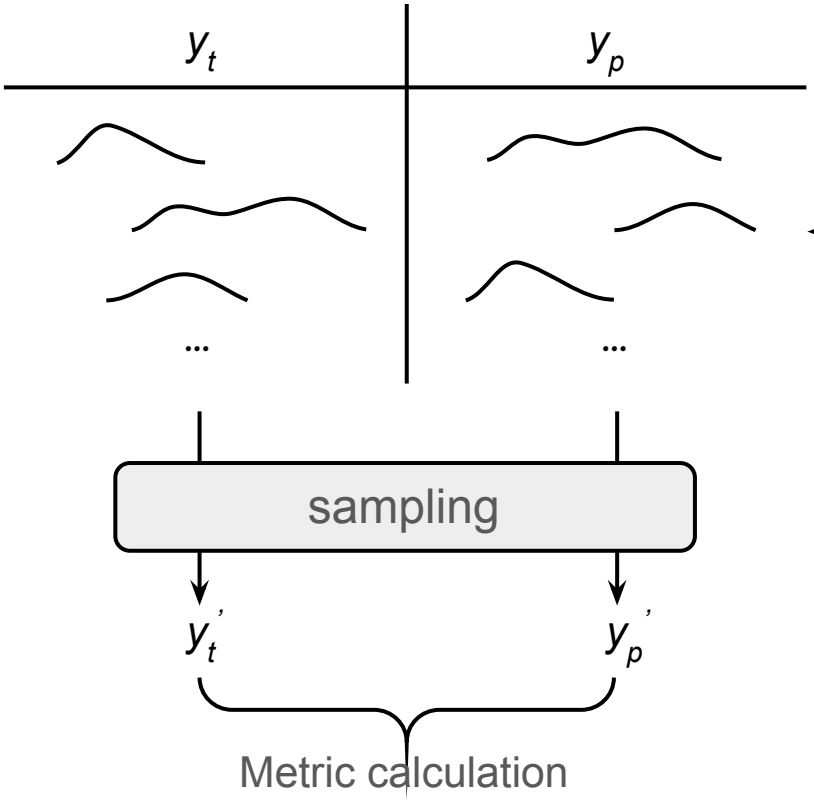


Upper performance bounds

- How to estimate Bayes Error?
  - Use a BE estimator
    - Dimensionality
    - Only error rate
    - No estimators for regression
  - Multiple expert annotators as a surrogate
    - Incompatible with self-assessment

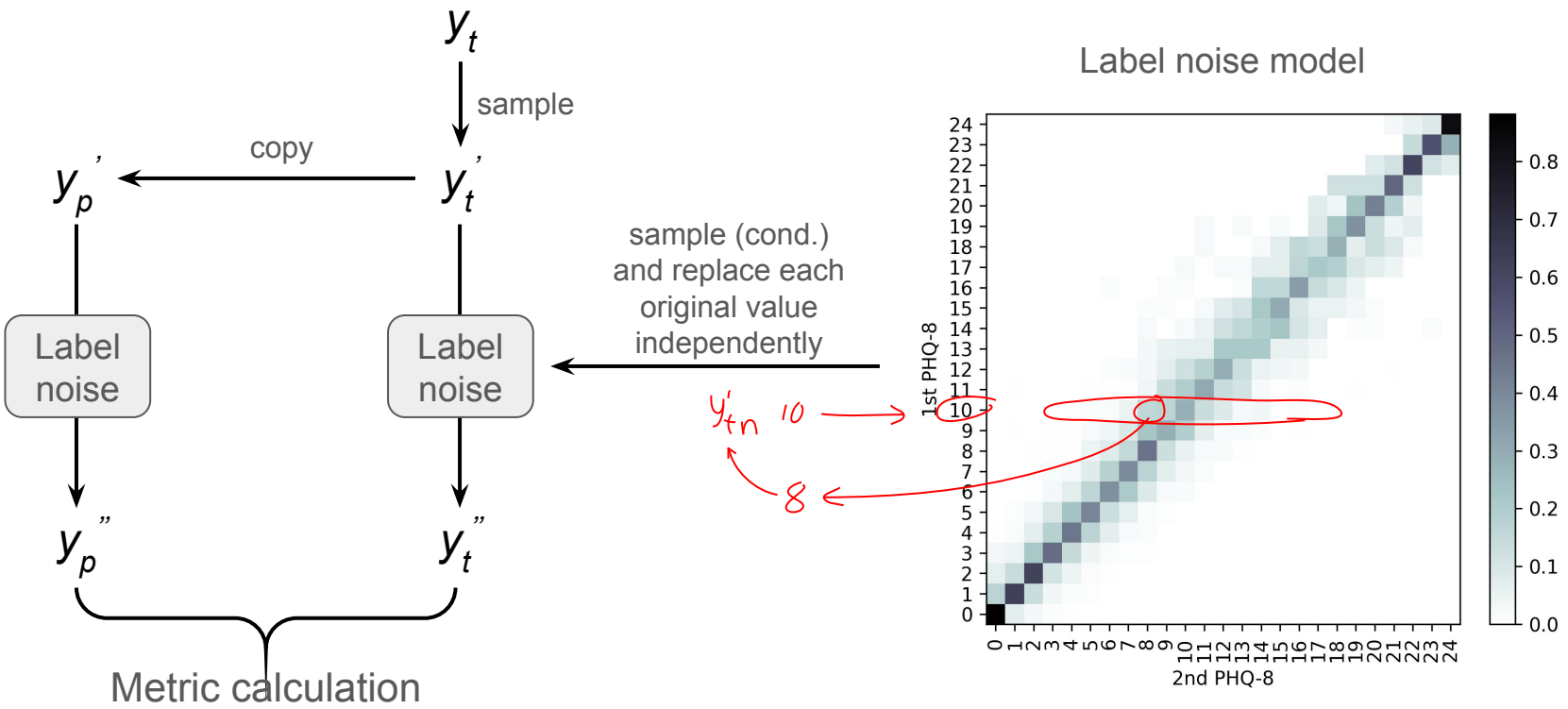


Our approach



Label noise model ← Repeated PHQ-8 administration

Our approach



Low Noise Model

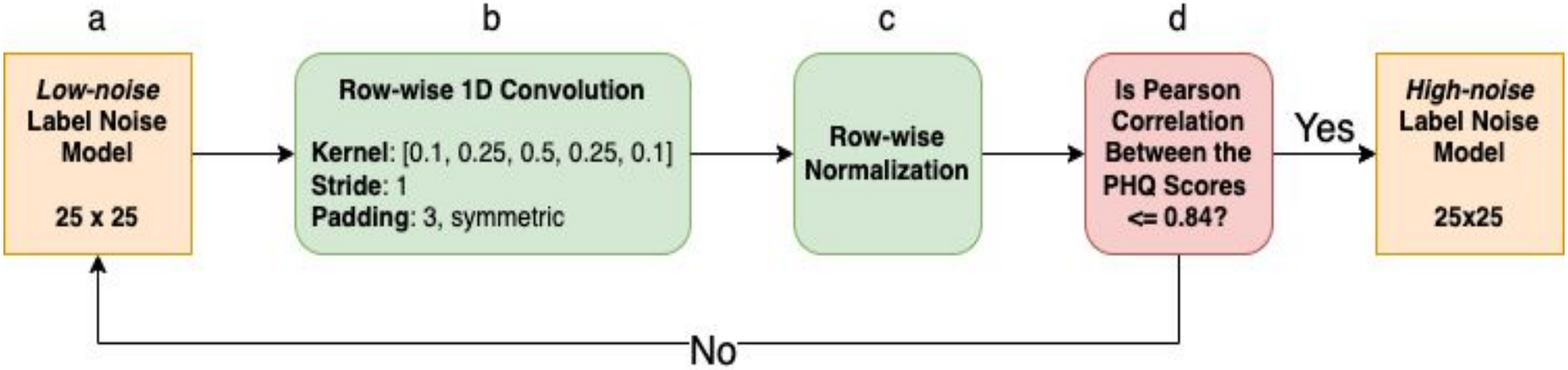
- Represents scenarios with minimal label noise.
- Derived from highly controlled data collection processes.
- **Data Source**:
  - Based on 5,625 sessions from the H2Ddual dataset.
  - PHQ-8 labels collected under optimal conditions:
    - Same session test-retest measurements with short intervals.
    - High-quality control measures to minimize external noise.
- **Characteristics**:
  - Strong test-retest reliability.
  - Reflects the upper limit of label quality achievable in research settings.
  - Probability distributions of PHQ-8 values show narrow spreads.
- **Purpose**:
  - Establishes a baseline for evaluating model performance in ideal conditions.
  - Helps identify how real-world noise diverges from ideal conditions.

High Noise Model (1)

- Simulates conditions with substantial label noise.
- Represents scenarios with lower test-retest reliability.
- **Creation Process**:
  - Developed by degrading the Low Noise model iteratively:
    - Applied a 1D convolution operation using a kernel [0.1, 0.25, 0.5, 0.25, 0.1].
    - Normalized the resulting values to simulate noise distribution.
    - Process stopped when the Pearson correlation coefficient between test-retest PHQ-8 scores dropped to 0.84.
- **Data Source**:
  - Derived from the same H2Ddual dataset as the Low Noise model.
  - Emulates lower-quality data collection processes.

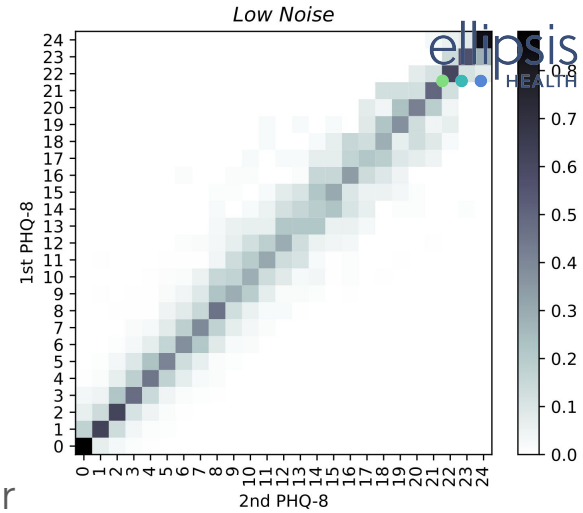
High Noise Model (2)

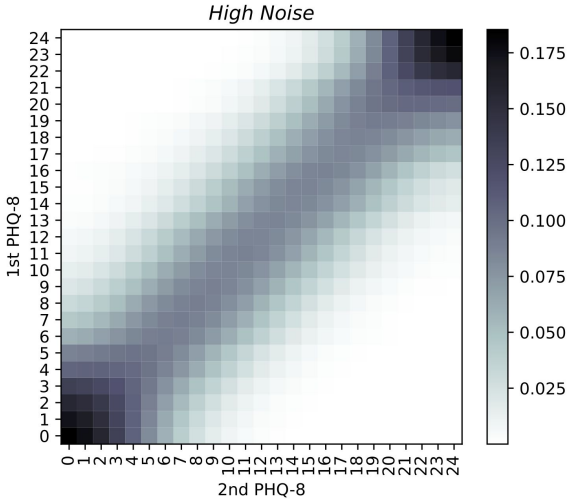
- **Characteristics**:
  - Broader probability distributions of PHQ-8 values.
  - Mimics lower test-retest reliability similar to studies using phone interviews for PHQ-8.
- **Purpose**:
  - Reflects noisy real-world conditions where external factors influence label reliability.
  - Useful for evaluating model robustness in less-controlled environments.



Comparison Between Models

- **Noise Levels**:
  - Low Noise: High reliability, narrow test-retest spread.
  - High Noise: Reduced reliability, wider test-retest spread.
- **Impact on Model Evaluation**:
  - The choice of noise model affects the estimated upper and lower bounds of model performance.
  - High noise reduces maximum achievable performance and increases performance uncertainty.
- **Applications**:
  - Low Noise: Useful for benchmarking models under ideal conditions.
  - High Noise: Highlights robustness challenges in realistic scenarios.





Methodology

- Each dataset underwent human and automated quality control to ensure high-quality input.
- Removed samples from participants who did not adhere to the recording or survey tasks (e.g., missing voice recordings).
- Surveys were administered twice (H2Ddual dataset) to measure label noise through test-retest reliability.
- PHQ-8 labels in other datasets derived from single-session surveys or verbal questionnaires.
- Speech samples ranged from 1 minute (prompted app-based datasets) to 40 minutes (case management recordings).
- Datasets captured diverse demographic groups and interaction modes (e.g., app-based, human-to-human).

Rigorous Scientific Evaluation (1)

- Blind sets: No speaker overlap in train/dev/test partitions
- **Regression Metrics**:
  - Mean Absolute Error (MAE): Measures prediction accuracy.
  - Concordance Correlation Coefficient (CCC): Assesses consistency between predicted and actual PHQ-8 scores.
- **Binary Metrics**:
  - ROC AUC (Receiver Operating Characteristic Area Under Curve): Measures the model's ability to differentiate between classes.
  - Accuracy: The fraction of correct predictions.

Noise Impact Analysis:

- Evaluated models using both Low Noise and High Noise scenarios.
- Measured how noise levels influence:
  - Upper performance bounds: The theoretical maximum achievable performance.
  - Lower performance bounds: Baseline performance achievable with random predictions.

Rigorous Scientific Evaluation (2)

Bootstrapping:

- Repeated sampling (10,000 iterations) to generate robust performance distributions.
- Used to calculate confidence intervals for performance metrics.

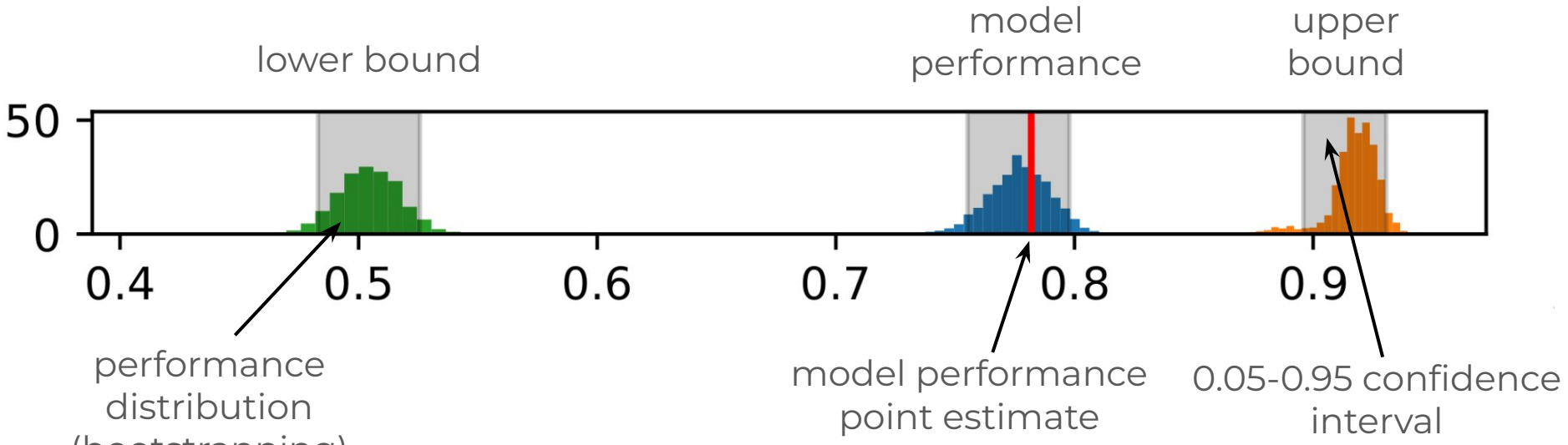
Evaluation Workflow:

- **Step 1**: Generate upper and lower bounds for each dataset using label noise models.
- **Step 2**: Compare actual model performance against these bounds.
- **Step 3**: Use metrics to assess discrepancies and robustness across datasets and subgroups.

Probabilistic Performance Bounds:

- Upper bounds calculated by simulating "perfect" predictions aligned with the noise model.
- Lower bounds calculated by assuming random predictions based solely on label distributions.
- Bounds provide a realistic range for model performance expectations.

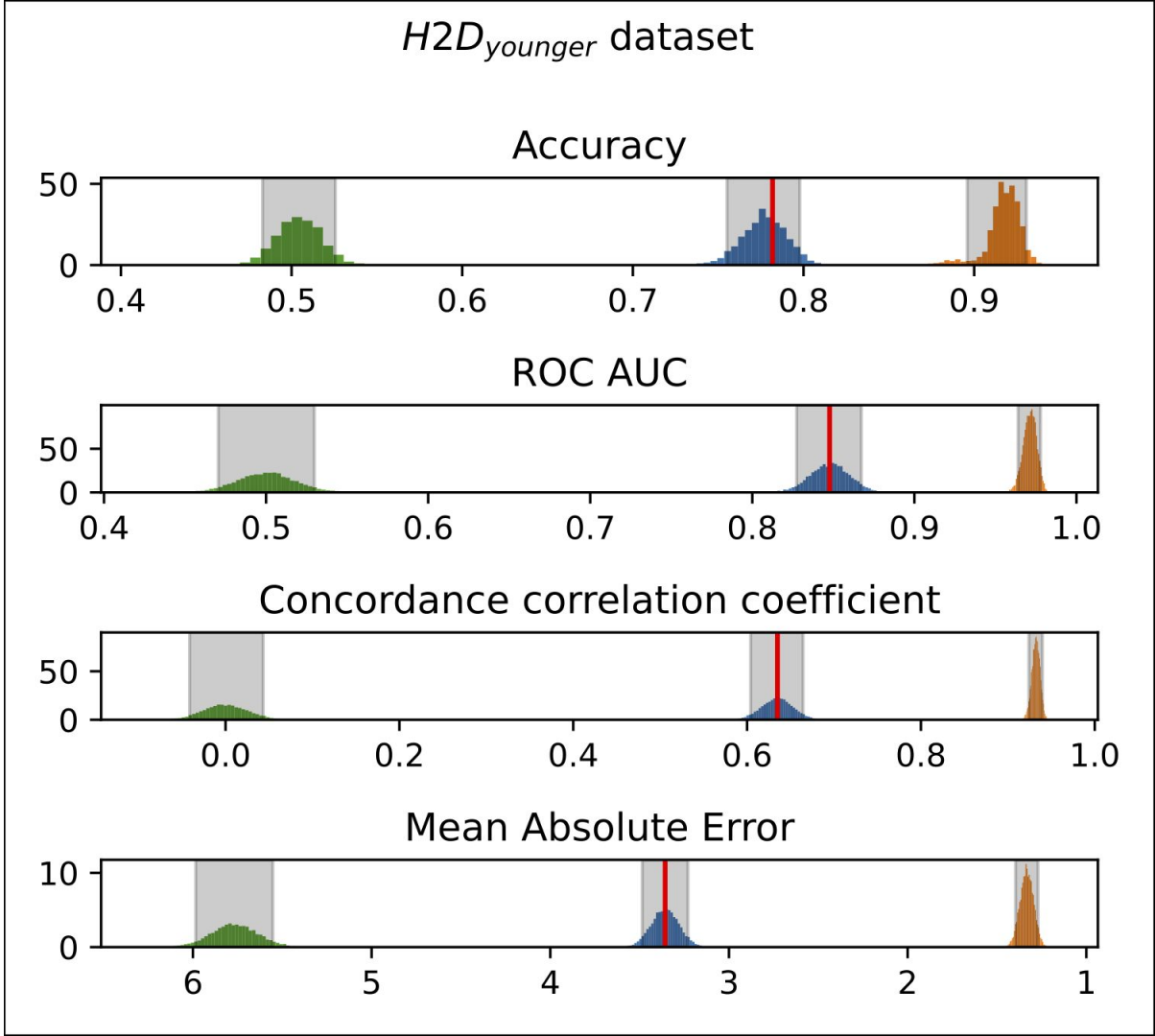
Results Visualization



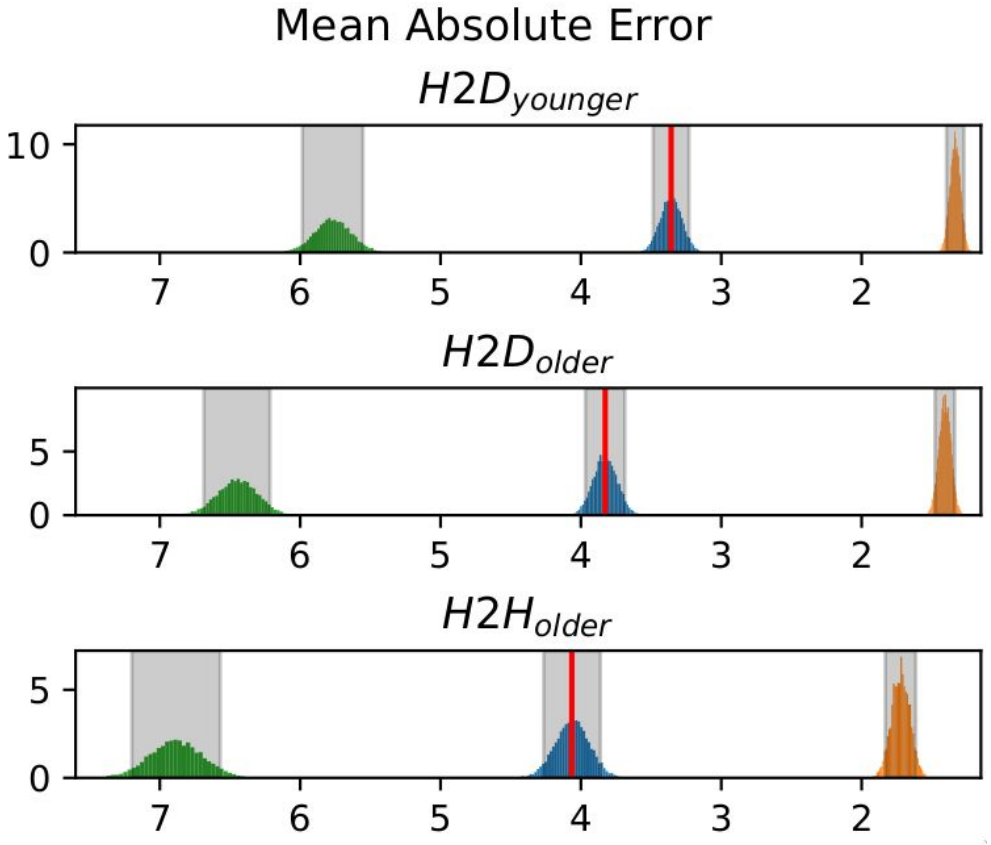
- Width of all distributions affected by:
  - y_t size (bigger datasets = lower variance)
  - Label distribution in y_t
- Upper bound affected by the noise model
- Other sources of irreducible error omitted

Results

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



Results



- Different datasets and models
- Substantial differences in the position of lower bound between the datasets
- Substantial difference in the position of upper bound between the third dataset and the rest
- Picture more complex than looking only at point estimates of model performance

Results

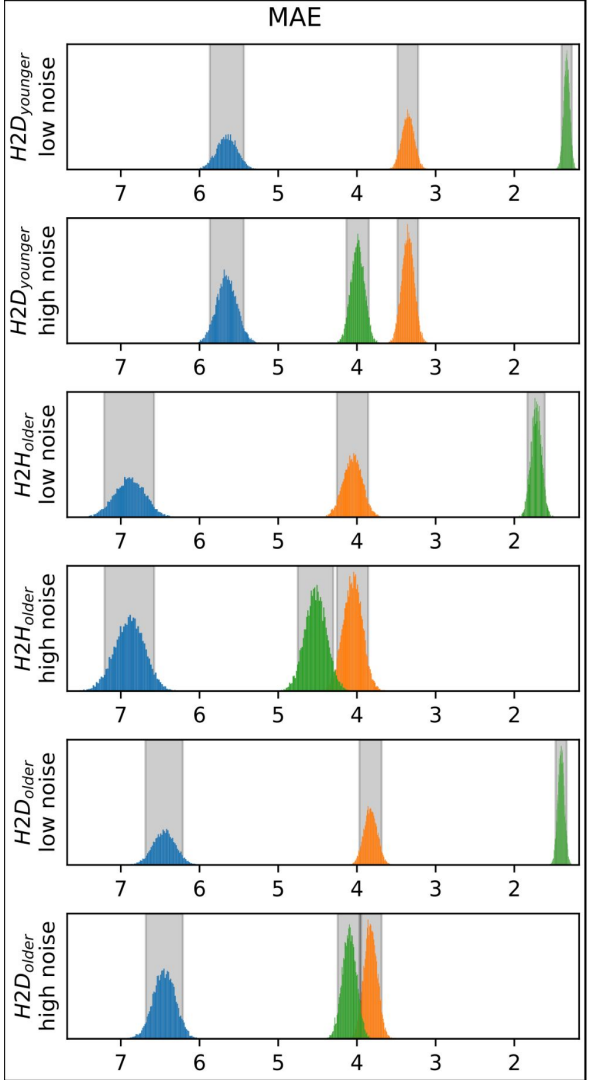
Low Noise: Narrow performance bounds, better alignment with actual data.

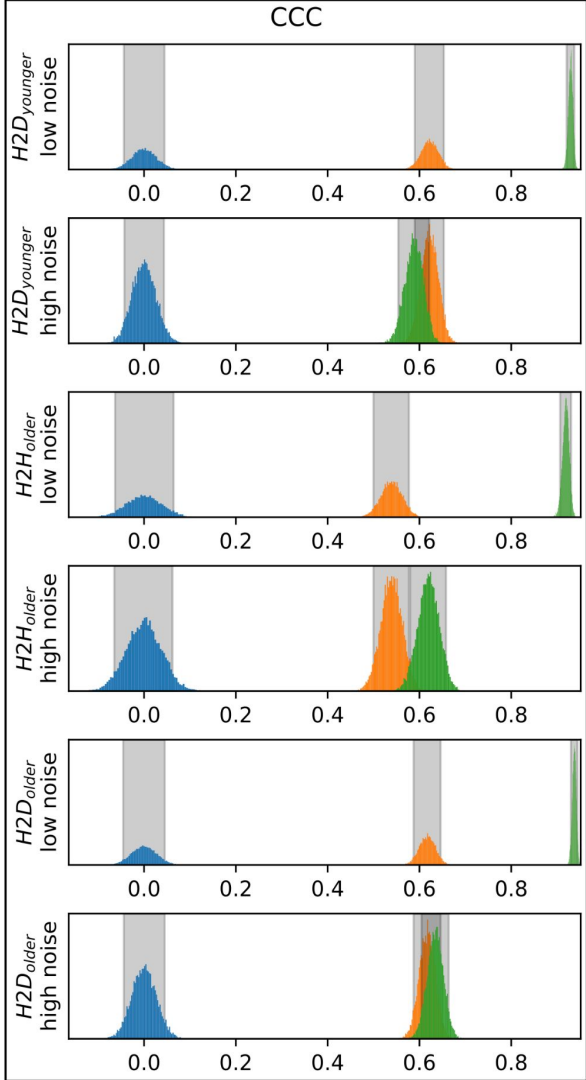
High Noise:

- Wider bounds, mismatched with cleaner datasets.
- Highlights challenges in noisy real-world conditions.

Observations:

- MAE and CCC show significant deviations under high noise.
- Clear performance gaps across datasets with varying noise models.





Results

High noise impacts ROC AUC and Accuracy.

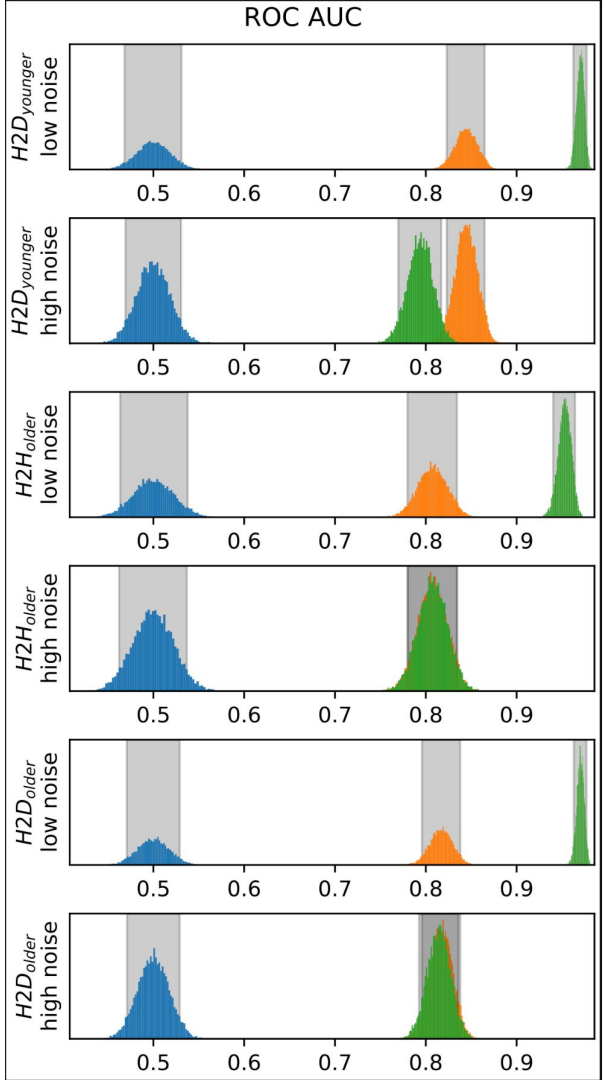
Performance differences are dataset-dependent.

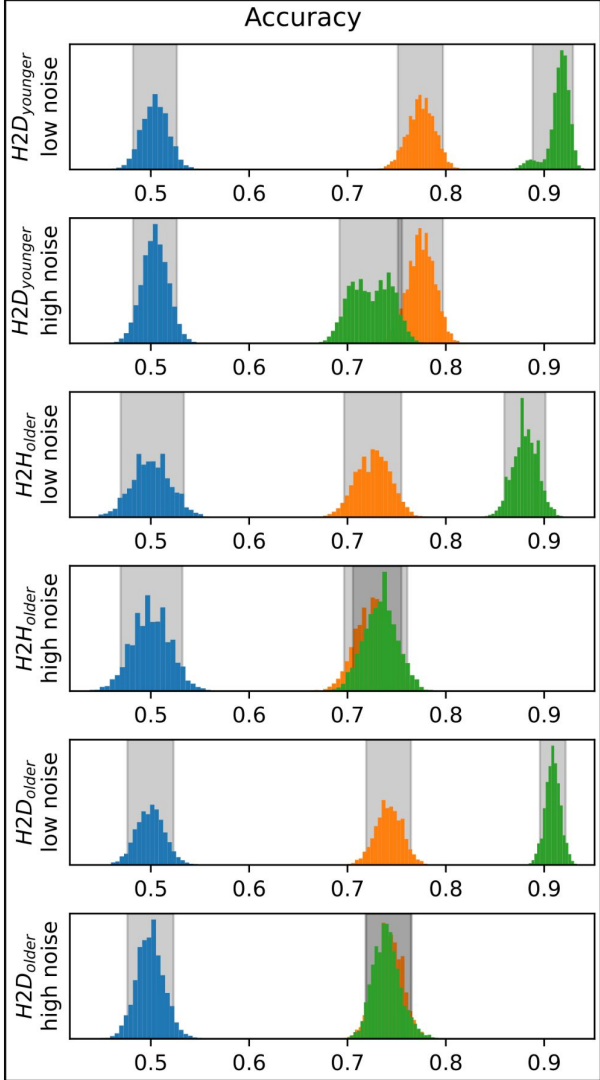
Probabilistic bounds help identify noise effects.

Challenges:

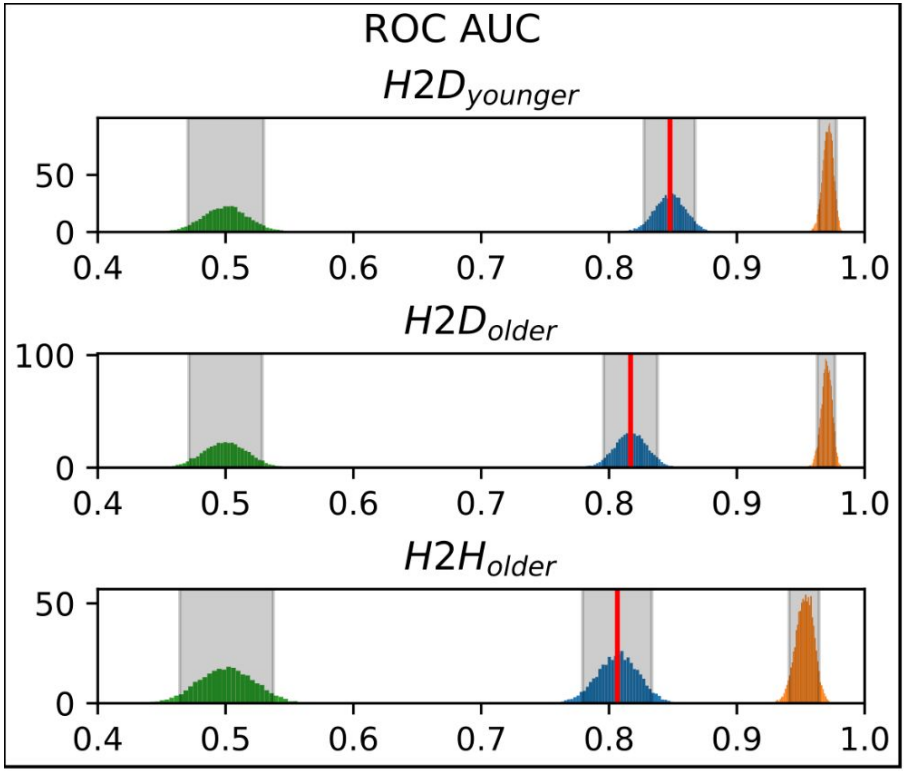
- Noise reduces model evaluation reliability.
- High noise datasets may not reflect practical applications.

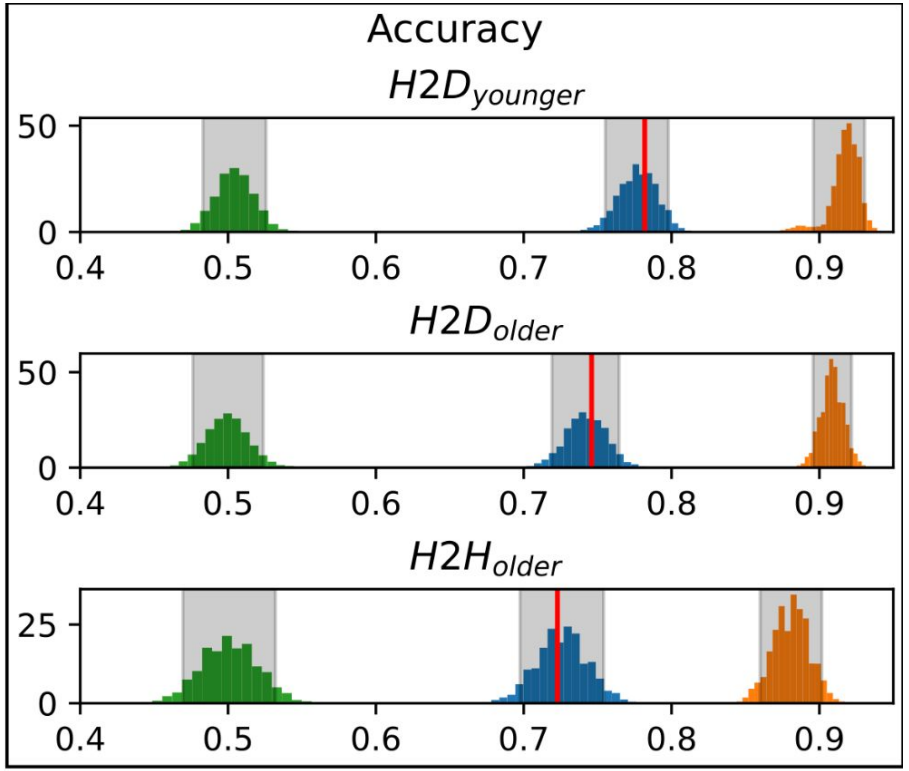
Recommendation: Focus on data collection quality for real-world scenarios.



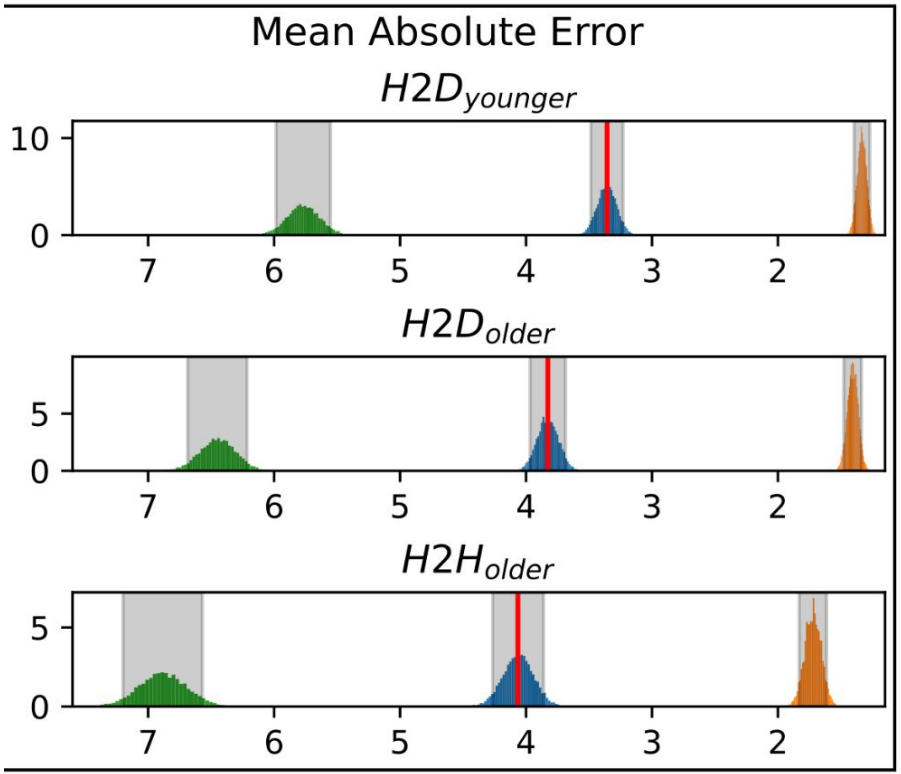


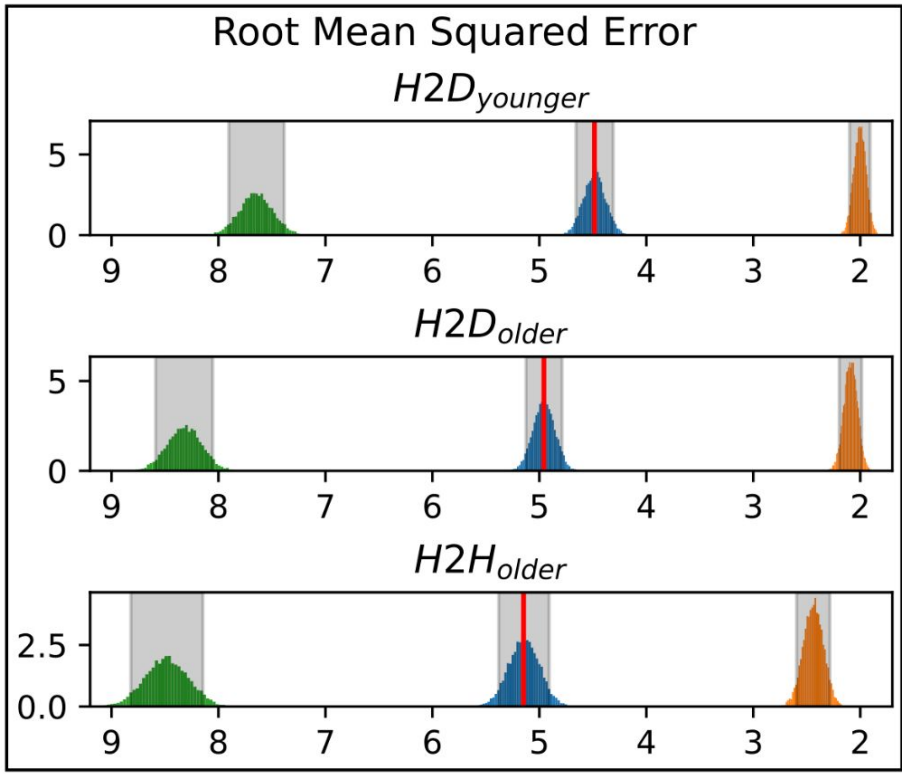
Results



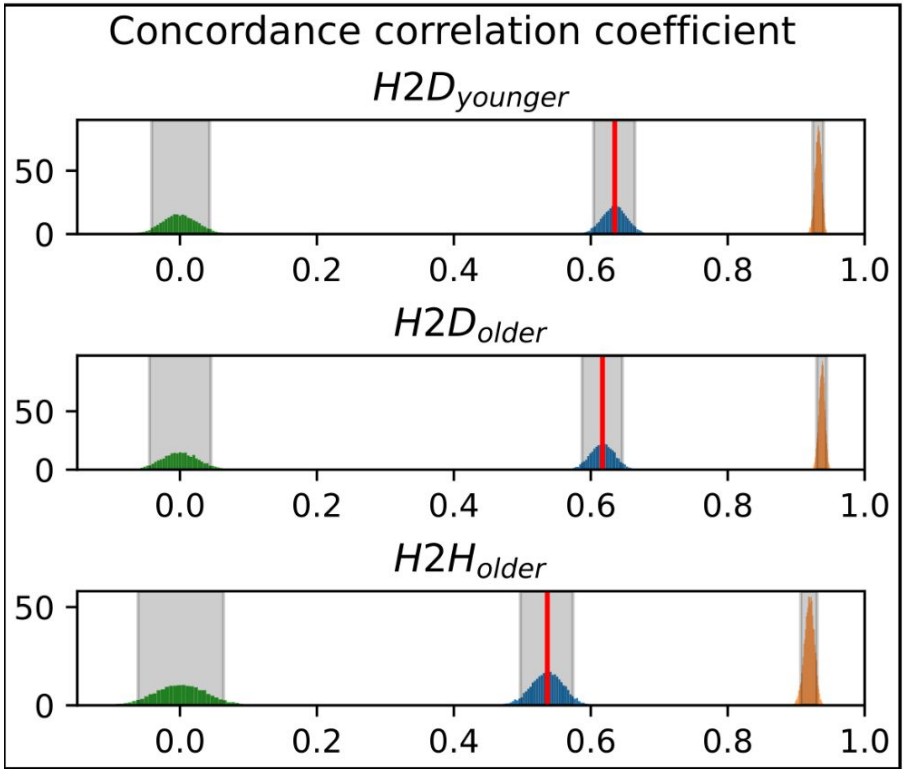


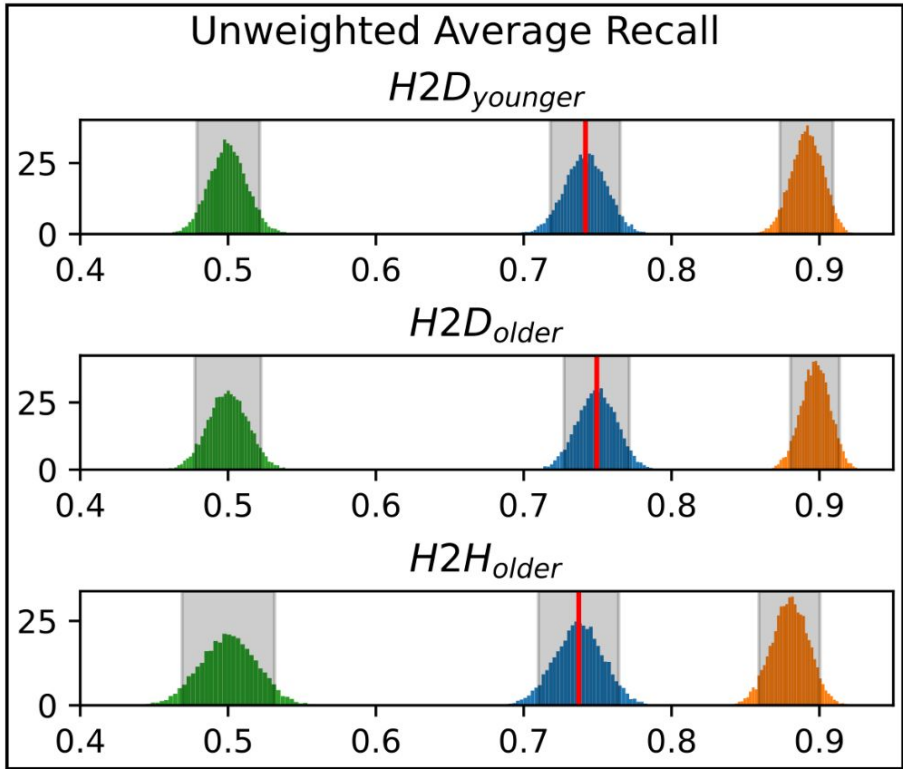
Results





Results





Conclusions

- 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
- **Importance of Data Collection**:
  - Methodology directly impacts label reliability and model evaluation.
- **Test-Retest Reliability**:
  - Variations in label noise highlight the need for robust evaluation methods that account for real-world challenges.
- **Generalization**:
  - These models provide frameworks that can be adapted to other datasets or scenarios with similar challenges.

Conclusions

- Models ran on conversations after survey removal
- Fused output of language and acoustic models for best performance
- No loss in performance from Dev to Blind Test (!)
- Excellent generalization of models over metadata subgroups
    - age, gender, MA/nonMA, BH/nonBH, Social Vulnerability Index
- AUC in 0.80s
    - AUC largely stable across cut-off thresholds (5, 10, 15, 20)
- MAE values in 3s and 4s
- Obtained gain from fusion
- Fine-tuning experiments show gains in CCC

Questions

Feel free to contact me at marija.stanojevic@ellipsishealth.com

We are hiring: https://www.ellipsishealth.com/join-us

See more details about me at
https://marija-stanojevic.github.io/