ellipsis HEALTH

Robust Speech and Natural Language Processing Models for Depression Screening

Yang Lu, Amir Harati, Tomasz Rutowski, Ricardo Oliveira, Elizabeth Shriberg

2020 IEEE Signal Processing in Medicine and Biology Symposium

December 5th 2020

Depression

Major depressive disorder

Prolonged, feelings of sadness and/or a loss of interest in activities. Emotional, physical problems. May feel life is not worth living.

American Psychiatric Association, 2020 (adapted and abbreviated)

Depression is a leading cause of disability worldwide and a major contributor to overall global burden of disease *

* World Health Organization, 2019-2020

Depression

Screening
Treatment

An estimated 60% of patients at risk for depression do not get screened

Successful treatment exists

Cues to depression

Prosodic
(rate, pitch, energy)

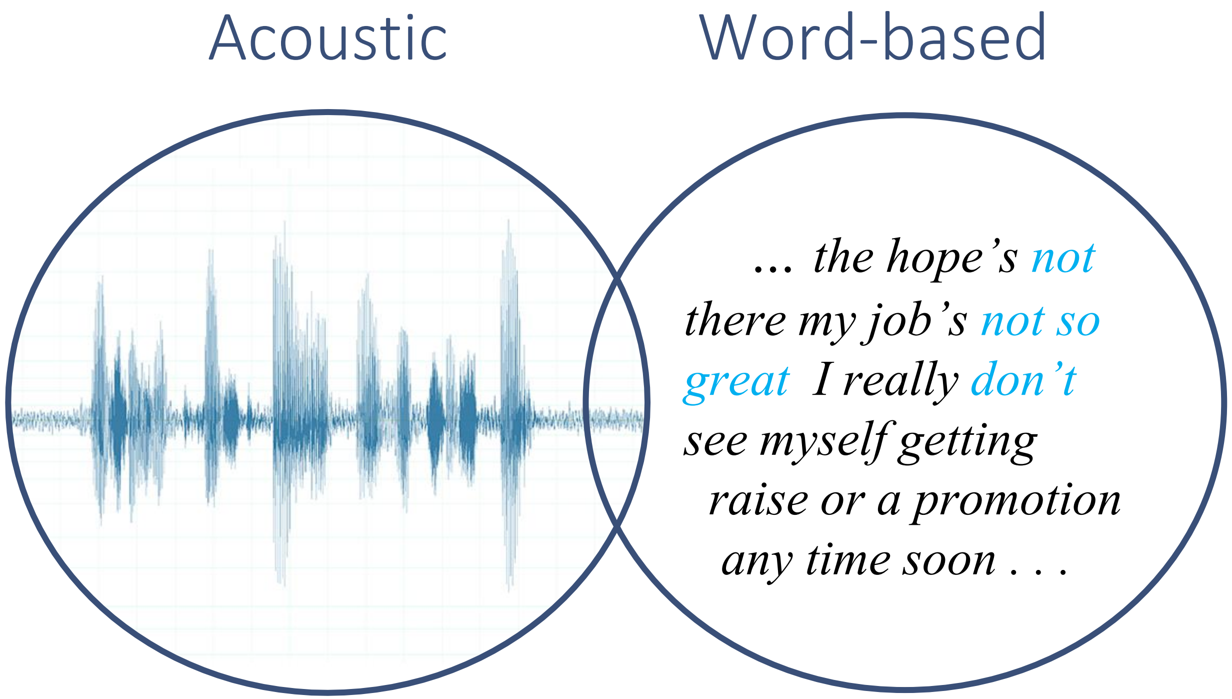
Glottal features

Formant-based

Spectral

Cepstral

. . .



Acoustic

Word-based

… the hope's not there my job's not so great I really don't see myself getting raise or a promotion any time soon . . .

Negative sentiment

Fewer positive words

Self-focus

First person pronouns

Word pattern dictionaries

Simplified syntax

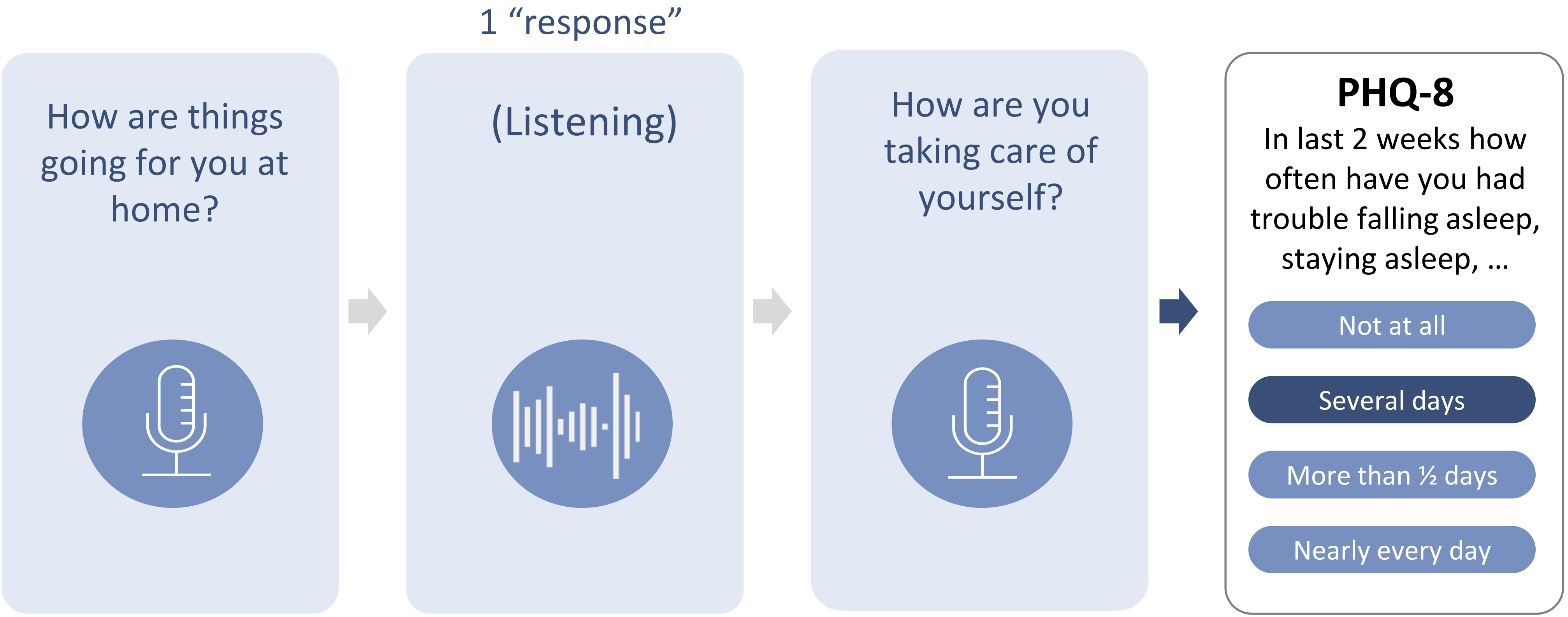
. . .

Data Collection for Depression Screening by Speech

Data: Speech

Speaking part, ~4-5 min. Users choose topics; start/end speech

Label: PHQ-8 Score



1 “response”

How are things going for you at home?

(Listening)

How are you taking care of yourself?

PHQ-8

In last 2 weeks how often have you had trouble falling asleep, staying asleep, …

Not at all

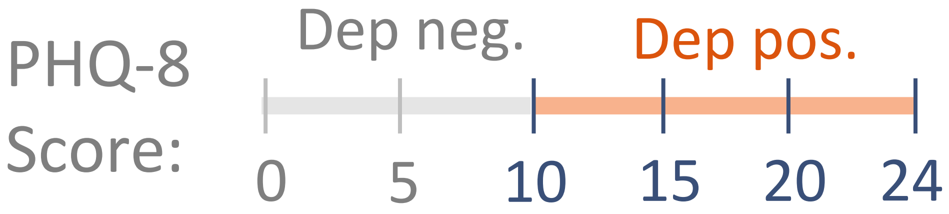
[✓ Several days]

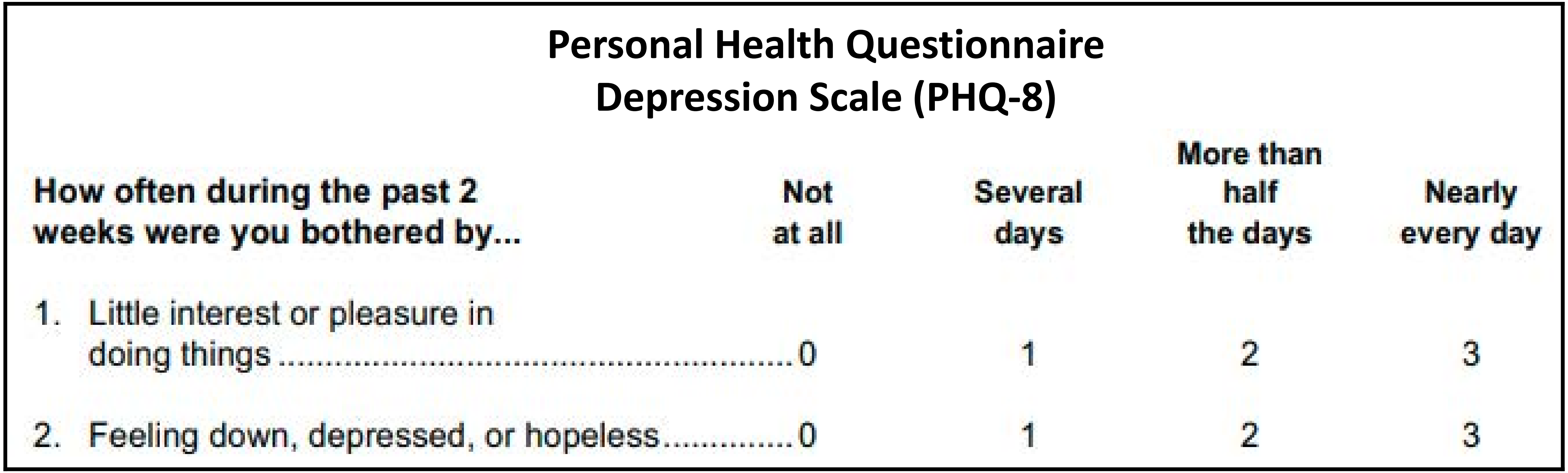
More than ½ days

Nearly every day

Label: PHQ-8 best approximation of patient state

Data: Speech





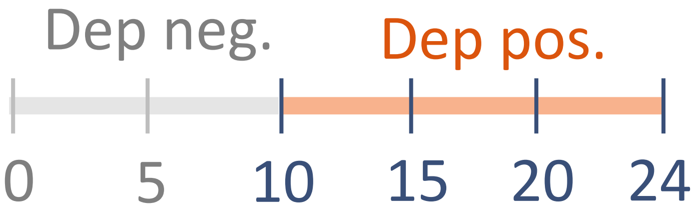
**Personal Health Questionnaire
Depression Scale (PHQ-8)**

How often during the past 2 weeks were you bothered by...	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless..............	0	1	2	3

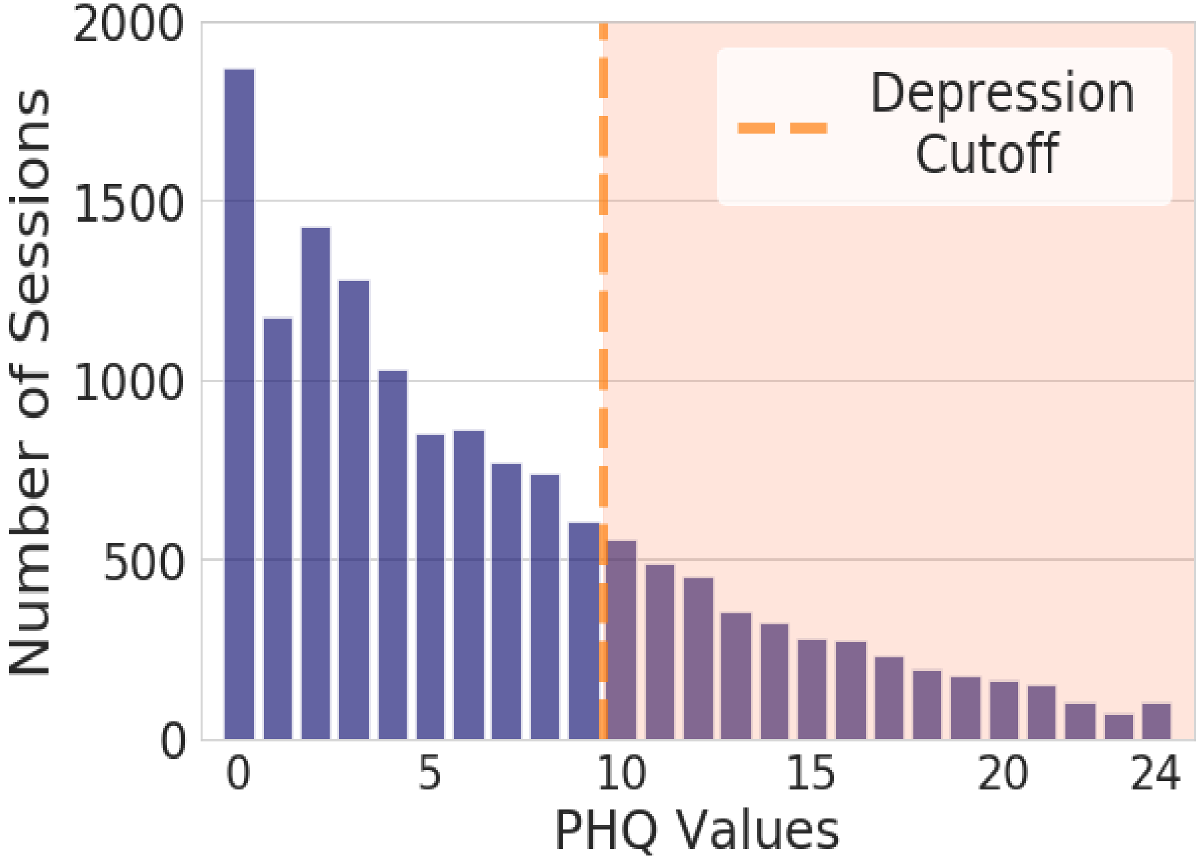
Label: PHQ-8 best approximation of patient state

Data: Speech

PHQ-8 Score:



PHQ-8 Distribution of our corpus



Personal He

Depressic

How often during the past 2 weeks were you bothered by...

1. Little interest or pleasure in doing things ....

2. Feeling down, depressed, or hopeless..............0 1 2 3

Label: PHQ-8 best approximation of patient state

Data: Speech

- American English
- Age 18-65, mean age 30
- Speakers chose topics

11K Unique Speakers	88K Question Responses	16K Sessions Mean 4.5min
1.6K Hours	26% Depressed	76% Train partition

Label: PHQ-8 best approximation of patient state

Data: Speech

- American English
- Age 18-65, mean age 30
- Speakers chose topics
- User and session metadata
  - Gender, Location
  - Time of collection

11K
Unique Speakers

88K
Question Responses

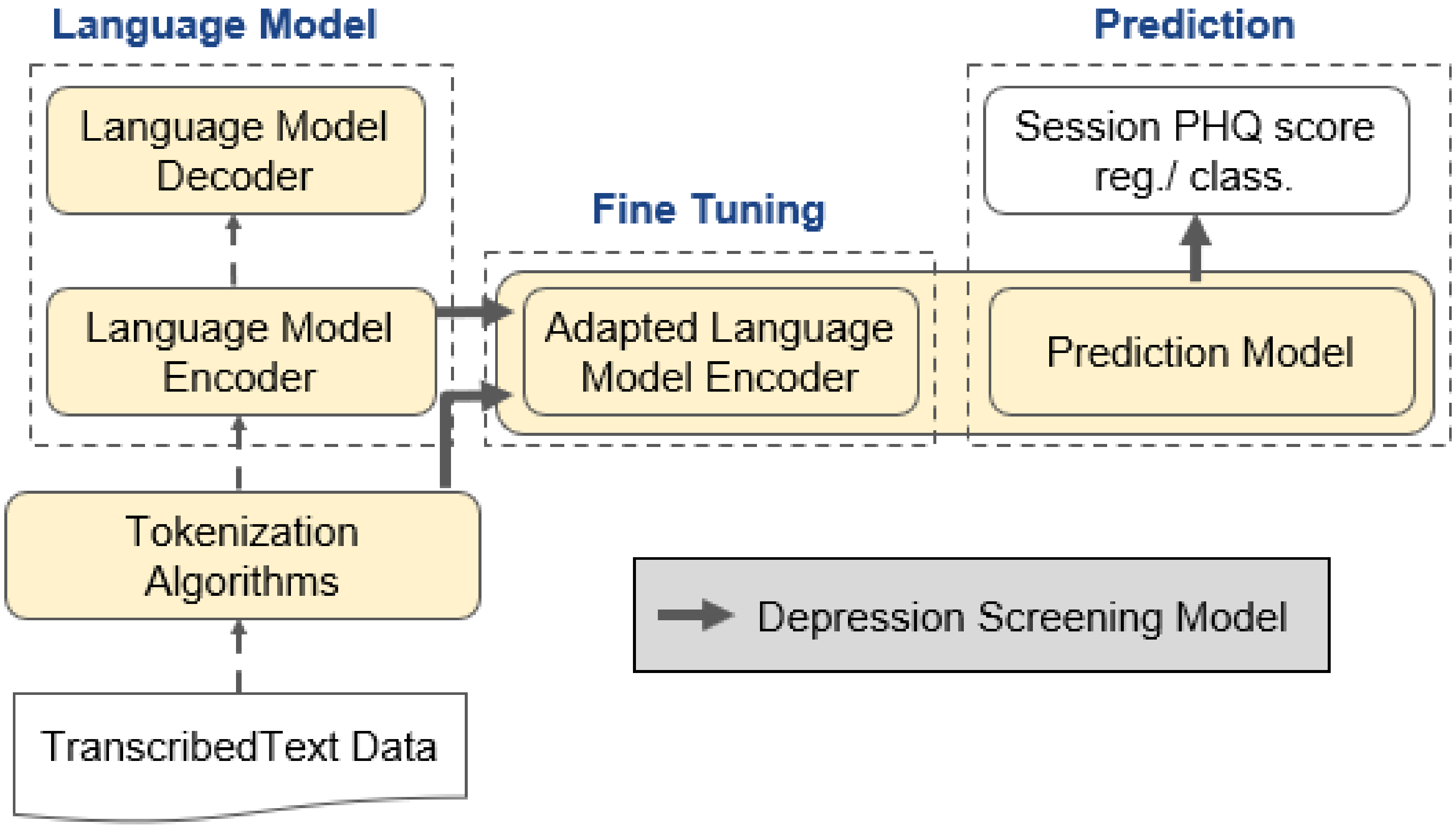
16K
Sessions
Mean 4.5min

1.6K
Hours

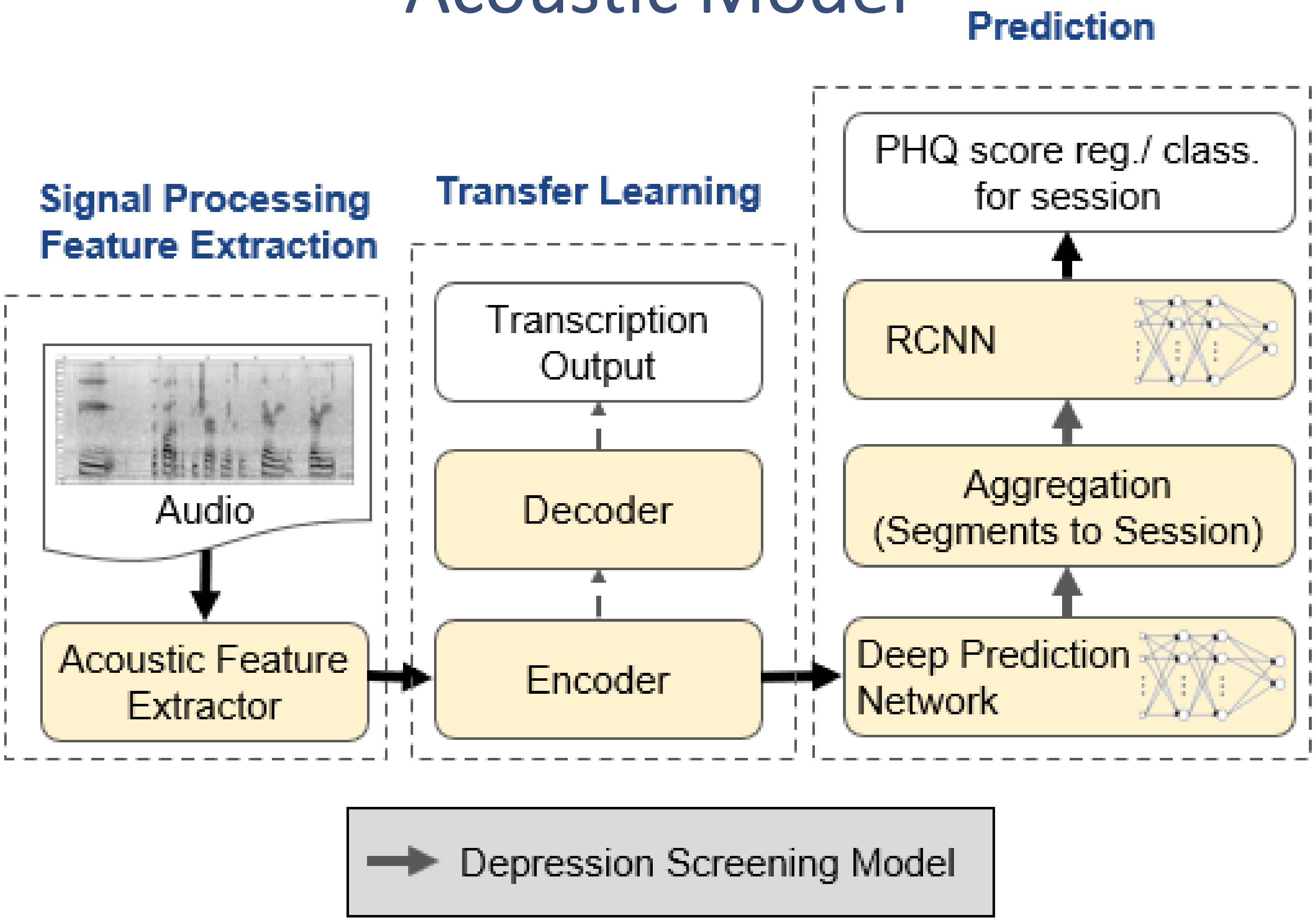
26%
Depressed

76%
Train partition

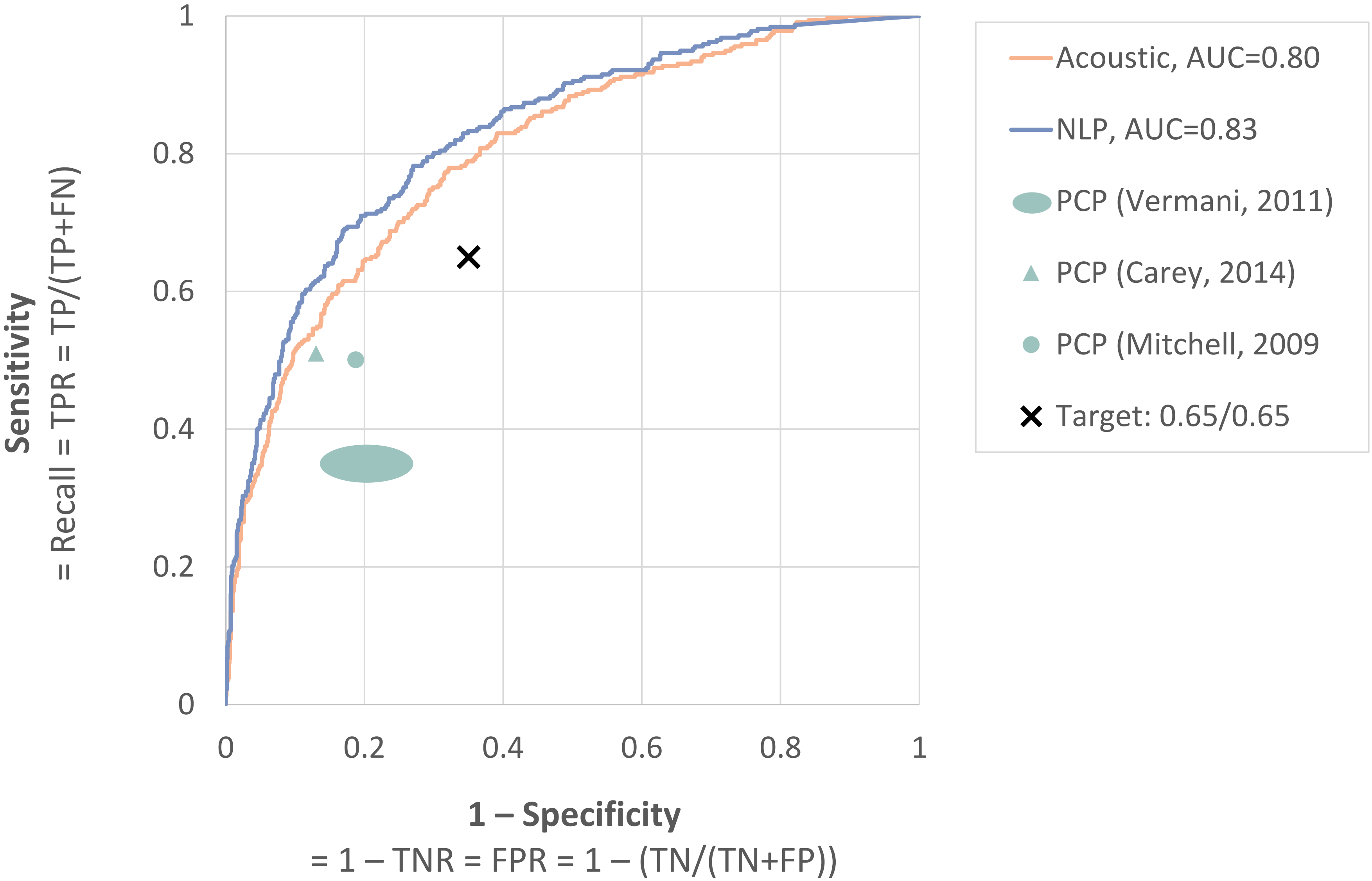
NLP Model


Language Model
Language Model Decoder
Language Model Encoder
Tokenization Algorithms
TranscribedText Data
Fine Tuning
Adapted Language Model Encoder
Prediction
Session PHQ score reg./ class.
Prediction Model
Depression Screening Model

Acoustic Model


Prediction
Signal Processing Feature Extraction
Transfer Learning
PHQ score reg./ class. for session
Transcription Output
RCNN
Audio
Decoder
Aggregation (Segments to Session)
Acoustic Feature Extractor
Encoder
Deep Prediction Network
Depression Screening Model

Acoustic and NLP Model Performance



Model generalization over various factors (no retraining)

Metadata	Categories	Mean PHQ	Acoustic model AUC	NLP model AUC
Base performance over all test set		**5.93**	**0.779**	**0.825**
Gender	Male:	5.74	0.769	0.819
	Female	6.77	0.774	0.820
Age group	18-25	7.32	0.792	0.828
	26-35	6.40	**0.752***	0.820
	36-45	5.60	0.790	0.808
	46-65	4.78	0.792	0.819
Smoking	Non-smoker	6.44	0.803	0.836
	Smoker	7.47	0.767	0.808
Ethnicity	Caucasian	6.05	0.796	0.826
	African American	5.63	0.777	0.812
	Hispanic	6.73	**0.676***	0.788
	Asian American	5.61	0.789	0.841
	Mixed	7.22	0.768	0.827

* Significant at $p<.05$ in DeLong test for AUC

Model generalization over various factors (no retraining)

Metadata	Categories	Mean PHQ	Acoustic model AUC	NLP model AUC
Base performance over all test set		**5.93**	**0.779**	**0.825**
Gender	Male:	5.74	0.769	0.819
	Female	6.77	0.774	0.820
Age group	18-25	7.32	0.792	0.828
	26-35	6.40	**0.752***	0.820
	36-45	5.60	0.790	0.808
	46-65	4.78	0.792	0.819
Smoking	Non-smoker	6.44	0.803	0.836
	Smoker	7.47	0.767	0.808
Ethnicity	Caucasian	6.05	0.796	0.826
	African American	5.63	0.777	0.812
	Hispanic	6.73	**0.676***	0.788
	Asian American	5.61	0.789	0.841
	Mixed	7.22	0.768	0.827

* Significant at $p<.05$ in DeLong test for AUC

Model generalization over various factors (no retraining)

ellipsis HEALTH

Metadata	Categories	Mean PHQ	Acoustic model AUC	NLP model AUC
Base performance over all test set		**5.93**	**0.779**	**0.825**
Gender	Male:	5.74	0.769	0.819
	Female	6.77	0.774	0.820
Age group	18-25	7.32	0.792	0.828
	26-35	6.40	**0.752***	0.820
	36-45	5.60	0.790	0.808
	46-65	4.78	0.792	0.819
Smoking	Non-smoker	6.44	0.803	0.836
	Smoker	7.47	0.767	0.808
Ethnicity	Caucasian	6.05	0.796	0.826
	African American	5.63	0.777	0.812
	Hispanic	6.73	**0.676***	0.788
	Asian American	5.61	0.789	0.841
	Mixed	7.22	0.768	0.827

* Significant at $p<.05$ in DeLong test for AUC

Model generalization over various factors (no retraining)

Metadata	Categories	Mean PHQ	Acoustic model AUC	NLP model AUC
Base performance over all test set		**5.93**	**0.779**	**0.825**
Gender	Male:	5.74	0.769	0.819
	Female	6.77	0.774	0.820
Age group	18-25	7.32	0.792	0.828
	26-35	6.40	**0.752***	0.820
	36-45	5.60	0.790	0.808
	46-65	4.78	0.792	0.819
Smoking	Non-smoker	6.44	0.803	0.836
	Smoker	7.47	0.767	0.808
Ethnicity	Caucasian	6.05	0.796	0.826
	African American	5.63	0.777	0.812
	Hispanic	6.73	**0.676***	0.788
	Asian American	5.61	0.789	0.841
	Mixed	7.22	0.768	0.827

* Significant at p<.05 in DeLong test for AUC

Model generalization over various factors (no retraining)

Metadata	Categories	Mean PHQ	Acoustic model AUC	NLP model AUC
Base performance over all test set		**5.93**	**0.779**	**0.825**
Gender	Male:	5.74	0.769	0.819
	Female	6.77	0.774	0.820
Age group	18-25	7.32	0.792	0.828
	26-35	6.40	**0.752***	0.820
	36-45	5.60	0.790	0.808
	46-65	4.78	0.792	0.819
Smoking	Non-smoker	6.44	0.803	0.836
	Smoker	7.47	0.767	0.808
Ethnicity	Caucasian	6.05	0.796	0.826
	African American	5.63	0.777	0.812
	Hispanic	6.73	**0.676***	0.788
	Asian American	5.61	0.789	0.841
	Mixed	7.22	0.768	0.827

* Significant at p<.05 in DeLong test for AUC

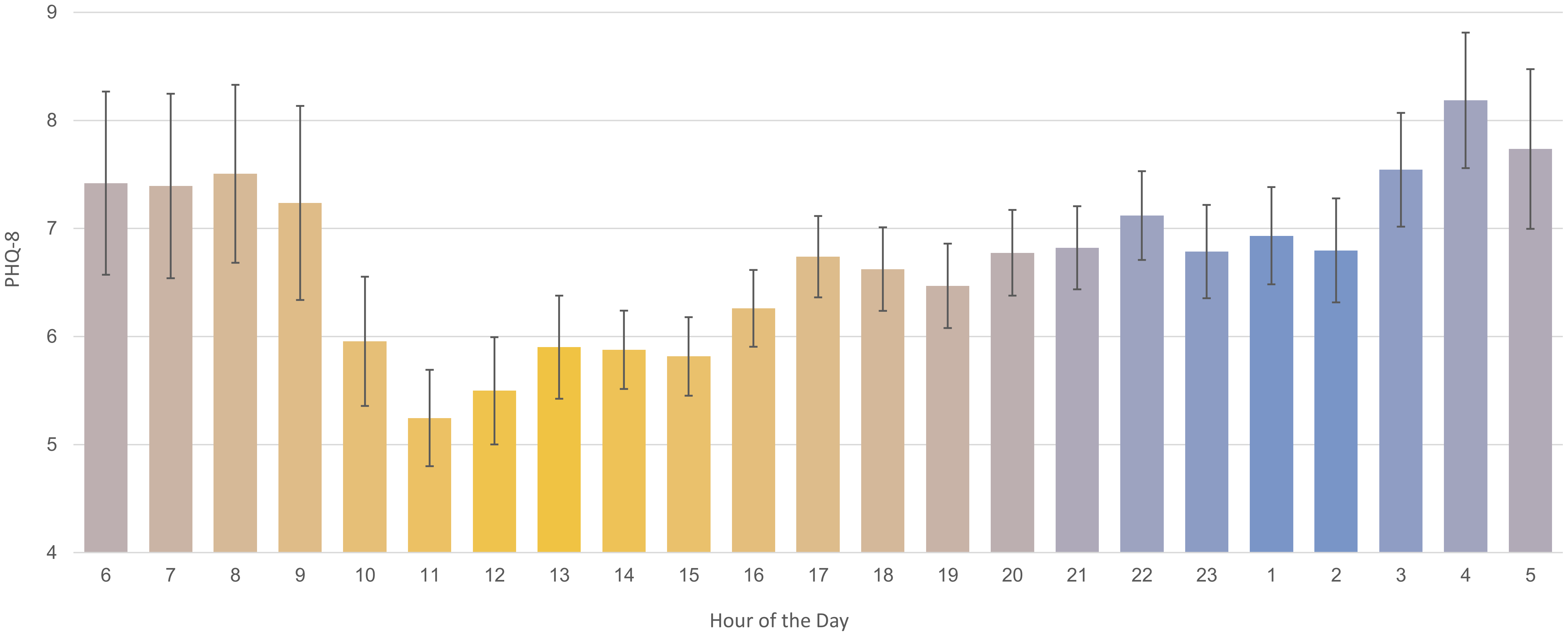
Model generalization over various factors (no retraining)

Metadata	Categories	Mean PHQ	Acoustic model AUC	NLP model AUC
Base performance over all test set		**5.93**	**0.779**	**0.825**
Gender	Male:	5.74	0.769	0.819
	Female	6.77	0.774	0.820
Age group	18-25	7.32	0.792	0.828
	26-35	6.40	**0.752***	0.820
	36-45	5.60	0.790	0.808
	46-65	4.78	0.792	0.819
Smoking	Non-smoker	6.44	0.803	0.836
	Smoker	7.47	0.767	0.808
Ethnicity	Caucasian	6.05	0.796	0.826
	African American	5.63	0.777	0.812
	Hispanic	6.73	**0.676***	0.788
	Asian American	5.61	0.789	0.841
	Mixed	7.22	0.768	0.827

* Significant at p<.05 in DeLong test for AUC

Generalized context factors: effect of time of day

Mean PHQ-8 and 95% confidence intervals by local time of day of recording



Conclusion

- Speech offers promise for digital health screening and monitoring at scale
- Our Acoustic and NLP models outperform target sensitivity/specificity
- **Models show promise for generalization over basic demographics**

Areas for further exploration

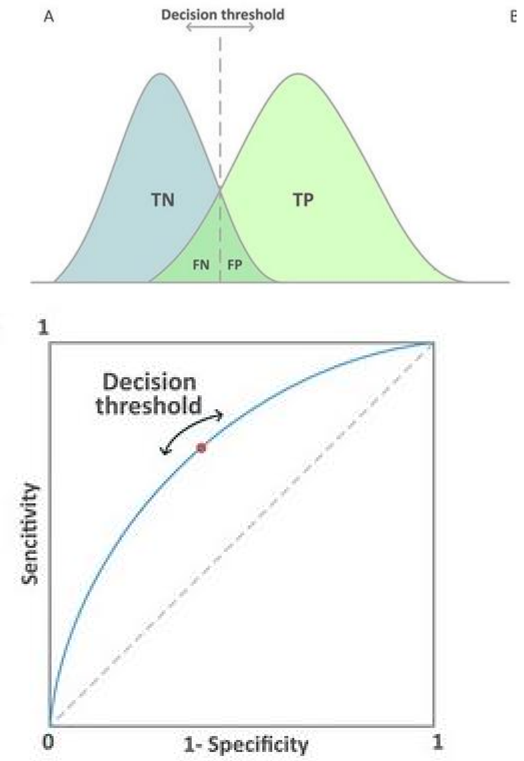
- Explainable AI
- NLP and Acoustic features affecting model performance
- Performance and robustness of combined model

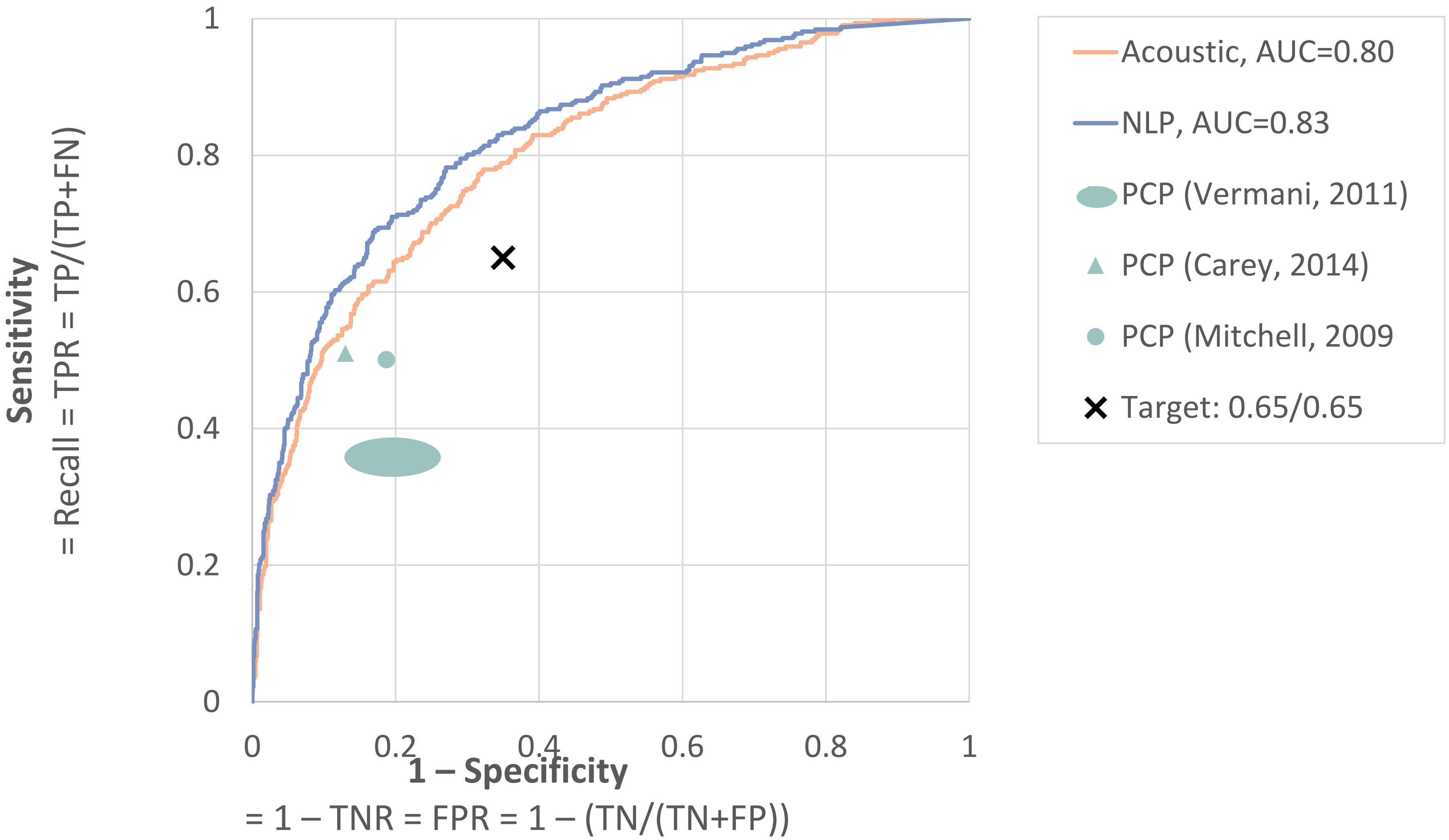
Thank You

yang@ellipsishealth.com

Acoustic and NLP Model Performances

ROC/AUC and Sensitivity Specificity trade off





Label: PHQ-8 best approximation of patient state

Data: Speech

- American English
- Age 18-65, mean age 30
- Speakers chose topics
  - self-care
  - home life
  - relationships
  - work

11K
Unique Speakers

16K
Sessions
4~5min each

1K+ hours
Total Speech Length

	Total	Train Dep-	Train Dep+	Test Dep-	Test Dep+
Sessions	15950	9266	3606	2425	653
Hours	1130	795	335	234	69
Words	11.68M	6.40M	2.64M	1.87M	0.53M

Model generalization over various factors (no 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

* Significant at $p<.05$ in DeLong test for AUC