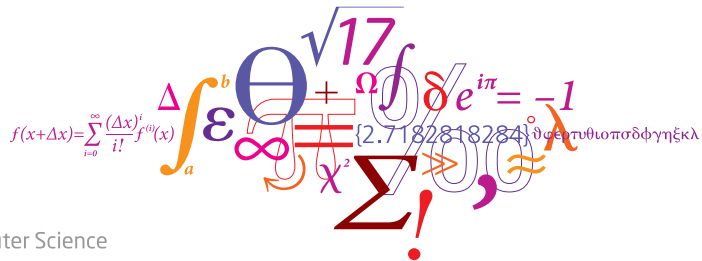
On Quality of Voice-biomarker Software: Considerations before Adoption in Healthcare

Sneha Das, Assistant Professor

Department of applied Mathematics and Computer Science,

Technical University of Denmark (DTU)

sned@dtu.dk

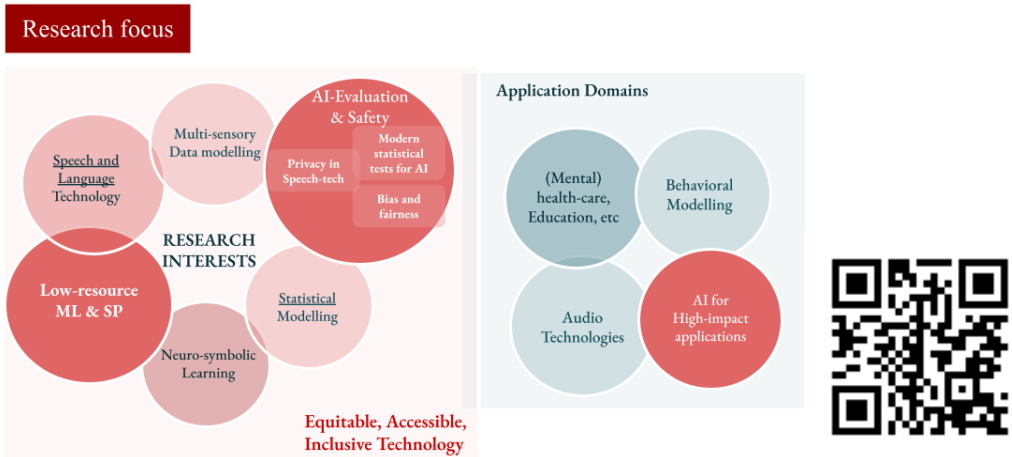


DTU Compute

Department of Applied Mathematics and Computer Science

Who am I?

- Engineer by training, and I work with computers and AI.



Outline

- Terminology
- Evaluation of Voice-biomarkers AI - Why?
- Standard metrics (in healthcare) - Classification
- Standard metrics (in healthcare) - Regression
- Design of evaluation
- Future directions
- Conclusion

Quality of voice-biomarker software

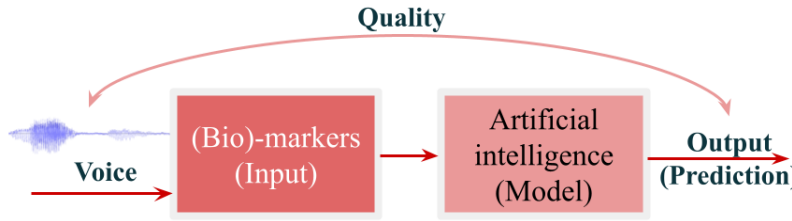


Figure: Terminology

- **Software tool**: AI model
- **Voice-biomarkers**: input features extracted from voice
- **Quality**: Testing, Evaluation, validation - Assess system performance and impact on patient outcomes.

Quality of voice-biomarker software $\Longrightarrow$ Evaluation of AI models

Why the need for strong evaluation?

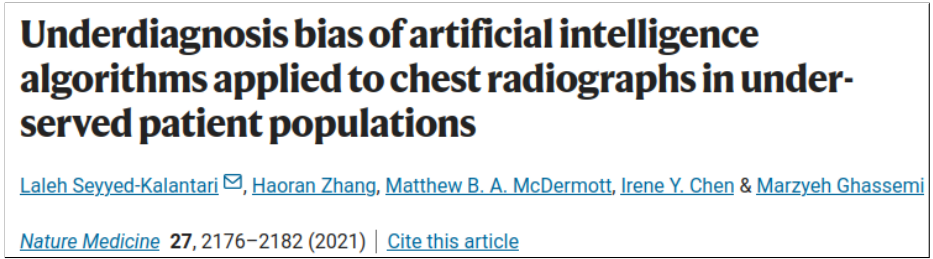
- Impact of misdiagnosis - healthcare is a critical sector
- Methods evolved from simpler regression-based methods to AI (Eg: foundation models)
- Voice is complex: variability, influence of environmental factors
- Disorders are complex

News • Coronavirus infection

AI shortcuts could lead to misdiagnosis of Covid-19

University of Washington researchers have discovered that AI models—like humans—have a tendency to look for shortcuts. In the case of AI-assisted disease detection, these shortcuts could lead to diagnostic errors if deployed in clinical settings.

Cases of AI issues in healthcare application (I)



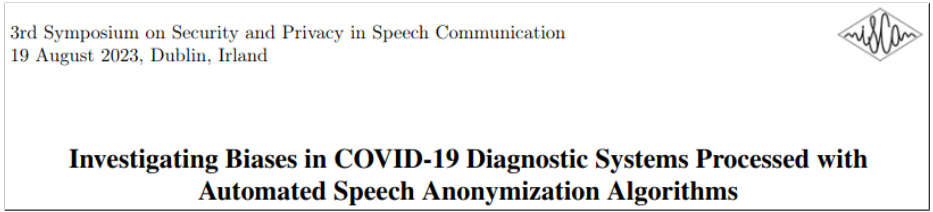
Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen & Marzyeh Ghassemi

Nature Medicine **27**, 2176–2182 (2021) | Cite this article

"We find that classifiers produced using state-of-the-art computer vision techniques consistently and selectively underdiagnosed under-served patient populations and that the underdiagnosis rate was higher for intersectional under-served subpopulations, for example, Hispanic female patients."

Cases of AI issues in healthcare application (II)



3rd Symposium on Security and Privacy in Speech Communication
19 August 2023, Dublin, Irland

Investigating Biases in COVID-19 Diagnostic Systems Processed with Automated Speech Anonymization Algorithms

"Our findings suggest the existence of diagnostic biases related to age and SNR of the recording, which become more prominent after anonymization."

Speech (vs) Voice for biomarker extraction?

Why the need for strong evaluation? - Regulation, compliance & trust

nature > npj digital medicine > perspectives > article

Perspective | Open access | Published: 06 September 2024

Navigating the EU AI Act: implications for regulated digital medical products

Mateo Aboy ✉, Timo Minssen & Effy Vayena

npj Digital Medicine **7**, Article number: 237 (2024) | Cite this article

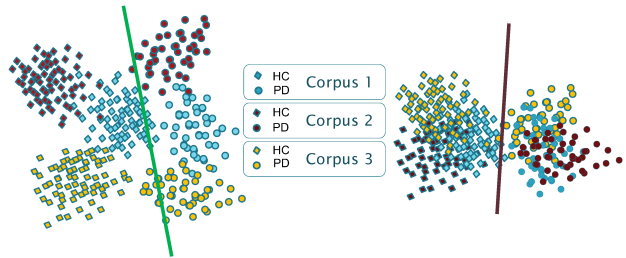
- **Regulatory Requirements:** Compliance with healthcare standards. AI -act
- **Patient Trust:** Transparency and reliability.
- **Bias identification**
- **Data Privacy**: Compliance with GDPR and HIPAA.

Standard summary metrics - classification

Binary classification task:

1. **Accuracy**: Fraction of samples correctly classified ($\frac{TP+TN}{TP+FP+TN+FN}$)
2. **Precision (PPV)**: Fraction of positively classified samples that are indeed positive.($\frac{TP}{TP+FP}$)
3. **NPV**: Fraction of negatively classified samples that are indeed negative.($\frac{TN}{TN+FN}$)

Ibarra EJ, Arias-Londoño JD, Zañartu M, Godino-Llorente JI. **Towards a corpus (and language)-independent screening of parkinson's disease from voice and speech through domain adaptation.** Bioengineering. 2023 Nov 15;10(11):1316.

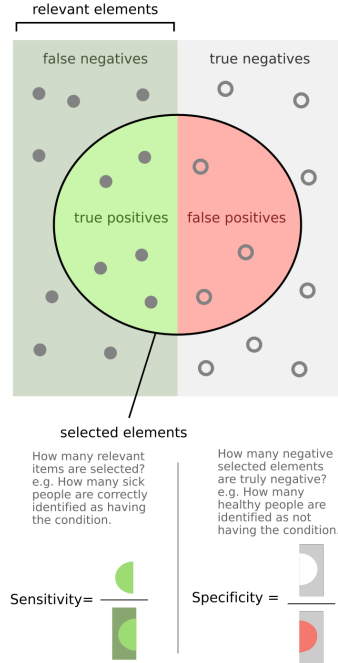


[Reference] Colliot O, editor. **Machine Learning for Brain Disorders.**

Standard summary metrics - classification

Binary classification task:

1. **Sensitivity**: Number of actually positive samples detected ($\frac{TP}{TP+FN}$)
2. **Specificity**: Number of actually negatived samples retrieved ($\frac{TN}{TN+FP}$)
3. **F1-score**: Harmonic mean of precision and sensitivity.



https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

Standard summary metrics - Confusion matrix

Confusion matrix: Summarizes results in a matrix form

Caution: Summary metrics should not be used in isolation.
Consider metrics both a) accounting for the model performance (Eg: sensitivity, specificity) b) accounting for the target population (Eg: PPV, NPV)

		Predicted condition	
	Total population = P + N	**Predicted positive**	**Predicted negative**
Actual condition	**Positive (P)** [a]	**True positive** (TP), hit[b]	**False negative** (FN), miss, underestimation
	Negative (N)[d]	**False positive** (FP), false alarm, overestimation	**True negative** (TN), correct rejection[e]

https://en.wikipedia.org/wiki/Confusion_matrix

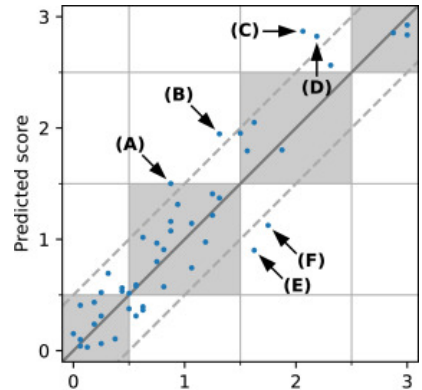
Standard summary metrics - Regression

For regression outcomes:

1. **R2 score**: Fraction of variance in the dependent variable explained by the estimate ($\hat{y}$), (wrt) the variance explained by the mean.
2. **MAE & RMSE**: Mean absolute error and root-mean squared error.
3. **Qualitative analysis of errors.**

Again: Metrics should be used in combination!

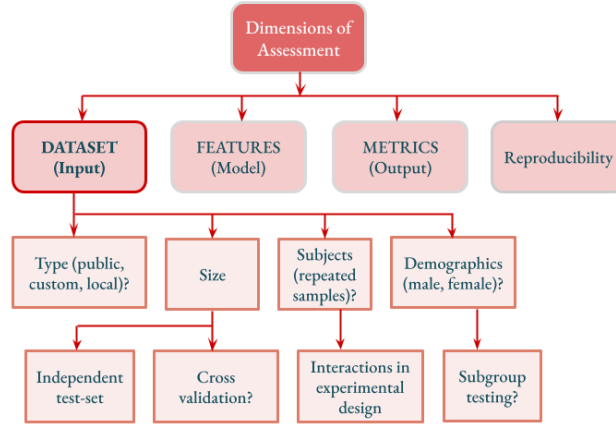
Hidaka S, Lee Y, Nakanishi M, Wakamiya K, Nakagawa T, Kaburagi T. **Automatic GRBAS scoring of pathological voices using deep learning and a small set of labelled voice data.** Journal of Voice. 2022 Nov 25.



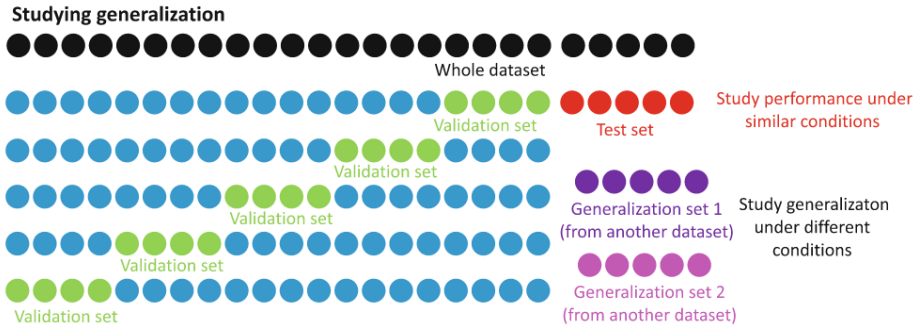
Evaluating the learning process

Design of the evaluation pipeline **matters!** – to ensure no data leakage

- Split the voice dataset: training, validation and test sets.
- IID (Independent and identically distributed) samples
- Control for age, gender, geography, ethnicity and other sensitive attributes!



Evaluating the learning process



[Reference] Varoquaux G, Colliot O. **Evaluating machine learning models and their diagnostic value. Machine learning for brain disorders.** 2023 Jul 23:601-30.

Evaluating the learning process

Other steps to avoid to prevent data leakage:

Box 3: Some Common Causes of Data Leakage

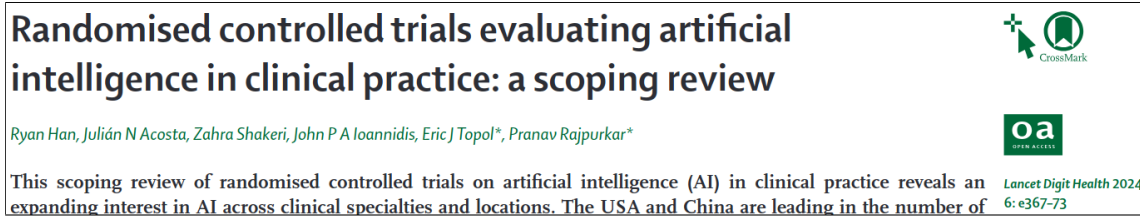
- Perform feature selection using the whole dataset.
- Perform dimensionality reduction using the whole dataset.
- Perform parameter selection using the whole dataset or the test set.
- Perform model or architecture search using the whole dataset or the test set.
- Report the performance obtained on the validation set that was used to decide when to stop training (in deep learning).
- For a given patient, put some of its visits in the training set and some in the validation set.
- For a given 3D medical image, put some 2D slices in the training set and some in the validation set.

Challenges in quality assessment

- **Data Quality** and Availability: Limited annotated medical datasets.
- **Variability** in medical data.

[Reference] Varoquaux G, Colliot O. **Evaluating machine learning models and their diagnostic value. Machine learning for brain disorders.** 2023 Jul 23:601-30.

Future discourse on Quality of voice-biomarker AI



Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review

Ryan Han, Julián N Acosta, Zahra Shakeri, John P A Ioannidis, Eric J Topol*, Pranav Rajpurkar*

This scoping review of randomised controlled trials on artificial intelligence (AI) in clinical practice reveals an expanding interest in AI across clinical specialties and locations. The USA and China are leading in the number of

Lancet Digit Health 2024; 6: e367–73

Before the large-scale deployment voice-biomarker AI model/software in hospitals and clinical, should consider an RCT.

Question: What will the form of such an RCT be? Any examples in within the medical voice-community (Eg: UEP)?

Summary

- **Key Takeaways**
  - Evaluation ensures reliable and safe voice-biomarker AI software.
  - Ask the right questions before deploying and adopting in healthcare.
- **Call to Action**
  - Collaborative efforts among AI researchers, clinicians-doctors and policy makers.

Questions and Discussions

Thank you for your attention!

sned@dtu.dk