Interspeech 2022
18-22 September 2022, Incheon, Korea

Are reported accuracies in the clinical speech machine learning literature overoptimistic?

Visar Berisha, Chelsea Krantsevich, Gabriela Stegmann, Shira Hahn, Julie Liss

Arizona State University
Aural Analytics

visar@asu.edu, chelsea.krantsevich@auralanalytics.com,
gabriela.stegmann@auralanalytics.com, shira.hahn@auralanalytics.com, jmliss@asu.edu

Abstract

Building clinical speech analytics models that will reliably translate in-clinic requires a realistic characterization of their performance. So, how well do we estimate the accuracy of published models in the literature? We evaluate the relationship between sample size and reported accuracy across 77 journal publications that use speech to classify between healthy controls and patients with dementia. The studies are combined across three meta-analyses that use the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol. The results show that reported accuracy declines as a function of increasing sample size, with small sample size studies yielding an overoptimistic estimate of the accuracy. For correctly trained models, this is unexpected as the ability of a machine learning model to predict group membership ought to remain the same or improve with additional training data. We posit that the overoptimism is the result of a combination of publication bias and overfitting and suggest mitigation strategies.

Index Terms: clinical speech analytics, robust machine learning, dementia, MCI, natural language processing

1. Introduction

There has been considerable interest recently in using the speech signal as a biomarker for different health conditions. The promise is that any neurological, mental health, or physical disturbances that impact the speech production process can be detected from patients' speech patterns. To that end, there has been significant interest in developing speech-based ML models for diagnosis, prognosis, and tracking of mental health [1], cognition [2, 3], motor disease [4, 5], etc.)

The methodology in this literature largely follows the traditional supervised learning paradigm. The authors begin with a labeled dataset consisting of speech samples and diagnostic or disease severity labels. From the speech, the authors either extract or learn a feature representation in service of building a machine learning model for predicting the label [1]. This paradigm has enjoyed considerable success in more traditional applications of speech analytics (e.g. automatic speech recognition (ASR)). However the clinical machine learning literature is markedly different since clinical databases are much smaller and more heterogeneous. For example, consider the studies for speech-based classification between healthy controls and patients with dementia in the meta-analyses in [6, 7, 8]. The largest samples are on the order of tens to hundreds of minutes of speech; this is orders of magnitude less than the data required to build consumer-grade ASR models [9, 10].

The lack of data challenges the development of robust supervised models that generalize in several ways. First, small databases are more susceptible to overfitting if the same data is used to select features and learn models [11]. Even when correctly using cross-validation, repeated use of small data for model/hyperparameter selection can lead to "overfitting to the dataset" [12]. Finally, small sample sizes lead to increased variability in the held-out error [13]. In other words, two different train-test splits by two different research groups working on the same data can yield wildly different estimates of model accuracy; the higher one is more likely to be published due to the file-drawer effect [14].

All of these are variations on the "selective inference" problem, where the ML designer makes a design choice (e.g. which features to use, which hypotheses class to consider, how to set a hyperparameter, how to split the data, etc.) prior to building the model and reporting an accuracy [15]. These choices are often made by reusing the same dataset multiple times explicitly, or implicitly by relying on the results of others that have used the data before. Theoretical analyses of the impact of adaptive reuse of the same data for inference shows that it leads to more optimistic estimates of a model's true accuracy [16, 15], especially for small sample sizes and high-dimensional data (e.g. Freedman's paradox is demonstration of this [17]).

From the perspective of a reader of a clinical speech paper, it's impossible to tell whether the reported accuracy is realistic or optimistic as they typically don't have insight into the string of choices made to arrive at the final model. But we can interrogate the literature as a whole to determine whether there is a trend between reported accuracy and sample size. If the effects of selective inference exist in the literature, we would expect that small sample size studies should report higher accuracies than large sample size studies. This trend has been observed in other clinical applications with high-dimensional inputs and small sample sizes [11, 18]. For correctly trained models, this is unexpected as ML model accuracy ought to remain the same or improve with additional training data.

We evaluate the relationship between sample size and reported accuracy in the clinical speech literature. We consider two applications: classification between healthy controls and patients with Alzheimer's disease (AD) and classification between healthy controls and patients with other forms of cognitive impairment (CI). This is an extension of our perspectives article in [19], where we first discussed the issues arising in estimating the accuracy of high-dimensional ML models. We extend that work by expanding on the set of studies considered to evaluate the negative relationship between sample size and reported accuracy, conducting a statistical analysis, and providing a more targeted discussion of how we can mitigate these issues. Our analysis finds a significant relationship between sample size and accuracy, with reported accuracy declining by 4% per unit increase in the $\log(\text{sample size})$ for both sets of studies.

10.21437/Interspeech.2022-691

2. Methods

We consider studies in three meta-analyses (MAs), each of which followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist. All three meta-analyses, described in turn, provided the data used in our analysis of the relationship between sample size and accuracy.

- **MA 1:** The study in [6] reviewed original research articles published between 2000 and 2019 covering multiple databases (ACM, IEEE, PsycINFO, PubMed, Embase, Web of Science). It identified 51 articles that satisfied pre-defined inclusion criteria. Tables 5 and 7 in [6] report on the sample size and reported accuracy of the algorithms presented in the studies under consideration.
- **MA 2:** The study in [7] reviewed articles published between 2013 and 2019 by searching in Web of Science, PubMed, and Ovid. It identified 33 eligible studies that satisfied pre-defined inclusion criteria, different from those in Study 1. Table 1 in [7] reports on the sample size and reported accuracy of the algorithms presented in the studies under consideration.
- **MA 3:** The study in [8] reviewed original research articles published between 2010 and 2020 covering PubMed, CINAHL, and PsychINFO. It identified 35 articles that satisfied the pre-defined inclusion criteria. Table 2 in [8] reports on the sample size and reported accuracy of the algorithms in the studies under consideration.

2.1. Data

The union of these three MAs results in a total of 77 unique studies that we individually analyze. Across these studies, we analyze the performance of binary classification models between a cognitively intact control group (Control) and a clinical group with cognitive impairment. We consider two types of clinical groups: those with an Alzheimer's disease (AD) diagnosis and those with other classifications of Cognitive Impairment (CI). The CI group may include participants with mild cognitive impairment (MCI), subjective cognitive impairment (SCI), functional memory disorders (FMD), neurodegenerative memory disorders (ND), and mild dementia. Several of the studies report on multiple classification algorithms between different pairs of clinical groups. The inclusion criteria for the classifiers we include in our analysis is as follows:

- Classification model performance must be provided in terms of model accuracy (% correct) or model accuracy must be estimable based on information provided in the paper. Accuracy was used as the performance metric of interest as it was most commonly reported across all studies.
- If multiple models (e.g., using different feature subsets or different classification algorithms) are provided in the papers, the model with the largest average performance is reported.
- If multiple modalities are considered in the paper, only the speech-based model results are reported.

The tables listed in the descriptions of the MAs above were used to obtain the sample size and reported accuracy for the classification algorithms reported in the studies. For any studies where dataset information was incompletely reported, sample size was unclear, accuracy was not provided, or individual models were incompletely explained in the meta-analyses, we downloaded the original paper for additional information.

The table listing the studies considered in the paper is provided as supplementary material online[1]. This table lists all classification models with information about sample size and reported accuracy. The *Study* column cites each study according to the numbers referenced in the meta-analysis from which it comes. The studies not included in our analysis are marked in red along with an explanation of why they were not included in the last column. The sample sizes for the control group and the clinical groups include all samples used for training and testing.

[1] http://vees.ar/IS22SpeechMLTable.pdf

Table 1: *Results of the statistical model evaluating the relationship between reported accuracy and sample size in the 3 MAs.*

Var.	Est.	Std Err.	t-val	p-val
Intercept	1.02	0.05	22.58	< 0.0001
log(SampleSize)	-.04	0.01	-3.61	< 0.0001
AnalysisType (Control vs CI)	-.09	0.02	-5.04	< 0.0001

2.2. Statistical modeling

The sample consisted of 59 classifiers across 3 meta-analyses which reported the sample sizes and prediction accuracies for two types of analyses: control vs AD and control vs CI. The goal was to estimate the relationship between the sample sizes and the reported accuracies; therefore a regression model was used with the reported accuracies as the outcome and the log of the sample size as the predictor. Given that the two types of analyses were expected to have different reported accuracies, where the Control vs AD group was expected to achieve higher accuracies than the Control vs CI group, the analysis type was included as a factor. Additionally, the interaction term between the log of the sample size and analysis and the inclusion of the meta-analysis groups as factors were evaluated.

Given that the sample size spanned multiple levels of magnitude, a log transformation was used. A linear regression was fit where the outcome was the reported accuracy, and the predictor was the log of the sample size. The analysis type (control vs AD and control vs CI) was included as a factor variable, and the two types of analyses were significantly different; therefore, this variable was retained in the model. The interaction term between the analysis type and log of the sample size was estimated; since it was not significantly different from 0, it was removed from the model. Finally, the three meta-analyses were included as factors in the model; since they were not significantly different from each other, they were removed. The final model was a linear regression with the accuracy as the outcome, and $\log(\text{sample size})$ and the analysis type as predictors.

3. Results

Table 1 shows the parameter estimates for the final model. The estimate for log(sample size) was -.04 (SE = .01, $t = -3.61$, $p < .0001$), indicating that for every unit increase in log(sample size), the accuracy decreased by .04. The estimate for the analysis type was -.09 (SE = .02, $t = -5.04$, $p < .0001$), indicating that the Control vs CI analyses had .09 lower accuracies than the Control vs AD analyses when holding sample size constant. A residual analysis was conducted using the Shapiro-Wilk test; it was not significant ($W = .98$, $p = 0.36$), indicating that residuals were not different from normal distribution. Fig. 1 shows the data with the regression lines and confidence intervals.

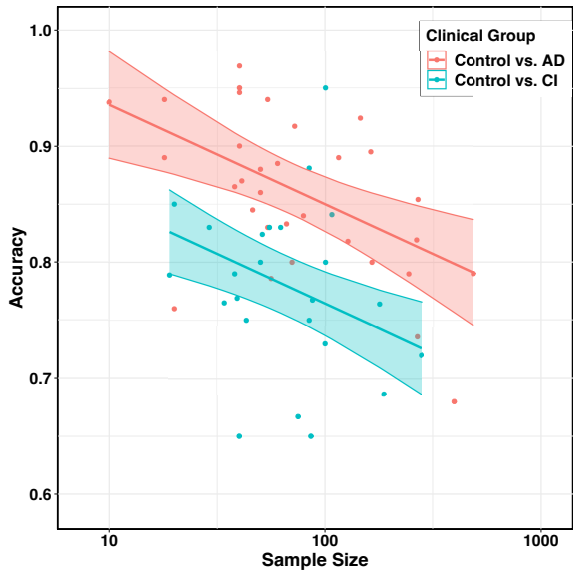

Figure 1: *The negative association between reported accuracy and sample size for two types of classification problems across three meta-analyses considered in the paper. AD = Alzheimer's Disease; CI = Other forms of cognitive impairment.*

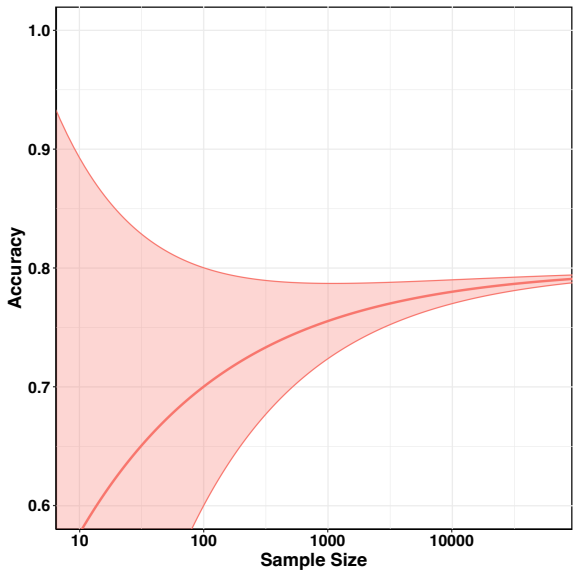

Figure 2: *A prototypical learning curve and confidence intervals.*

4. Discussion

4.1. Why is there a negative association between sample size and accuracy and why is it unusual?

The negative association between sample size and reported accuracy is seen in other machine learning studies in digital health. For example, an analysis of over 200 studies that propose ML models to predict disease status from neuroimaging similarly showed a negative association between the reported accuracy and the sample size across schizophrenia, MCI, Alzheimer's disease, major depressive disorder, and attention deficit hyperactivity disorder [18]. A similar analysis was found for classification models that predict Autism spectrum disorder using other modalities [11].

These findings are unexpected given what we know from theoretical and empirical investigations of learning curves in machine learning [20]. For properly trained ML models, as the sample size increases, the accuracy of the model increases monotonically according to a power law model, as in the example in Fig. 2. The figure shows a prototypical learning curve with average accuracy approaching a theoretical limit (0.8 in this example) and confidence intervals around this average. Yet we observe a decreasing trend in Fig. 1, contrary to the shape of the learning curve in Fig. 2. So, what explains the deviation from expectations?

We propose that model overfitting and publication bias are responsible for the negative association we observe. A common pre-processing step in the development of machine learning models is feature selection. For feature selection, it's known that using combined train and test data for feature selection and parameter tuning, followed by k-fold cross validation to estimate model accuracy results in optimistic estimates of model performance, especially for models trained with a small sample size [11]. More generally, repeated use of the same dataset over time - either explicitly or implicitly by relying on others that have analyzed the data previously - to improve models can lead to a similar bias [16]. The prototypical learning curve in Fig. 2 assumes that there is no information leakage between the training set used to develop the model and test set used to estimate the error at each sample size. This assumption no longer holds for either of the scenarios we describe above.

Another explanation for the counter-intuitive observation in Fig. 1 is based on publication bias. It's known that small sample size studies lead to more variable estimates of the generalization error [19]. We can observe this in our prototypical learning curve in Fig. 2 where the confidence interval around the average decreases as a function of sample size. Consider multiple research labs working with a small dataset. Even if each group follows best-practices for ML model design, the small sample size will lead to a highly variable estimate of model accuracy across the groups. However, the groups with higher estimates of the accuracy are more likely to publish their model [14]. Under this scenario, the readers of this literature will observe the declining upper envelope of the learning curve.

4.2. Why is this important?

Overestimation of true model accuracy in the published literature provides readers with unrealistic expectations of how well these models will work once deployed; this has important downstream consequences. Model accuracy is a convenient metric for succinctly describing how well an ML model works and it's commonly cited when the press reports on digital health. These articles raise the public's expectations for the capabilities of ML. If these expectations aren't met, confidence in the technology and scientific literature wanes [21].

The readership of the clinical speech machine learning literature is increasingly more interdisciplinary; stakeholders and decision makers may come away with unrealistic expectations of what clinical information can be gleaned from speech. This can prematurely lead to new companies, new investments, or large-scale implementations in hospital systems. There are already several examples of ML models in healthcare that were deployed after initial evaluation on small data sets revealed good performance, only to find out that the initial performance estimates were overoptimistic [22, 23].

4.3. What can we do about it?

To machine learn or not to machine learn? Cognition is a complex construct [24]; this means that the decision boundary between "healthy" and "cognitively impaired" is also likely complex. Small speech databases simply don't support the learning of such models. To compound the issue, small sample sizes make overfitting easier and increase variability, contributing to the overoptimism. For small sample sizes (e.g in the tens), algorithm developers should forgo supervised learning on broad diagnostic labels for more manageable problems. For example, there is a need to develop robust speech/language *features* known to change with cognitive impairment (e.g. circumlocution). The ground truth for these models can be attained directly from the audio and/or transcript, making it much easier to obtain than clinical labels. Furthermore, data from healthy individuals, which is much easier to obtain, can be used to augment clinical databases for development of these features, for more accurate estimates of performance, and for developing normative ranges.
A new pipeline for clinical speech analytics: The challenge in discovering clinically-relevant features, and accurately assessing the performance of models trained with those features, is that speech signals are high-dimensional, the underlying patterns are complex, and datasets are small. In contrast to applications like ASR or speaker recognition, there is no standard representation used for clinical applications. As a result, researchers rely on existing open-source feature extraction frameworks re-purposed from other applications [25]. Efforts focused on clinically-relevant speech representations would go a long way towards developing a new "pipeline" tailored to specific clinical applications. Smaller-dimensional, clinically relevant, and individually validated features built on a clinical foundation are less likely to lead to overfit or variable models.

The Levelt model of speech production, which models speech as the output of a 3-stage process (Conceptualization $\rightarrow$ Formulation $\rightarrow$ Articulation), provides a good conceptual framework for what to measure [26]; several researchers have adopted it in their clinical speech analytics models already [1, 27]. However, what's missing is a "measurement model" (e.g. how should we compute the features) of speech that maps to this framework. For example, there are hundreds of ways in which researchers could characterize the Articulation stage - acoustic models based on MFCCs, mel-spectra, etc; or hundreds of ways to characterize the Formulation stage - language models like BERT, ELMO, etc. So what's the right way?

While there is not yet a consensus on the answer to this question, we note that there are several desiderata when designing new clinically-important features:

- Low-dimensional, well-characterized features: Small sample sizes will continue to be a reality of clinical applications in the near future. Models built on low-dimensional, well-understood features are more likely to generalize [19].
- Amplify behaviors of interest: Certain characteristics of disease - especially in the early stages - are not readily apparent in passively-collected speech. New feature representations should be paired with maximum-performance tasks (e.g. diadochokinetic elicitation) for assessment of speech [28].
- Interpretable and individually-validated features: Perceptual analysis of clinical speech has a long and rich history that dates back to the 1800's [29]. We should align our representation of speech with this literature so that models can be validated relative to the existing clinical knowledge base. Speech features that are individually validated are more likely to generalize across populations and clinical conditions.

These desiderata lead to new challenges for the clinical speech analytics community to address. This will not preclude use of state-of-the-art models in the development of clinical speech analytics, but will challenge us to use such models creatively when there is a dearth of clinical data. For example, deep learning models *trained using only healthy speech* can be validated and then used to assess the fidelity of consonant-vowel transitions or hypernasality in clinical populations [30, 31, 32]. As the scale of clinical data grows, so can the complexity of the models brought to bear on the problem.
Changing how we think about model validation: The current approach to model validation tends to be superficial; most papers only report on "model accuracy" estimated using cross-validation. As we see in this work, this metric can be easily mis-estimated, especially in the small sample-size regime. Supervised models, especially those based on deep learning, are susceptible to perturbations in the input [33]. That is, small changes in the input features (even if they are imperceptible) can change the output of a pre-trained network. This is problematic for speech features since there is considerable evidence that commonly used features for clinical applications are highly variable day-to-day [25]. So, accuracy estimates of models built using features from day one no longer hold on day two. Open-source models that can be evaluated by other teams for reproducibility on different databases can help mitigate this problem.

A recently-proposed framework for validation of biometric monitoring technologies provides a way forward [34]. The framework proposes three levels of evaluation: Verification of hardware, analytical Validation, and clinical Validation (V3). Verification evaluates the fidelity of the raw sample data produced by the sensor technology used to acquire it in the environment in which it is to be used. Analytical validation evaluates the performance of the algorithms that process the raw sensor data to produce physiological or behavioral metrics. Clinical validation demonstrates that models track with a clinically-relevant ground truth or a clinically accepted measure.

Clinical speech analytics researchers tend to skip verification and analytical validation steps in favor of aiming to directly evaluate the clinical validity of their model. However, we posit that this is limiting as the intermediate validation steps can help better characterize the underlying features. Note that the analytical validation step is agnostic to the underlying clinical condition; rather, it focuses on estimating a behavioral measure of interest. For example, speaking rate and articulatory precision are impacted across many conditions [4, 35]. Having a tool that is validated at the behavior-measurement level will allow researchers to use it as a feature across many conditions for model-building. Applying this level of validation to the new pipeline proposed in the previous section will result in a reusable framework for discovering and translating speech analytics across the clinical landscape.

5. Conclusions

Our results show that reported accuracy in the clinical speech analytics literature declines as a function of increasing sample size, with small sample size studies yielding an overoptimistic estimate of model accuracy. This is inconsistent with the theoretical and empirical observations of learning curves in machine learning. We explain the overoptimism as a result of publication bias and overfitting and suggest several mitigation strategies.

6. Acknowledgements

This work is funded in part by NIH - NIDCD R01 DC006859.

7. References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[2] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[3] G. Stegmann, S. Hahn, S. Bhandari, K. Kawabata, J. Shefner, C. J. Duncan, J. Liss, V. Berisha, and K. Mueller, "Automated semantic relevance as an indicator of cognitive decline: Out-of-sample validation on a large-scale longitudinal dataset," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 14, no. 1, p. e12294, 2022.

[4] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. Rutkove, K. Shelton, C. J. Duncan, and V. Berisha, "Early detection and tracking of bulbar changes in als via frequent and remote speech analysis," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–5, 2020.

[5] A. Zhan, S. Mohan, C. Tarolli, R. B. Schneider, J. L. Adams, S. Sharma, M. J. Elson, K. L. Spear, A. M. Glidden, M. A. Little *et al.*, "Using smartphones and machine learning to quantify parkinson disease severity: the mobile parkinson disease score," *JAMA neurology*, vol. 75, no. 7, pp. 876–880, 2018.

[6] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, 2020.

[7] U. Petti, S. Baker, and A. Korhonen, "A systematic literature review of automatic alzheimer's disease detection from speech and language," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1784–1797, 2020.

[8] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article," *Frontiers in Psychology*, vol. 12, p. 645, 2021.

[9] S. H. K. Parthasarathi and N. Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6670–6674.

[10] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," *CoRR*, vol. abs/1801.00841, 2018. [Online]. Available: http://arxiv.org/abs/1801.00841

[11] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PloS one*, vol. 14, no. 11, p. e0224365, 2019.

[12] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015.

[13] C. Flint, M. Cearns, N. Opel, R. Redlich, D. Mehler, D. Emden, N. R. Winter, R. Leenings, S. B. Eickhoff, T. Kircher *et al.*, "Systematic misestimation of machine learning performance in neuroimaging studies of depression," *Neuropsychopharmacology*, vol. 46, no. 8, pp. 1510–1517, 2021.

[14] R. Rosenthal, "The file drawer problem and tolerance for null results." *Psychological bulletin*, vol. 86, no. 3, p. 638, 1979.

[15] J. Taylor and R. J. Tibshirani, "Statistical learning and selective inference," *Proceedings of the National Academy of Sciences*, vol. 112, no. 25, pp. 7629–7634, 2015.

[16] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[17] D. Freedman, "A note on screening regression equations," *the american statistician*, vol. 37, no. 2, pp. 152–155, 1983.

[18] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *Neuroimage*, vol. 145, pp. 137–165, 2017.

[19] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, and J. Liss, "Digital medicine and the curse of dimensionality," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–8, 2021.

[20] T. Viering and M. Loog, "The shape of learning curves: a review," *arXiv preprint arXiv:2103.10948*, 2021.

[21] S. Cave, C. Craig, K. Dihal, S. Dillon, J. Montgomery, B. Singler, and L. Taylor, "Portrayals and perceptions of ai and why they matter," 2018.

[22] A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza *et al.*, "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients," *JAMA Internal Medicine*, vol. 181, no. 8, pp. 1065–1070, 2021.

[23] C. Ross and I. Swetlitz, "IBM's Watson supercomputer recommended 'unsafe and incorrect'cancer treatments, internal documents show," *Stat*, vol. 25, 2018.

[24] A. W. Ellis and A. W. Young, *Human cognitive neuropsychology: A textbook with readings*. Psychology Press, 2013.

[25] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. B. Rutkove, K. Kawabata, S. Bhandari, K. Shelton, C. J. Duncan, and V. Berisha, "Repeatability of commonly used speech and language features for clinical applications," *Digital biomarkers*, vol. 4, no. 3, pp. 109–122, 2020.

[26] W. J. Levelt, *Speaking: From intention to articulation*. MIT press, 1993.

[27] R. Voleti, J. M. Liss, and V. Berisha, "A review of automated speech and language features for assessment of cognitive and thought disorders," *IEEE journal of selected topics in signal processing*, vol. 14, no. 2, pp. 282–298, 2019.

[28] R. D. Kent, J. F. Kent, and J. C. Rosenbek, "Maximum performance tests of speech production," *Journal of speech and hearing disorders*, vol. 52, no. 4, pp. 367–387, 1987.

[29] P. Broca *et al.*, "Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech)," *Bulletin de la Société Anatomique*, vol. 6, pp. 330–357, 1861.

[30] V. Mathad, J. Liss, K. Chapman, N. Scherer, and V. Berisha, "Consonant-vowel transition models based on deep learning for objective evaluation of articulation," *arXiv preprint arXiv:2203.10054*, 2022.

[31] M. Saxon, J. Liss, and V. Berisha, "Objective measures of plosive nasalization in hypernasal speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6520–6524.

[32] V. C. Mathad, N. Scherer, K. Chapman, J. M. Liss, and V. Berisha, "A deep learning algorithm for objective assessment of hypernasality in children with cleft palate," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 10, pp. 2986–2996, 2021.

[33] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.

[34] J. C. Goldsack, A. Coravos, J. P. Bakker, B. Bent, A. V. Dowling, C. Fitzer-Attas, A. Godfrey, J. G. Godino, N. Gujar, E. Izmailova *et al.*, "Verification, analytical validation, and clinical validation (v3): the foundation of determining fit-for-purpose for biometric monitoring technologies (biomets)," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–15, 2020.

[35] V. Berisha, J. Liss, T. Huston, A. Wisler, Y. Jiao, and J. Eig, "Float like a butterfly sting like a bee: Changes in speech preceded parkinsonism diagnosis for muhammad ali." in *INTERSPEECH*, 2017, pp. 1809–1813.