Safeguarding against spurious AI-based predictions: The case of automated verbal memory assessment

Chelsea Chandler and **Peter W. Foltz**
Department of Computer Science &
Institute of Cognitive Science,
University of Colorado Boulder
chelsea.chandler@colorado.edu
peter.foltz@colorado.edu

Alex S. Cohen
Department of Psychology &
Center for Computation & Technology,
Louisiana State University
acohen@lsu.edu

Terje B. Holmlund
Department of Clinical Medicine,
University of Tromsø -
The Arctic University of Norway
terje.holmlund@uit.no

Brita Elvevåg
Department of Clinical Medicine,
University of Tromsø -
The Arctic University of Norway &
Norwegian Centre for eHealth Research
brita@elvevaag.net

Abstract

A growing amount of psychiatric research incorporates machine learning and natural language processing methods, however findings have yet to be translated into actual clinical decision support systems. Many of these studies are based on relatively small datasets in homogeneous populations, which has the associated risk that the models may not perform adequately on new data in real clinical practice. The nature of serious mental illness is that it is hard to define, hard to capture, and requires frequent monitoring, which leads to imperfect data where attribute and class noise are common. With the goal of an effective AI-mediated clinical decision support system, there must be computational safeguards placed on the models used in order to avoid spurious predictions and thus allow humans to review data in the settings where models are unstable or bound not to generalize. This paper describes two approaches to implementing safeguards: (1) the determination of cases in which models are unstable by means of attribute and class based outlier detection and (2) finding the extent to which models show inductive bias. These safeguards are illustrated in the automated scoring of a story recall task via natural language processing methods. With the integration of human-in-the-loop machine learning in the clinical implementation process, incorporating safeguards such as these into the models will offer patients increased protection from spurious predictions.

1 Introduction

Artificial intelligence (AI)-based systems that incorporate language and behavioral data hold promise of increasing sensitivity, equity, and access in the assessment and treatment of mental illness through the use of remote and continuous monitoring via clinical decision support systems. This is due to the fact that the pattern and content of language, as well as additional measures of behavior, such as timing and neuropsychological task scores, provide rich information that can be traced back to an individuals' overall mental state.

In order to demonstrate clinical translational value there are numerous risks and factors that are necessary to consider. First, it is important to collect data from large samples of the population across differing ages, cultures, genders, clinical conditions, and stages of disorder. Second, it is critical to create models that are explainable, transparent, and generalizable (Chandler et al., 2020b) in order to nurture trust from both patients and clinicians. And finally - the area that this paper will address - it is necessary to add safeguards to models such that they are capable of flagging cases that show attribute noise (i.e., abnormalities in feature values) or class noise (i.e., erroneous or missing class labels), and of determining the extent to which models will generalize to unseen data. These safeguards will enable a human-in-the-loop system where humans are required to review data abnormalities.

AI is used in a wide range of applications within mental health, notably within clinical research settings where data are used to aid in understanding the nature of diagnoses and to improve diagnostic accuracy (for reviews see Shatte et al., 2019; Su et al., 2020; Thieme et al., 2020), as well as in making complex and potentially lifesaving de-

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology
To appear online during NAACL-HLT 2021, June 11, 2021.

cisions (e.g., in suicidology - for review see Cox et al., 2020). Acoustic measurements of speech have been analyzed in automated applications for detecting Mild Cognitive Impairment and dementia (Roark et al., 2011; König et al., 2015), as well as serious mental illness (Cohen et al., 2019) and depression (McGinnis et al., 2019).

In the domain of techniques that specifically leverage natural language processing (NLP), there are a growing number of reports of using these methods on social media data, notably to data mine publicly shared written reports of mood on platforms such as Twitter and Reddit (Zirikly et al., 2019; Peng et al., 2019; Wu et al., 2012). There is also a growing interest in using such methods on electronic medical records to assist in the extraction of diagnostic information or to enhance understanding of medical conditions (Ryu et al., 2016; Wang et al., 2012; Metzger et al., 2017). A broad range of NLP metrics such as incoherence and tangentiality have been used to automatically assess the clinical state of patients with schizophrenia (Elvevåg et al., 2007) and predict the risk of psychosis onset (Bedi et al., 2015; Rosenstein et al., 2015; Corcoran et al., 2018). Deep language models and NLP feature-based models have also been shown to differentiate the language of healthy controls from those diagnosed with Mild Cognitive Impairment or dementia (Orimaye et al., 2018; Eyigoz et al., 2020).

There is clear evidence that the clinical data used in AI-based research applications hold predictive power in detection and diagnosis, prognosis, support and treatment, and as a second opinion measurement for illness severity, but it is unclear about the degree to which these models will be stable on new data. Many psychiatric studies that harness AI tend to do so on relatively small datasets (i.e., 10-100 participants) in fairly homogeneous populations (e.g., the WEIRD (Western, Educated, Industrialized, Rich, and Democratic) phenomenon - Henrich et al., 2010; and the predominance of male participants in psychiatric research studies - Longenecker et al., 2010). These shortcomings may lead to insufficient accuracy on unseen data retrieved from different experimental settings (e.g., in a lab vs. remote; prompted free speech vs. natural; as a component of a larger testing battery vs. on its own), populations (e.g., southern vs. northern; different English speaking countries; monolingual vs. multilingual participants), and clinical states (e.g., hallucinating vs. not hallucinating). One must keep in mind that in small datasets, spurious features may not be generalizable to a larger population, especially if they are not of any apparent clinical relevance (Chandler et al., 2020b; Whelan and Garavan, 2014). While these research experiments are noteworthy, they must be re-evaluated on larger and more diverse sets of participants to test for robustness and generalizability.

Incorrect or ill-advised decisions and predictions in psychiatry can be dangerous and life altering for patients, and the difficulty in decision making is further confounded by the very short time frame in which changes in mental state occur and the associated clinical decisions must be made. Thus, we must build systems that have the ability to instantaneously flag data abnormalities - both in the research phase and when translated into real clinical use - and pass these cases on for human review. Furthermore, rather than selecting a preferred machine learning model based on metrics such as accuracy, sensitivity, or correlation as is common in AI and NLP applications, we must seek to understand the underlying mechanisms and the context in which they will be used (Ethayarajh and Jurafsky, 2020; Hand, 2006).

Researchers in machine learning have proposed assessing models with stability metrics which define ways to quantify and compare the stability of results rather than simply focusing on the aforementioned metrics (Turney, 1995; Lange et al., 2002). Specifically, Zhu and Wu (2004) differentiated data-based noise and outliers into class noise and attribute noise, and advocated for analyzing their effects on machine learning models separately. Uncertainty estimation, as well as in- and out-of-distribution error detection has been critically important in the use of AI in a wide range of applications such as self driving cars (Mohseni et al., 2020), general medicine (Kompa et al., 2021), education (Foltz et al., 2013), and in many other domains.

In this paper we illustrate an example of NLP and machine learning methods applied to the automated scoring of a story recall task, a core component of psychiatric neuropsychological assessments. We focus on two approaches to safeguarding such a model: 1) the detection of attribute and class noise that can affect the predictions of a model and 2) the evaluation of the extent to which the model may or may not generalize to unseen data. We first applied methods to determine where noise exists with an outlier detection algorithm and data visualization. For the issue of model generalizability, we studied the effect of dataset size on the results, and we illustrate how such results change as we randomly remove portions of our training data. Additionally, we show the results of this particular story recall model applied to a new collection of data. We advocate that these computational safeguards, which have major implications in regard to their use in human-in-the-loop clinical support systems, must be placed on each machine learning model that is developed to automate or assist in clinical assessments.

2 Experimental overview

2.1 The *d*MSE

The data in the present work were collected from a mobile phone application (the *delta* Mental Status Examination, henceforth called *d*MSE) designed to assess patient state via various neuropsychological assessments, with many relying on patient language (Chandler et al., 2020a; Cohen et al., 2019; Holmlund et al., 2019; Holmlund et al., 2020). A total of 12 behavioral assessment tasks were employed to specifically assess the language, cognition, motor skill, and mental state of patients - areas where assessment is critical in those with serious mental illness - and integrated into the *d*MSE smart device application. The behavioral assessment tasks were similar to standardly employed neuropsychological tests (for an overview of neuropsychological testing, see Lezak et al., 2012), but adapted such that they could be remotely and frequently self-administered with variations of each task presented over time (Chandler et al., 2020a; Holmlund et al., 2019). As an automated measurement tool that can be used remotely, frequently and self-administered, this approach has the potential to enable greater access to mental health services. It permits patients to be monitored longitudinally outside of clinical institutions and can alert clinicians to critical changes in mental states, thereby providing greater availability to assistance, regardless of age, gender, ethnicity, location, or socioeconomic status.

The data comprised N = 25 patients and N = 79 presumed healthy undergraduate students from Louisiana State University who all provided informed written consent. These participants completed N = 118 and N = 226 sessions (i.e., one completion of the full battery of tasks in a single use of the application) with the *d*MSE, with an average of 4.72 (stdev = 1.14) and 2.90 (stdev = 0.90) per person, respectively. The patients were severely mentally ill outpatients on the psychosis spectrum. Two-thirds of the patients met the criteria for schizophrenia (N = 16), and the remaining met the criteria for major depressive disorder (N = 8) and bipolar disorder (N = 1). This study was approved by the Louisiana State University Institutional Review Board (#3618) and participants provided their informed written consent before participation. The application was designed specifically for use in remote settings, such as rural Louisiana and Northern Norway, where access to in-person clinical support can be quite difficult.

2.2 The story recall regression model

The machine learning model we use to illustrate safeguarding techniques automatically scored a variant of the immediate and delayed Logical Memory story recall task (of the Wechsler Memory test; Wechsler, 1997) that was employed in the *d*MSE. The story recall task is critical in neuropsychological assessment as memory function is of core interest in the evaluation of many neurodevelopmental, neurodegenerative and neuropsychiatric conditions, as well as in brain injuries (Baddeley and Wilson, 2002). Further, it is of enormous interest in mental illness research because of its value as a critical endophenotype (Cirillo and Seidman, 2003), as well as the fact that the process of recollecting has similarities to what is required by patients when their medical history is taken.

In our version of this task, a participant listens to a short story of on average 74 words (min = 62, max = 87) and then is asked to retell it both immediately and after a delay of 30 minutes in as much detail as possible, thus following the same format as the traditional Wechsler version. Stories were either narrative or instructional. The narrative stories contain two characters, a setting, an action that caused a problem, and a resolution. The instructional passages described how to accomplish some sort of goal, such as how to assemble a skateboard or how to clean a fish bowl. This *d*MSE story recall task was developed such that there could be many different versions capable of being scored with automated NLP methods (e.g., Chandler et al., 2021, Holmlund et al., 2020) rather than traditional rubric-based methods.

Three trained human raters with clinical experience assigned scores to the recall transcriptions based on the quality and amount of details (e.g., characters, events, dates, descriptors, feelings) recalled. The rubric was on a scale from 1 to 6, with 1 indicating no details were recalled, and 6 indicating all major and almost all minor details were recalled. Each participant completed one immediate narrative recall, one immediate instructional recall, and one delayed narrative recall per session. After the removal of responses with no words, the dataset contained N = 846 samples (N = 285 immediate narrative, N = 285 immediate instructional, and N = 276 delayed narrative).

A ridge regression model was created to predict the rating a trained professional would assign to a story recall. The model was trained on (1) the number of word types (i.e., unique words) in the recall, (2) the number of common word types between the original story and the recall, and (3) the BERTScore (Zhang et al., 2020) between the original story and the retell (the model was created in the same manner as that of Chandler et al., 2019 besides a change in the last feature from the word mover's distance to BERTScore). BERTScore is a similarity metric that was created to produce a score of how close a machine generated translation is to the gold standard(s) of some piece of text. Specifically, it creates a matrix of BERT (Devlin et al., 2019) cosine distances between words in one text to words in another. Alignment between words in both texts is produced greedily with the maximum cosine distance for each word in one text to another in the reference. All distances are averaged and inverse document frequency weightings are optionally incorporated.

The ridge regression model was trained and tested using 10-fold cross-validation and controlled such that sessions from the same participant did not occur simultaneously in both the train and test sets. The rating prediction model resulted in an average Pearson r correlation with human ratings of r = 0.91. These results indicate that we can automatically derive a range of semantic and surface level features from spoken recalls, and that these features can be harnessed to accurately predict the ratings of expert humans.

3 Effects of attribute and class noise

We begin our analysis of computational safeguards by discussing the determination of attribute and class noise in the context of model stability. Model stability analysis allows us to establish how unusual variations in input data will affect the output of the model. Put simply, we wish to find where in the feature space models may be the most unstable. We illustrate an approach that will allow researchers to detect attribute and class noise in data that could be due to construct-irrelevance or errors in assumptions.

Attribute 1: Number word types	Attribute 2: Number common word types	Attribute 3: BERTScore	Class rating
4	2	0.91	1
3	'x'	0.70	1
36	25	0.93	6
4	3	0.71	***6***

Table 1: Hypothetical subset of story recall data showing attribute noise (underlined) and class noise (bold and italicized). First, 0.91 in the first row constitutes potential attribute noise as the average BERTScore for examples with a rating of 1 is 0.80 (stdev = 0.05), and furthermore the average BERTScore for examples with 4 word types and 2 common words is 0.79 (stdev = 0.04) and 0.80 (stdev = 0.05), respectively. Thus, it is far out of the expected distribution. Second, 'x' in the second row constitutes attribute noise because this attribute expects numbers and there is a string in its place. Thus, it is erroneous. The class label of 6 in the last row constitutes class noise as the distribution of the feature values resembles a much lower recall score.

Specifically, attribute noise is where values of individual attributes do not make sense; whether they are erroneous or missing. Class noise is where a label does not make sense given the distribution of the features for other data with the same label; whether it is mislabeled or contradictory. In order to make the notions of attribute noise and class noise concrete, see Table 1 for a hypothetical distribution of the story recall data with an emphasis on what could potentially constitute both types of noise. In this section, we explore instability that could be due to outliers in training data, disagreement between features, or incorrect assumptions of the data.

Our first outlier analysis was based on research-stage settings where we have access to both attribute values and class labels. While this exact approach may not always be feasible in the eventual clinical application stage (since there are not always ground truth class labels available), the approach itself can nonetheless be harnessed in

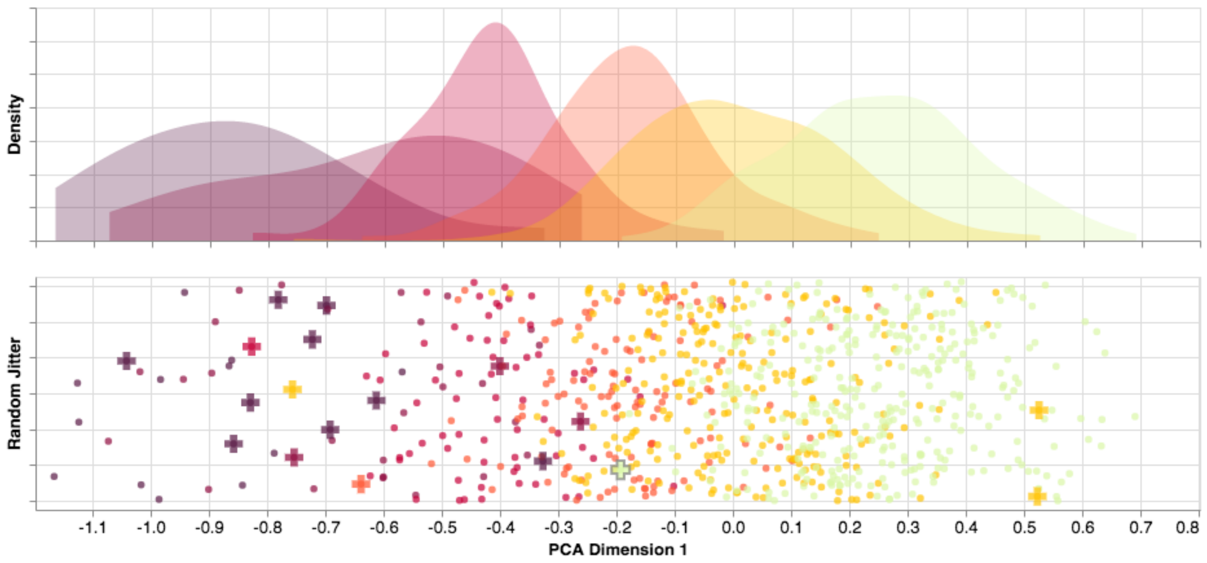

Figure 1: Distribution of the first dimension of Principal Component Analysis (PCA) of the 3 features of the story recall data separated by rating. The darker colored peak on the left represents the lowest rating (1 point) which increases by one point per peak to the lighter colored peak on the right hand side (6 points). Outliers found with the Isolation Forest algorithm are shown with a cross and the color of the cross represents the human rating given to that example.

the same manner but with the omission of ratings, classes, or labels. Here, we discovered outliers using the Isolation Forest algorithm (Liu et al., 2018). Most outlier detection algorithms first find the normal region of data and subsequently define anything outside of this defined region to be an outlier. The Isolation Forest algorithm, on the other hand, discovers minority data points that have attribute values that differ from those of the usual instances. Specifically, the algorithm isolates examples by selecting an attribute at random and then selecting a random split value between the maximum and minimum values of the selected feature. Anomalous examples will have shorter paths from the root to the leaves in their isolation trees than the normal examples since they need fewer partitions to be isolated. This algorithm is well-suited for high dimensional datasets and has proven to be an effective way of detecting outliers and anomalies (Ding and Fei, 2013). Furthermore, it works especially well for behavioral data as “normal” regions tend to be more variable than in other domains.

The current outlier analysis was specifically based on the number of types (i.e., unique words), the number of common types between the original story and the recall, the BERTScore between the original story and the recall, and the human rating given to the recall. Figure 1 shows the results of applying the Isolation Forest algorithm to the story recall data. It is shown that 18 outliers were detected. Such instances would be flagged for human review, where researchers can determine if attribute or class noise is present and either fix the erroneous values or exclude them from the modeling in the case that the examples are entirely invalid. When the approach is used in clinical settings to flag attribute noise, clinicians can review the raw data and make determinations for themselves rather than relying on a machine prediction.

Out of the 18 examples flagged by the Isolation Forest algorithm, 9 were determined to be invalid responses (i.e., participants stating that they simply do not remember or responses that are insufficient for data analysis) and 9 were valid responses with either sparse amounts of language or large amounts of language but poor performance. The average absolute error on the outliers was 1.34 (stdev = 0.80); the valid response outliers generated a higher absolute error (average = 1.63, stdev = 0.91) than the invalid response outliers (average = 1.05, stdev = 0.63). The performance of the model on outlier data is far lower than the models overall performance.

As the contamination threshold of the Isolation Forest algorithm is increased (i.e., the criteria for an outlier is relaxed), additional responses are chosen that mirror the behavior of these 18. This is a parameter that would need to be tuned such that

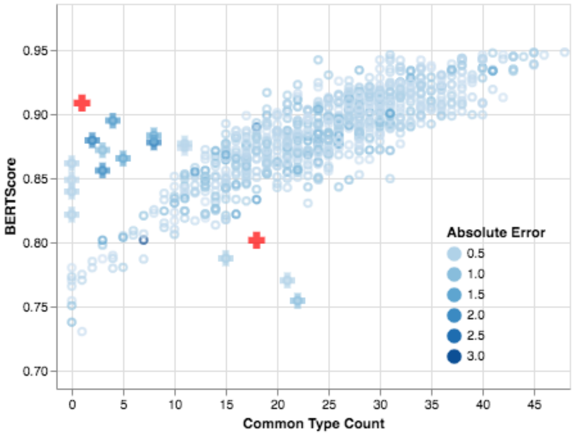

Figure 2: Scatter plot depicting the relationship between the number of common word types and the BERTScore of each example. The color represents the absolute error between the model rating and the human rating in each instance. Cross symbols indicate attribute noise (with two specific examples colored red and detailed at a high level in the text).

all true outliers are detected yet it does not extend into the normal data range. Furthermore, this parameter will need to be learned by investigating the true distribution of these phenomena and will depend on the application. Interestingly, the model performance did not change with the removal of these outliers. As approximately 2% of the data was flagged in this experiment, the model behaved indifferently to their exclusion. The exclusion of extremes (which help the performance of the model) combined with noise (which harm the performance of the model) potentially balanced out the effects of both. This Isolation Forest analysis can be performed on the same data without ratings in the eventual clinical stage to find attribute noise and extremes. We also present an analysis of features alone that can be done in any stage of the modeling process.

A basic noise detection approach that can be used at any stage of the modeling process is to simply find the examples with low attribute agreement (assuming that the attributes are collinear). Figure 2 depicts the distribution of two of the most predictive features of the story recall rating prediction model (the number of common word types and the BERTScore between the original story and the recall). There is a steady agreement between the two features, with some outliers (marked with crosses) outside of the diagonal where the features do not agree. The color of the circles represent how far off the model rating was from the human rating. Two examples with exceptionally high error (~2.5-3.0) are identified in red. The bottom-most red example is a response with a mixture of correct and incorrect (random) details, as well as incoherent language. The top-most red example is a response with a high BERTScore even though only a recitation of the title of the story was spoken. This high disagreement between features in turn uncovered a faulty feature score potentially due to flawed weighting parameters in the BERTScore model. We have shown that examples located off of the diagonal in plots such as these should be passed on for human evaluation as disagreement in two objective collinear attributes of story recall may raise concern.

Finding these outliers is critical because if a model has not been exposed to certain combinations of features or labels in its training set, then we cannot assume that it will produce accurate predictions in such settings. Outliers are important to detect both in the research stage in order to update or exclude certain examples from affecting the model in a negative manner and in the clinical setting so that spurious decisions are not made on abnormal data.

4 Effects of model generalizability

As previously stated, one of the most critical safeguards to spurious AI-based predictions is using large, diverse, and representative data (Cirillo et al., 2020), but this is not always possible. When using human behavioral data in machine learning algorithms, researchers inadvertently make the assumption that there is one canonical representation of specific groups of humans (i.e., those with serious mental illness), but this is simply not true. Those with psychiatric disorders exhibit extremely diverse symptoms and behaviors. Human behavior displays patterns indicative of a chaotic system (Paulus and Braff, 2003; Guess and Sailor, 1993), which holds true for behavior within one person as well as behavior within a group. To approach the topic of generalizable data, we first explored whether choosing different subsets within a training dataset would affect the output of the resulting model and whether there are spurious results when using smaller subsets.

The story recall regression model was trained on N = 846 samples, a large size relative to clinical experiments in the mental health domain. We used stratified sampling to create smaller subsets of the data that retain the proportions of each rating

Percent of data (N)	Average model rating correlation (stdev)	Average BERTScore correlation (stdev)	Average common types correlation (stdev)
100% (846)	0.91	0.86	0.82
75% (634)	0.91 (0.01)	0.86 (0.01)	0.82 (0.01)
50% (423)	0.90 (0.01)	0.82 (0.01)	0.79 (0.01)
25% (212)	0.88 (0.02)	0.81 (0.03)	0.79 (0.02)

Table 2: The change in the average and standard deviation (stdev) of the correlations between (1) the human rating and the model rating, (2) the human rating and BERTScore, and (3) the human rating and common word types as smaller subsets of the data are randomly chosen in a stratified manner for training and testing. The first column displays the percent of data and the number of data points used in each data reduction setting.

and tested how the model behaved on these smaller subsets. Table 2 depicts the changing accuracy of the model and correlations of features to human ratings when these smaller subsets of the data were used for training and testing. We found the average correlation over a 10-fold cross-validation of the sampled subsets controlled such that sessions from the same participant did not occur simultaneously in both the training and testing sets. So as to show the low effect on the randomness involved in sampling smaller subsets, we report these metrics after 10 random re-samplings. It is shown that this regression model is stable when smaller subsets of the training data are used. Had the model shown significant drops in accuracy when restricting the dataset size, it could be concluded that the model was unstable or had overfit the training data.

Since experiments based on subsets of data retrieved from the same experimental population and setting do not necessarily show the true extent of model generalizability, we also performed transfer tests of the story recall model. Specifically, a second dataset was collected from inpatients at a substance abuse program in Louisiana (N = 99), most of whom suffered from co-occurring mood, psychotic, anxiety and personality spectrum disorders, as well as an additional collection from presumed healthy undergraduates at Louisiana State University (N = 124). Together, the inpatients and the presumed healthy undergraduates completed N = 1254 story recalls. A ridge regression model with the same NLP features as previously reported was trained on the initial dataset and tested on the new dataset, as well as vice versa. The first experiment resulted in a Pearson r correlation of 0.86 and the reverse an r of 0.84. Here, we conclude that the story recall regression model will generalize to differing clinical populations as well as illness severities. The same may not hold true for differing cultural populations as language differences may prove to be a confounding variable in transferring such a model. We thus advocate testing models on each new population prior to implementation.

Neuropsychological task scoring is a much more objective application area than other modeling applications in this field in which less is known and gold standard labels are often disagreed upon (e.g., disease detection, mental state tracking, and so on). Thus, generalizability is much more critical to test in these other applications and will potentially not yield such robust conclusions. Nonetheless, the understanding of when a model will yield accurate output and when it will not is an extremely important endeavor. Finding representative data is of the utmost importance in machine learning. In some cases, such as the story recall regression model, it is best to get as much data from as many people as possible. In other cases, especially when dealing with extreme diversity between individuals or subsets of individuals, it may be best to only use data that behaves in a similar fashion to the example currently being tested.

5 Discussion

Mental health is extremely dynamic as it can change on the scale of seconds, minutes, hours, or days, and language offers an objective and potentially unobtrusive way to assay such changes. Mental state in some conditions can change quickly with fatal consequences (e.g., suicide attempts) and more frequent monitoring of language and behavioral data, combined with machine learning methods, has the potential to offer clinicians unprecedented support in tracking patient state. Language can be harnessed for many applications as it offers a quantitative conceptualization of a person's underlying thought processes and mental health. Tracking such phenomena is extremely important yet increasingly complex, and as such there is a need for greater reliance on model outputs in this field.

In experiments involving NLP methods, it is common to deal with high dimensionality from features like word embeddings, parser outputs, and so on, which makes interpretation and understanding of models difficult. Features often go beyond normal distributions and as such there tends to be high variability in data distributions. Thus, it is especially important to create methods and tools that allow us to better understand the feature space and determine whether attributes or classes may violate assumptions.

An eventual goal of this line of work is to have a human-in-the-loop system where models analyze streams of high dimensional patient data and produce predictions of mental state and well-being. In the research stage of this implementation, real data must be analyzed to determine what normal distributions of attributes and classes appear to be. Aberrant instances of patient data can be flagged and reviewed by researchers to either update or exclude from models. Researchers must also test their models' generalizability by collecting additional samples or performing validation techniques to verify performance on unseen data. This process will allow for models to be based on the most accurate and representative data.

In the eventual clinical decision support system implementation, models must be realized such that attribute outliers are not predicted on, but rather the raw data is passed to a clinician to make a judgment. If the outlier is due to faulty feature values, clinicians can update these values or they can create their own labels and update the system such that future similar cases would not necessarily need to be verified by a human. In such a situation, there is a "best of both worlds" where models can execute the tasks that they are best at (high dimensional data analysis) and humans can execute the tasks that they are best at (handling anomalies and interpreting patient data).

For NLP and machine learning methods to be adopted in current research experiments as well as in eventual clinical practice, they require critical peer evaluation. What is needed is transparency in terms of data collection, validation, reproducibility, and clinical agreement in the association of language features to underlying illness. This paper showcases how essential it is that clinicians are involved in all stages of development. As such, it is a large step towards bringing more ethics and transparency into AI-based studies in mental health. Ethics review boards must demand this type of transparency and fairness in the creation of models so that systems that harness machine learning can be implemented in real clinical practice with low risk. Some discussion of this path forward has been brought to light by Friesen et al. (2021) who reported on IRBs as a means of ethics oversight in health research that harnesses AI.

6 Conclusion

This paper illustrates the importance of understanding the assumptions and distributions that underlie training data and the algorithms used, as well as the need to flag data that have characteristics that violate these assumptions. Not only is this knowledge important, but so too is having the tools to do this. We found model instabilities in a story recall regression model with the use of outlier detection algorithms and error analyses with respect to varying input. We advocate that approaches such as these be incorporated into machine learning and NLP-based clinical research and implementation. With the complexities inherent to models based on many features, high numbers of parameters, highly variable human behavioral data, and extremely high (and potentially fatal) stakes for mistakes, it is critical to establish methods beyond model designer intuition in assuring robustness and that predictions cannot be made on out of range data or data that lies in areas of instability. It should now be obvious that high predictive power on a relatively small dataset does not entail clinical relevance or generalizability, and that it is essential to use larger data sets, have more data collection outside of controlled settings, incorporate modeling safeguards, and use human-in-the-loop methodologies at all steps of the process.

Acknowledgements

Parts of this project were funded by grant 231395 from the Research Council of Norway awarded to Brita Elvevåg.

References

Alan Baddeley and Barbara A. Wilson. 2002. Prose recall and amnesia: implications for the structure of working memory. *Neuropsychologia*, 40:1737–1743.

Gillinder Bedi, Facundo Carrillo, Guillermo A. Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B. Mota, Sidarta Ribeiro, Daniel C. Javitt, Mauro Copelli, and Cheryl M. Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1.

Chelsea Chandler, Peter W. Foltz, Jian Cheng, Jared C. Bernstein, Elizabeth P. Rosenfeld, Alex S. Cohen, Terje B. Holmlund, and Brita Elvevåg. 2019. Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 137–147.

Chelsea Chandler, Peter W. Foltz, Alex S. Cohen, Terje B. Holmlund, Jian Cheng, Jared C. Bernstein, Elizabeth P. Rosenfeld, and Brita Elvevåg. 2020a. Machine learning for ambulatory applications of neuropsychological testing. *Intelligence-Based Medicine*, 1–2.

Chelsea Chandler, Peter W. Foltz, and Brita Elvevåg. 2020b. Using machine learning in psychiatry: The need to establish a framework that nurtures trustworthiness. *Schizophrenia Bulletin*, 46:11–14.

Chelsea Chandler, Terje B. Holmlund, Peter W. Foltz, Alex S. Cohen, and Brita Elvevåg. 2021. Extending the usefulness of the verbal memory test: The promise of machine learning. *Psychiatry Research*, 297.

Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, and Nikolaos Mavridis. 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine*, 3.

Michael A. Cirillo and Larry J. Seidman. 2003. Verbal declarative memory dysfunction in schizophrenia: from clinical assessment to genetics and brain mechanisms. *Neuropsychol. Rev*, 13:43–77.

Alex S. Cohen, Taylor L Fedechko, Elana K. Schwartz, Thanh P. Le, Peter W. Foltz, Jared Bernstein, Jian Cheng, Terje B. Holmlund, and Brita Elvevåg. 2019. Ambulatory vocal acoustics, temporal dynamics and serious mental illness. *Journal of Abnormal Psychology*, 128:97–105.

Cheryl M. Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C. Javitt, Carrie E. Bearden, and Guillermo A. Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17:67–75.

Christopher R. Cox, Emma H. Moscardini, Alex S. Cohen, and Raymond P. Tucker. 2020. Machine learning for suicidology: A practical review of exploratory and hypothesis-driven approaches. *Clin Psychol Rev*, 82.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhiguo Ding and Minrui Fei. 2013. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. In *PIFAC Proceedings Volumes*, volume 46, pages 12–17.

Brita Elvevåg, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg. 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93:304–316.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4853.

Elif Eyigoz, Sachin Mathur, Guillermo Cecchi Mar Santamaria, and Melissa Naylor. 2020. Linguistic markers predict onset of alzheimer's disease. *EClinicalMedicine*, 28:304–316.

Peter W. Foltz, Mark Rosenstein, and Karen E. Lochbaum. 2013. Improving performance of automated scoring through detection of outliers and understanding model instabilities. In *Presented at the National Council on Measurement in Education Conference*, San Francisco, CA.

Phoebe Friesen, Rachel Douglas-Jones, Mason Marks, Robin Pierce, Katherine Fletcher, Abhishek Mishra, Jessica Lorimer, Carissa Véliz, Nina Hallowell, Mackenzie Graham, Mei Sum Chan, Huw Davies, and Taj Sallamuddin. 2021. Governing ai-driven health research: Are irbs up to the task? *Ethics Hum Res*, 43:35–42.

Doug Guess and Wayne Sailor. 1993. Chaos theory and the study of human behavior: Implications for special education and developmental disabilities. *The Journal of Special Education*, 27:16–34.

David J. Hand. 2006. Classifier technology and the illusion of progress. *Statistical Science*, 21:1–14.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world. *Behav Brain Sci*, 33:61–83.

Terje B. Holmlund, Chelsea Chandler, Peter W. Foltz, Alex S. Cohen, Jian Cheng, Jared C. Bernstein, Elizabeth P. Rosenfeld, and Brita Elvevåg. 2020. Applying speech technologies to assess verbal memory in patients with serious mental illness. *npj Digital Medicine*, 3.

Terje B. Holmlund, Peter W. Foltz, Alex S. Cohen, Håvard D. Johansen, Randi Sigurdsen, Pål Fugelli, Dagfinn Bergsager, Jian Cheng, Jared Bernstein, Elizabeth Rosenfeld, and Brita Elvevåg. 2019. Moving psychological assessment out of the controlled laboratory setting and into the hands of the individual: Practical challenges. *Psychological Assessment*, 31:292–303.

Benjamin Kompa, Jasper Snoek, , and Andrew L. Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4.

Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, and Renaud David. 2015. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring*, 1:112–124.

Tilman Lange, Mikio L. Braun, Volker Roth, and Joachim M. Buhmann. 2002. Stability-based model selection. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pages 633—-642.

Muriel D. Lezak, Diane B. Howieson, Erin D. Bigler, and Daniel Tranel. 2012. *Neuropsychological assessment (5th Ed.)*. Oxford University Press.

Fei T. Liu, Kai M. Ting, and Zhi-Hua Zhou. 2018. Isolation forest. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, pages 413–422, Pisa, Italy.

Julia Longenecker, Jamie Genderson, Dwight Dickinson, James Malley, Brita Elvevåg, Daniel R. Weinberger, and James Gold. 2010. Where have all the women gone?: participant gender in epidemiological and non-epidemiological research of schizophrenia. *Schizophrenia Research*, 119:240–245.

Ellen W. McGinnis, Steven P. Anderau, Jessica Hruschak, Reed D. Gurchiek, Nestor L. Lopez-Duran, Kate Fitzgerald, Katherine L. Rosenblum, Maria Muzik, and Ryan S. McGinnis. 2019. Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood. *IEEE Journal of Biomedical and Health Informatics*, 23:2294–2301.

Marie-Hélène Metzger, Nastassia Tvardik, Quentin Gicquel, Côme Bouvry, Emmanuel Poulet, and Véronique Potinet-Pagliaroli. 2017. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a french pilot study. *International Journal of Methods in Psychiatric Research*, 26:e1522.

Sina Mohseni, Mandar Pitale, Vasu Singh, and Zhangyang Wang. 2020. Practical solutions for machine learning safety in autonomous vehicles. In *The AAAI Workshop on Artificial Intelligence Safety (Safe AI)*.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Chee Piau Wong. 2018. Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia. *PLoS ONE*, 13:e0205636.

Martin T. Paulus and David L. Braff. 2003. Chaos and schizophrenia: does the method fit the madness? *Neuroscience Perspectives*, 53:3–11.

Zhichao Peng, Qinghua Hu, and Jianwu Dang. 2019. Multi-kernel svm based depression recognition using social media data. *International Journal of Machine Learning and Cybernetics*, 10:43–57.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:2081—2090.

Mark Rosenstein, Peter W. Foltz, Lynn E. DeLisi, and Brita Elvevåg. 2015. Language as a biomarker in those at high-risk for psychosis. *Schizophrenia Research*, 165:249—250.

Euijung Ryu, Alanna M. Chamberlain, Richard S. Pendegraft, Tanya M. Petterson, William V. Bobo, , and Jyotishman Pathak. 2016. Quantifying the impact of chronic conditions on a diagnosis of major depressive disorder in adults: a cohort study using linked electronic medical records. *BMC Psychiatry*, 16.

Adrian B. R. Shatte, Delyse M. Hutchinson, and Samantha J. Teague. 2019. Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, 49:1426–1448.

Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. 2020. Deep learning in mental health outcome research: a scoping review. *Transl Psychiatry*, 10.

Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction*, 27:Article 34.

Peter Turney. 1995. Technical note: Bias and the quantification of stability. *Machine Learning*, 20:23–33.

Zhuoran Wang, Anoop D. Shah, A. Rosemary Tate, Spiros Denaxas, John Shawe-Taylor, and Harry Hemingway. 2012. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One*, 7:e30412.

David Wechsler. 1997. *Wechsler Memory Scale - Third Edition, WMS-III: Administration and scoring manual*. The Psychological Corporation.

Robert Whelan and Hugh Garavan. 2014. When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biol Psychiatry*, 75:746–748.

Jheng-Long Wu, Liang-Chih Yu, and Pei-Chann Chang. 2012. Detecting causality from online psychiatric texts using inter-sentential language patterns. *BMC Medical Informatics and Decision Making*, 12.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingquan Zhu and Xindong Wu. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22:177–210.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33.