Published in final edited form as:
Acta Neurol Scand. 2020 May ; 141(5): 388–396. doi:10.1111/ane.13216.

Identifying epilepsy psychiatric comorbidities with machine learning

Tracy Glauser[1], **Daniel Santel**[1], **Melissa DelBello**[2], **Robert Faist**[1], **Tonia Toon**[1], **Peggy Clark**[1], **Rachel McCourt**[1], **Benjamin Wissel**[1], **John Pestian**[1]

[1]Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

[2]Department of Psychiatry and Behavioral Neuroscience, University of Cincinnati, Cincinnati, Ohio

Abstract

Objective: People with epilepsy are at increased risk for mental health comorbidities. Machine-learning methods based on spoken language can detect suicidality in adults. This study's purpose was to use spoken words to create machine-learning classifiers that identify current or lifetime history of comorbid psychiatric conditions in teenagers and young adults with epilepsy.

Materials and Methods: Eligible participants were >12 years old with epilepsy. All participants were interviewed using the Mini International Neuropsychiatric Interview (MINI) or the MINI Kid Tracking and asked five open-ended conversational questions. N-grams and Linguistic Inquiry and Word Count (LIWC) word categories were used to construct machine learning classification models from language harvested from interviews. Data were analyzed for four individual MINI identified disorders and for three mutually exclusive groups: participants with no psychiatric disorders, participants with non-suicidal psychiatric disorders, and participants with any degree of suicidality. Performance was measured using areas under the receiver operating characteristic curve (AROCs).

Results: Classifiers were constructed from 227 interviews with 122 participants (7.5 ± 3.1 minutes and 454 ± 299 words). AROCs for models differentiating the nonoverlapping groups and individual disorders ranged 57%-78% (many with $P < .02$).

Discussion and Conclusion: Machine-learning classifiers of spoken language can reliably identify current or lifetime history of suicidality and depression in people with epilepsy. Data suggest identification of anxiety and bipolar disorders may be achieved with larger data sets. Machine-learning analysis of spoken language can be promising as a useful screening alternative

Correspondence: Tracy Glauser, Cincinnatil Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH 45229, USA. tracy.glauser@cchmc.org.

CONFLICT OF INTEREST
Drs. Glauser and Pestian assert that they have no conflicts of interest that could inappropriately bias this work except they are inventors on the issued patent: "Processing Text with Domain-Specific Spreading Activation Methods". Drs. Santel, DelBello, and Clark along with Mr Faist, Ms Toon, Mr Wissel, and Ms McCourt assert that they have no conflicts of interest that could inappropriately bias this work.

DATA AVAILABILITY STATEMENT
Research data are not shared.

when traditional approaches are unwieldy (eg, telephone calls, primary care offices, school health clinics).

Keywords

artificial intelligence; childhood absence epilepsy; natural language processing; psychiatric screening

1 | INTRODUCTION

Individuals with epilepsy are at an enhanced risk of developing a number of mental health comorbidities,[1] including depression,[2] agoraphobia,[3] anxiety disorders,[2] social phobia,[4] and suicide.[2] Children, in particular, show significantly higher rates of depressive disorders and anxiety disorders.[5] In 2007, a population study of psychiatric disorders in patients with epilepsy reported rates of 14.1 (7.0-21.1)% (95% CI) for depression (past 12 months), 12.8 (6.0-19.7)% for anxiety (past 12 months), and 25.0 (17.4-32.5)% for suicidal ideation (lifetime).[2] Yet, frequently general practice and specialist offices do not implement systematic screenings for detection of these comorbidities, and they often go undiagnosed[6] despite recommendations.[7]

Investigators have used machine-learning to identify psychiatric disorders from the spoken words of individuals. For example, machine-learning algorithms have been developed to "listen" to the spoken language of individuals and identify their risk of suicide,[8–10] obtaining areas under the receiver operator curve (AROCs) greater than 85%. Written language has also been analyzed by natural language processing and machine-learning techniques to identify suicidality and other psychiatric disorders. Suicide notes have been distinguished from fake suicide notes,[11] the works of suicidal songwriters and poets have been distinguished from the works of non-suicidal songwriters and poets,[12] and Twitter posts have been successfully classified into a number of psychiatric disorders.[13] There are indications that there are significant differences between spoken and written word, however, and it is not clear that the results are cross-applicable.[14,15]

To date, there have been few examples of using spoken language to identify psychiatric disorders beyond suicidality. Further, machine-learning has not yet been used to analyze the language of patients with epilepsy to identify psychiatric comorbidities. The goal of this work was to create machine-learning classifiers that identify current or lifetime history of comorbid psychiatric conditions in teenagers and young adults with epilepsy, using their language harvested during semi-structured interviews.

2 | METHODS

The objective of this study was to determine whether machinelearning models can be constructed to identify current or lifetime history of comorbid psychiatric conditions in people with epilepsy using the language gathered in semi-structured interviews.

2.1 | Population

Participants were eligible if they had a documented history of epilepsy and were older than 12 years of age. This single-site prospective study was approved by the Cincinnati Children's Hospital Medical Center institutional review board. Written informed consent was obtained from all participants (or their parent or legal guardian, when applicable) prior to enrollment.

After receiving informed consent, interviewers collected medical history from participants and subsequently screened each for psychiatric disorders using either the Mini International Neuropsychiatric Interview (MINI),[16,17] for participants aged eighteen years or older, or the MINI Kid Tracking,[18] for participants under eighteen. The MINI (Kid) is a diagnostic interview composed of modules that assess the severity and/or presence of a variety of psychiatric disorders and symptoms. For this study, the modules for mania, depression, suicidality (risk of suicide), agoraphobia, generalized anxiety disorder, social phobia, and panic disorder were chosen. The MINI provides results for both lifetime and current presence of disorders. Subjects were asked five open-ended Ubiquitous Questions (UQs)[10,19] to harvest the language that would serve as the input to the machine-learning model: Do you have hope? Do you have any secrets? Are you angry? Do you have any fear? Does it hurt emotionally? The interviewers were encouraged to ask follow-up questions to continue the conversation. All interviews were conducted by one of two experienced interviewers (a certified clinical research professional or a research nurse practitioner). A second interview following the same procedure, including updating of medical history, a repeat of the MINI assessment, and repeat of the UQs, was conducted with participants six months after the first interview.

For analysis, two types of groupings were created for classification. *Individual comorbidity* groups were defined such that they included a specific type of comorbidity, regardless of any other disorder. *Anxiety disorders* included any identified lifetime history of social phobia, panic disorder, generalized anxiety disorder, or agoraphobia. *Depressive disorders* included any lifetime history of depressive episodes or dysthymia. *Bipolar disorders* included any lifetime history of manic or hypomanic episodes or bipolar disorders, as identified by the MINI. The MINI categorizes suicidality into low, medium, high, and no risk. In this analysis, *suicidality* included low, medium and high risk. In addition, *mutually exclusive* comorbidity groups were defined: *suicidality* as defined above; *non-suicidal psychiatric disorders* for the presence of anxiety, depression, or bipolar disorders if there was not also an indication of suicidality; and *no psychiatric disorders.* For depression, suicidality, and bipolar, this study's analysis cohort included individuals with either current or lifetime history of the disorder. This was done due to the small sample size of participants with active illness for these groups. Since the anxiety disorder questions on the MINI only assess current status, no lifetime data is available for this group.

Statistical metrics of language use, words per sentence and syllables per word, are collected using the python library textstat.[20] These were collected for each individual disorder cohort to compare complexity of language between the presence and absence of a disorder.

2.2 | Data preparation and model creation

Machine-learning algorithms require measurable numeric quantities (features or explanatory variables) as inputs. The machine-learning classifiers investigated here were constructed to use language characteristics, specifically *n*-grams (contiguous sequences of *n* words) and frequencies of word categories as described by the Linguistic Inquiry and Word Count (LIWC) software. LIWC is a text analysis and word categorization program that counts words in psychologically meaningful categories. In this work, the LIWC word categories were extracted using the 2007 LIWC dictionary.[21] All unigram, bi-gram, and tri-gram frequencies were extracted for all interviews, and those that occurred in at least two different interviews were kept. Frequencies for all 64 LIWC word categories (excluding the LIWC categories involving punctuation and total document word counts) were also calculated.

In general, the vast majority of features would contribute little to identifying psychiatric disorders; that is, they add noise to the training set used to construct the machine-learning model. To train with only the set of features that contribute the most information to the classification, multiple feature reduction passes were used. The first feature reduction pass included a cut on features with a near-zero variance across the dataset. The nearZeroVar function in the R package caret[22] was used with a frequency cut (cutoff for the ratio of the most common value to the second most common value) of 19 (the default), with a unique cut (the cutoff for the percentage of distinct values out of the total number of samples) of 5.

Then, to select the optimal features to be used by the classifier, the remaining features were ranked for each constructed classifier based on their Kolmogorov-Smirnov (KS) test *P*-value,[23] which was evaluated for each feature by comparing the frequency distributions for the positive and negative classes for a given disorder. The optimal number of features, then, was that which maximized the performance of the machine-learning classifier.

A Support Vector Machine (SVM) model was used for classification.[24] SVMs have been proven useful for classification problems such as these,[8,10] due to their robustness to overfitting and ability to perform well in high-dimensional spaces.[25,26] Linear and radial kernels were all considered in training the classifiers, and the kernel with the best overall performance was chosen. The hyperparameters of the SVM kernels were tuned using 10-fold cross-validation with 5 repetitions and an adaptive grid search as implemented by the caret[22] package. The hyperparameters C (for both linear and radial kernels) and γ (radial only) describe the shape of the hyperplane that attempts to separate classes of data in an SVM. Tuning within cross-validation is required to find a boundary that provides good discrimination without overfitting. The number of features used in training models was also tuned and allowed to vary as powers of 2 between 16 and 512. This cross-validation strategy is used only for model tuning and is not used to validate performance.

Instead, two separate strategies were employed to validate the model. In the first, only the first interviews were considered. AROCs were evaluated using the machine-learning decision values evaluated on the left-out points in a leave-one-out (LOO) cross-validation,[27] where training and feature selection were performed within each of the training folds *(first interview LOO)*. That is, in each fold of the LOO analysis, the entire creation and training of the model were performed without any knowledge of the single left-out document. This was then repeated for every document. LOO validation maximizes the amount of data available for training at each step and represents a best-case performance rating for a given dataset and is especially useful in the case of limited data, but there are common fears that it may over-estimate performance. So, a second validation method was employed, and a training set was constructed using a combination of half of the first interviews and half of the second interviews. A validation set was constructed from the "non-training set" first and second interviews *(first and second train/validate).* The training set was used to train the SVM models, and AROCs were again evaluated using the decision values created on the validation set. Similarly, the held-out portion in this validation strategy was never used in construction of the model.

For all classifiers, AROCs were used to measure the performance of the classifiers. An excellent classifier would have an AROC >90%, while a random classifier would have an expected AROC = 50%. In 2005, Rice et al suggested that in psychology studies, an effect size $d = 0.8$, equivalent to an AROC of ≈71% is "about as high as they come".[28] For this study, we believe that a classifier with an AROC >65% is clinically meaningful. Confidence intervals were calculated by the pROC[29] package for R, using the DeLong method.[30]

SVM models were trained separately to distinguish between subjects within specific disorder groups and those without that disorder, regardless of other additional comorbid diagnoses, as well as to distinguish between the constructed non-overlapping *no psychiatric disorder, suicidality,* and *non-suicidal psychiatric disorder* groups. For visualization of the potential discrimination power we might see with the available data given the feature reduction processes, a linear discriminant analysis (LDA),[31] a supervised dimensionality reduction technique, was used. The created LDA model was not directly used for classification.

For this study, a successful classifier model was defined as having $P < .05$ for both the *first interview leave-one-out cross-validation* and the *first and second train/validate* validation schemes described above.

3 | RESULTS

3.1 | Population and data collection

Overall, 140 participants were enrolled between June 2016 and December 2017 and performed at least one interview. Due to technical issues, audio recordings or diagnostic information were lost for 18 first interviews, and not all participants completed two interviews. The analyzed sample consists of 122 first interviews and 105 seconds interviews. The overall cohort's demographics and seizure/epilepsy characteristics along with these variables by both group and individual comorbidities are shown in Tables 1 and 2. In all comorbidity groups, there were more female participants. There were no statistically significant differences in demographics, seizure type or epilepsy characteristics among the comorbid disorders. Of the 30 participants with some degree of suicidality determined by the MINI (Kid), at the first interview, the distribution was low (n = 24), medium (n = 4), and high (n = 2) risk. For the 21 participants with suicidality flagged at the second interview, the distribution was low (n = 15), medium (n = 3), and high (n = 3) risk.

The analysis involved participant language harvested from the dialogue using the five open-ended Ubiquitous Questions. Overall, the participant-interviewer dialogue lasted 7.5 ± 3.1 minutes and involved 454 ± 299 words from the study participants. Second interviews on average were longer and contained more words than first interviews (first: 6.7 ± 2.7 minutes and 385 ± 256 words, second: 8.1 ± 3.3 minutes and 534 ± 327 words). There was no significant difference between male and female participants in dialogue duration (male: 6.4 ± 2.4, female: 7.3 ±2.8 minutes) or the mean number of words spoken by the participant (male: 387 ± 247, female: 406 ± 298). The MINI took between 10 and 20 minutes to administer.

Total word counts for presence/absence of *bipolar disorders* was 421 ± 131 vs 381 ± 23 ($P = .84$); for *depressive disorders* 480 ± 54 vs 351 ± 24 ($P = .03$); for *suicidality* 475 ± 48 vs 350 ± 25 ($P = .02$); for *anxiety disorders* 354 ± 61 vs 384 ± 25 ($P = .6$). Syllables per word for presence/absence of *bipolar disorders* was 1.19 ± 0.03 vs 1.173 ± 0.006 ($P = .5$); for *depressive disorders* 1.191 ± 0.011 vs 1.169 ± 0.007 ($P = .1$); for *suicidality* 1.196 ± 0.010 vs 1.166 ± 0.007 ($P = .02$); for *anxiety disorders* 1.202 ± 0.019 vs 1.170 ± 0.006 ($P = .1$). Words per sentence for presence/absence of *bipolar disorders* was 37 ± 7 vs 49 ± 4 ($P = .1$); for *depressive disorders* 49 ± 5 vs 48 ± 5 ($P = .8$); for *suicidality* 49 ± 6 vs 48 ± 5 ($P = .9$); for *anxiety disorders* 59 ± 19 vs 47 ± 4 ($P = .5$).

3.2 | Data preparation and model creation

With the LIWC categories and *n*-grams combined, and before applying the feature reduction techniques, there were a total of 21 603 features identified over the 227 analyzed interviews. The removal of near-zero variance features reduced the total number to 1982. During SVM training, the radial kernel was found to consistently have the best overall classification performance. This is not unexpected; with proper hyperparameter tuning radial kernels should always perform at least as well as linear kernels.[32]

A plot of the LDA model applied to the data from the first and second interviews is shown in Figure 1. It was created with the three non-overlapping groups labeled, using the full set of 227 first or second interviews, and with the 1982 features available before applying the KS-test ranking. Visual inspection showed clear separation and clustering of the three groups, although there were some outliers. The *suicidality* datapoint on the far left (past the *no psychiatric disorder* cluster) was a very short interview (74 words), and likely was not long enough for an accurate classification. A few words that correlated with *no psychiatric disorder* participants (“mom,” “mother,” “mind,” “happy”) influenced the *suicidality* outlier’s datapoint classification.

3.3 | Model performance

The results and performance metrics of the *first interview LOO* and *first and second train/validate* classifiers are presented in Tables 3 and 4. Table 3 shows the results for classifying the three non-overlapping groups, and Table 4 presents the classification results for identifying the presence of individual disorders regardless of other present comorbidities. The performance metrics of the classifiers are presented along with the number of features used to train the models and the top five features correlated to each of the groups or individual disorders as determined by the SVM's weighting. The statistical significance of each model's performance was calculated compared to a purely random classifier.

3.4 | Non-overlapping group analysis

The suicidality classification model performed significantly better than random chance in both validation techniques (*P* = .00018 and *P* = .003) at distinguishing participants with *suicidality* from participants with *no psychiatric disorder* as shown in Table 3. The suicidality classification model was able to distinguish participants with *suicidality* from participants with *non-suicidal psychiatric disorders* at a statistically significant rate in one validation method and a trend in the other (AROC = 71%, *P* = .015; AROC = 67%, *P* = .09). In contrast, the suicidality classification model distinguished participants with *suicidality* from the combination of participants with *no psychiatric disorder* or *non-suicidal psychiatric disorders* at a statistically significant rate better than random chance in one validation method but a non-significant rate in the other (AROC = 57%, *P* = .22; AROC = 71%, *P* = .0011). Lastly, the classification model was able to distinguish participants with *non-suicidal psychiatric disorders* from participants with *no psychiatric disorder* at a statistically significant rate better than random chance in one validation method and a trend in the other (AROC = 73%, *P* < .0001; AROC = 64%, *P* = .09).

3.5 | Individual disorder analysis

The *depressive disorder* classification model performed significantly better than random chance in both validation techniques (*P* = .011 and *P* = .001) at distinguishing participants with *depressive disorders* from participants with no evidence on the MINI of current or lifetime history of depressive disorder (AROC = 65%, *P* = .011; AROC = 69%, *P* = .001) (Table 4). The individual *suicidality* classification model results in Table 4 are the same as the *suicidality* classification model results in Table 3 where the participants with *suicidality* are compared to the combination of participants with *no psychiatric disorder* or *non-suicidal psychiatric disorders.* The individual *anxiety disorder* and *bipolar disorder* classification models distinguished participants with the specific individual disorders from participants with no evidence on the MINI of current or lifetime history of the specific individual disorder at a statistically significant rate better than random chance in one validation method but a non-significant rate in the other validation method Table 4. However, there were only 15 participants with *anxiety disorders* and 6 participants with *bipolar disorders.*

4 | DISCUSSION

This study demonstrated through multiple analyses that machine-learning classifiers of spoken language can reliably identify current or lifetime history of suicidality and depressive disorders in people with epilepsy. Identification of anxiety and bipolar disorders using spoken language was achieved in some but not all of our analyses; this was most likely due to the small sample sizes in each of these cohorts. We anticipate that with larger numbers of participants with anxiety or bipolar disorders, these classifiers would become more reliable.

An initial reaction to the distinguishability of the language of the suicidality and depressive disorder cohorts (and to a lesser extent the anxiety and bipolar disorders cohorts) may be the suspicion that those with a history of those disorders have a lower facility with language in general. The analysis might, then, be picking up on this lower language functioning. On the contrary, no statistically significant difference was noticed in the depressive disorders, anxiety, or bipolar cohorts when comparing various language complexity metrics (words per sentence and syllables per word) to their interviews. The only difference noted in participants with current or lifetime history of suicidality was in syllables per word but without a difference in words per sentence.

Rather than language complexity, we propose a conceptual model where language use serves as a proxy for the expression of the function of integrated brain networks. A combination of genetics and environment lead to the development, structure, and subsequent function of these networks. A given structure of brain networks may have a "psychiatric susceptibility" (a long-term trait, eg, the long-term risk of suicidal ideation) which is just an inclination to a possible "psychiatric expression" (the current state, eg, actual suicidal ideation). At the same time, there is evidence that the expression of and reaction to language are also linked to the structure of a person's brain networks.[33] We propose, then, that in studying the differences in word use between cohorts, our classifiers are identifying differences in both the susceptibility *and* expression.

We propose that with the right dataset-a sufficient quantity of people with history of a disorder, current expression of a disorder, and no history of a disorder-this could be taken further to identify language features specifically associated with only susceptibility, only expression, or both for a psychiatric disorder.

The *depressive disorder* classifiers performed significantly better than random chance ($P < .02$) under both types of validation schemes. The results of the other classifiers are promising, but their performance is marginal compared to chance given the confidence intervals, in at least one validation scheme or other. The performance of the *suicidality* classifiers is not as strong as seen in some other similar studies,[9] but it is important to note that the vast majority of our participants with some amount of suicidality were identified as "low risk," whereas the participants in the other cited studies had been admitted to hospital EDs in situations related to suicidal ideation or attempts. Model performance can be expected to improve as the overall sample size and pool of higher-risk participants for training grows.

In general, there is not always an obvious interpretation for the way an SVM utilizes features. In fact, it is a linear combination of the set of training features that leads to classification of a data point. Additionally, the features presented in Tables 3 and 4 are small fractions of the total feature set used to train most of the presented models. As shown in Tables 3 and 4, despite exhibiting similar performance, the models trained for the different validation schemes often chose completely different features and completely different numbers of features. A rough indication of the importance of features can be gathered based on the weights chosen by the SVM's training, and the most heavily weighted are displayed in the tables. No obvious patterns emerge, and a full linguistic analysis is beyond the scope of this paper. However, a cursory examination of the features ranked by the KS tests does show a few patterns that agree with the literature. Many *n*-grams with personal pronouns occurred more often in *suicidiality* interviews: “I guess,” “because I,” “I I,” “that I,” “I do not” featured prominently. Other analyses have shown that pronouns play a larger role in the speech of those with suicidal tendencies. Coppersmith et al[13] evaluated Twitter data using LIWC and found that first person singular and third person plural pronouns were more commonly used in those who attempt suicide compared to non-suicidal controls. In studies of suicidal poets and songwriters, Stirman and Pennebaker[12] also saw an increased use of first person pronouns.

Despite this, most of the top features are not overtly descriptive. This highlights the power of machine-learning techniques to pick up on trends in language use that would potentially be missed or hard to interpret even by experts. This problem of explainability does present other issues, however, and is the subject of major ongoing research in the field of machine learning.[34]

According to the criteria outlined above, classifiers were successfully constructed to distinguish subjects with MINI diagnoses of *depressive disorders* and *suicidality* based on language use. Classification of comorbidities of *anxiety disorders, bipolar disorders* and the larger group of *non-suicidal psychiatric disorders* was promising, but more data are needed. This study provides further evidence that language can be used to identify psychiatric disorders, with many AROCs >70%. Other studies have shown that with sufficient high-quality data, this classification can be done with AROCs >90%, which would make this a viable alternative to other screening methods with similar or worse performance.

The promising results in classifying the group of *non-suicidal psychiatric disorders* suggest that in the future, with more data, it may be possible to identify aspects of language use that are common across multiple disorders, which may lead to identification of similarities in the disorders, whether from a root source or in common effects and may therefore provide guidance for therapeutic methods.

With promising or suggestive performance in classifying multiple types of disorders and an average administration time of seven minutes, the Ubiquitous Questions (UQs) or other free-form conversational interview combined with a machine-learning analysis could provide a useful alternative tool in situations traditional psychiatric screening and testing is unwieldy (such as telephone calls, primary care offices, school health clinics). These results provide support for more extensive research into using language to identify not just other psychiatric disorders comorbid with epilepsy, but the benefits of constructing corpora of data specific to psychiatric comorbidities of primary disease vs generalizable psychiatric disorder corpora.

This analysis should be considered a proof-of-concept, adding to previous research on suicidality classifiers to demonstrate that classifiers can also in principle be constructed that can classify language consistent with depression, anxiety, or generically a variety of non-suicidal mental disorders. Additionally, in most cases these classifiers were trained on participants that were not reporting any current distress, further suggesting that there might be a difference in the language of people with a remitted mental disorder compared to people that have never experienced a mental disorder. The sparsity of the data did not allow for this to be deeply investigated, but it merits further research. Future research could potentially enable discrimination between lack of disorder, remitted disorder, and current severity of disorder, but this will require large amounts of high-quality data.

While the interviews in this study were transcribed manually, as computerized speech-to-text capabilities continue to improve this step can likely be more fully automated. This would allow software solutions to provide almost instant classification results following an interview. Current plans are to complete third follow-up interviews with the same participants, at another six-month interval, to allow for a longitudinal analysis of the changes in language specific to psychiatric disorders and an examination of language's ability to predict future psychiatric states.

ACKNOWLEDGEMENTS

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines. We acknowledge and thank the Charles L. Shor Foundation for their financial support of the study.

Funding information

Charles L Shor Foundation

REFERENCES

1. Strine TW, Kobau R, Chapman DP, Thurman DJ, Patricia P, Balluz LS. Psychological distress, comorbidities, and health behaviors among US adults with seizures: results from the 2002 National Health Interview Survey. Epilepsia. 2005;46(7):1133–1139. [PubMed: 16026567]
2. Tellez-Zenteno JF, Patten SB, Jetté N, Williams J, Wiebe S. Psychiatric comorbidity in epilepsy: a population-based analysis. Epilepsia. 2007;48(12):2336–2344. [PubMed: 17662062]
3. Tucker GJ. Seizure disorders presenting with psychiatric symptomatology. Psychiatr Clin North Am. 1998;21(3):625–635. [PubMed: 9774800]
4. De Boer HM, Marco M, Sander JW. The global burden and stigma of epilepsy. Epilepsy Behav. 2008;12(4):540–546. [PubMed: 18280210]
5. Jones JE, Watson R, Sheth R, et al. Psychiatric comorbidity in children with new onset epilepsy. Dev Med Child Neurol. 2007;49(7):493–497. [PubMed: 17593119]
6. Druss BG, Walker ER. Mental disorders and medical comorbidity. Synth Proj Res Synth Rep. 2011;21:1–26.
7. Guilfoyle SM, Wagner JL, Smith G, Modi AC. Early screening and identification of psychological comorbidities in pediatric epilepsy is necessary. Epilepsy Behav. 2012;25(4):495–500. [PubMed: 23153713]
8. Pestian JP, Sorter M, Connolly B, et al. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. Suicide Life Threat Behav. 2017;47(1):112–121. [PubMed: 27813129]
9. Pestian J, Matykiewicz P, Cohen K, et al. Suicidal thought markers: a controlled trial examining the language of suicidal adolescents. 46th American Association of Suicidology Annual Conference, Austin. 2013.
10. Pestian JP, Grupp-Phelan J, Bretonnel Cohen K, et al. A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. Suicide Life Threat Behav. 2016;46(2):154–159. [PubMed: 26252868]
11. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: a content analysis. Biomed Inform Insights. 2010;2010(3):19–28. [PubMed: 21643548]
12. Wiltsey Stirman S, Pennebaker JW Word use in the poetry of suicidal and nonsuicidal poets. Psychosom Med. 2001;63(4):517–522. [PubMed: 11485104]

13. Glen C, Ryan L, Eric W, Tony W. Quantifying suicidal ideation via language usage on social media. Joint Stat Meet Proc Stat Comput Sect. 2015.

14. Rayson P, Wilson A, Leech G. Grammatical word class variation within the British National Corpus Sampler. 2001.

15. Leech G, Rayson P, Wilson A. Word Frequencies in Written and Spoken English: Based on the British National Corpus. London and New York: Routledge; 2014.

16. Yves L, Sheehan DV, Emmanuelle W, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. Eur Psychiatry. 1997;12(5):224–231.

17. Sheehan DV, Lecrubier Y, Harnett SK, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. Eur Psychiatry. 1997;12(5):232–241.

18. Sheehan DV, Sheehan KH, Douglas SR, et al. Reliability and validity of the mini international neuropsychiatric interview for children and adolescents (MINI-KID). The Journal of Clinical Psychiatry. 2010;71(3):313–326. [PubMed: 20331933]

19. John P A conversation with Edwin Shneidman. Suicide and life-Threatening Behavior. 2010;40(5):516–523. [PubMed: 21034214]

20. Bansal Stextstat. https://github.com/shivam5992/textstat 2019.

21. LIWC. LIWC: Linguistic Inquiry and Word Count, https://liwc.wpengine.com

22. Max K. Contributions, Jed Wing, Weston Steve, Williams Andre, et al. Caret: Classification and Regression Training 2016. R package version 6.0–73.

23. Wilcox R Kolmogorov-smirnov test. Encyclopedia of biostatistics. 2005.

24. Schölkopf B, Smola A. Support vector machines. Encyclopedia of Biostatistics. 1998.

25. Press William H, Teukolsky Saul A, Vetterling William T, Flannery BP. Section 16.5. support vector machines. Numerical Recipes: The Art of Scientific Computing. 2007.

26. Sebastiani F Machine learning in automated text categorization. ACM Comput Surv. 2002;34(l):1–47.

27. Bradley E, Tibshirani RJ. An Introduction to the Bootstrap. CRC Press; 1994.

28. Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. Law Hum Behav. 2005;29(5):615–620. [PubMed: 16254746]

29. Xavier R, Natacha T, Alexandre H, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. [PubMed: 21414208]

30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837. [PubMed: 3203132]

31. Radhakrishna Rao C The utilization of multiple measurements in problems of biological classification. J R Stat Soc Series B. 1948;10(2):159–203.

32. Sathiya KS, Chih-Jen L. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. 2003;15(7):1667–1689. [PubMed: 12816571]

33. Adam JM, Lisa P, Cherkassky VL, et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. Nature Human. Behaviour. 2017;1(12):911–919.

34. Ribana R, Bastian B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. arXiv preprint arXiv:1905.08883. 2019.

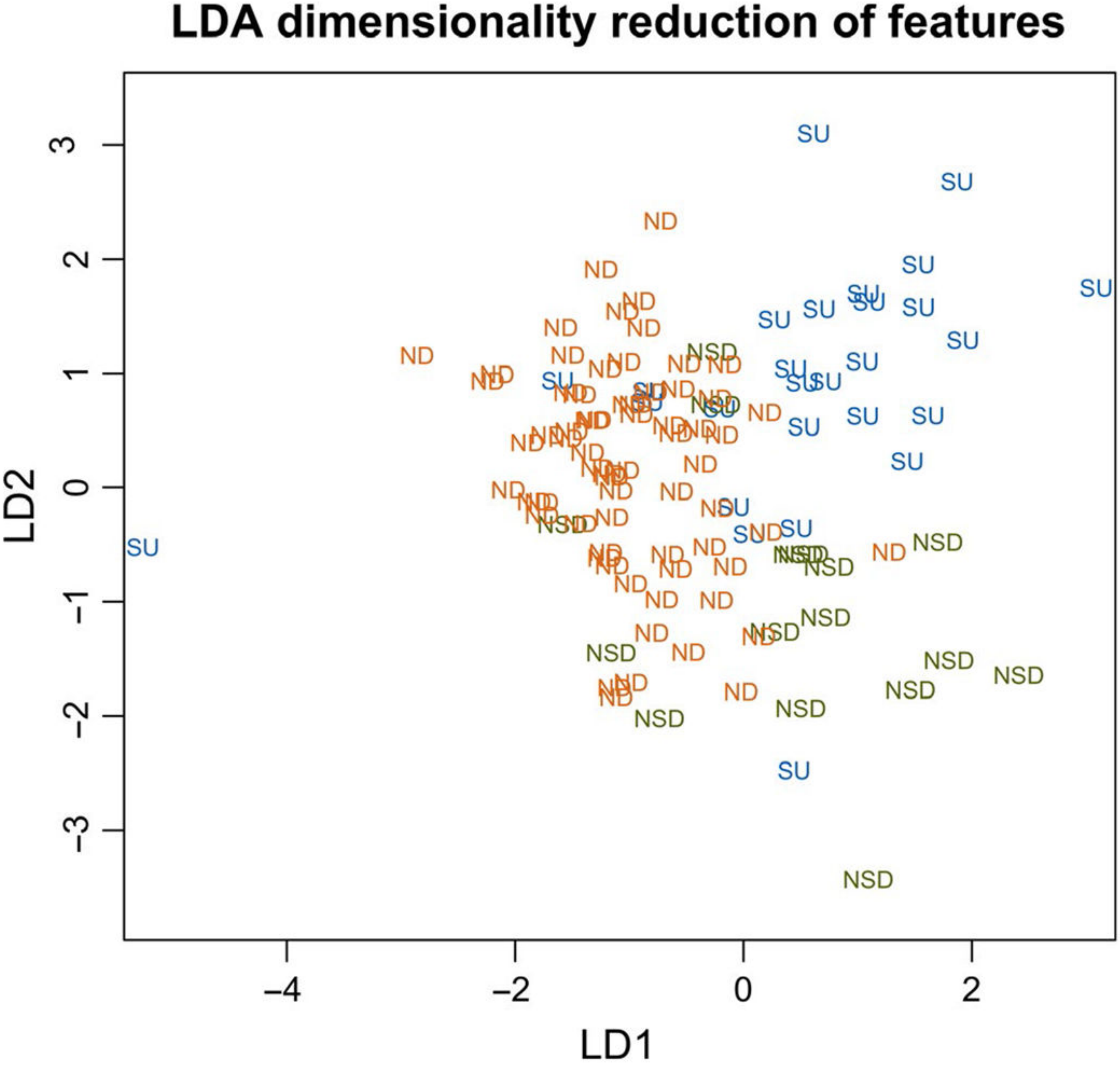

FIGURE 1.

Dimensionality reduction of 227 interviews and 1982 total features. A linear discriminant analysis (LDA) is used with supervision. The three classes are ND = no psychiatric disorder, SU = suicidality, NSD = non-suicidal psychiatric disorder; that is, all psychiatric disorders identified without a comorbidity of suicidality

TABLE 1.

Demographic information for study participants at the time of the first interviews. Non-suicidal psychiatric disorders include all psychiatric disorders identified without a comorbidity of suicidality. *No psychiatric disorder,* suicidality, and *non-suicidal psychiatric disorders* are mutually exclusive groups. Many participants were identified with multiple individual comorbidities

		Comorbidity by group			**Individual comorbidity**		
	Overall	**N psychiatric disorder**	**Non-suicidal psychiatric disorder**	**Suicidality**	**Depressive disorder**	**Anxiety disorder**	**Bipolar disorder**
Count	122	75	17	30	29	15	6
Age (years)	17 ± 2	17 ± 3	19 ± 2	18 ± 2	18 ± 2	19 ± 2	18 ± 2
Male sex (%)	38	43	29	30	24	27	33
Hispanic ethnicity (%)	11	13	0	13	10	13	17
Race							
White (%)	72	67	76	83	79	73	67
Black/African American (%)	15	16	18	10	14	7	17
Other or missing (%)	13	17	6	7	07	20	17
Epilepsy duration at interview (years)	9.8 ± 1.4	9.8 ± 1.5	9.8 ± 1.5	10.1 ± 0.9	9.9 ± 1.3	10.1 ± 0.9	10.5 ±0.6
Seizure type							
Generalized onset (%)	93	92	88	97	93	93	100
Focal onset (%)	5	4	12	3	7	7	0
Unknown onset (%)	2	4	0	0	0	0	0
Syndrome							
CAE (%)	88	88	88	87	83	80	83
None (%)	8	8	12	7	10	13	17
Other (%)	4	4	0	7	7	7	0

TABLE 2.

Demographic information for study participants at the time of the second interviews. *Non-suicidal psychiatric disorders* include all psychiatric disorders identified without a comorbidity of suicidality. *No psychiatric disorder, suicidality,* and *non-suicidal psychiatric disorders* are mutually exclusive groups. Many participants were identified with multiple individual comorbidities

		Comorbidity by group			Individual comorbidity		
	Overall	**No psychiatric disorder**	**Non-suicidal psychiatric disorder**	**Suicidality**	**Depressiv disorder**	**Anxiety disorder**	**Bipolar disorder**
Count	105	72	12	21	21	11	1
Age (years)	18 ± 2	18 ± 2	19 ± 2	18 ± 2	19 ± 2	19 ± 3	18
Male sex (%)	36	39	33	19	24	18	0
Hispanic ethnicity (%)	13	14	0	19	14	18	0
Race							
White (%)	70	68	67	81	76	55	100
Black/African American (%)	14	15	17	10	10	18	0
Other or missing (%)	15	17	17	10	14	27	0
Epilepsy duration at interview (years)	10.8 ± 1.0	10.8 ± 0.9	11.3 ± 1.3	10.7 ± 0.9	10.7 ± 1.2	11.2 ± 1.5	12
Seizure type							
Generalized onset (%)	98	97	100	100	100	100	100
Focal onset (%)	1	1	0	0	0	0	0
Unknown onset (%)	1	1	0	0	0	0	0
Syndrome							
CAE (%)	97	97	100	95	95	100	100
None (%)	2	3	0	0	0	0	0
Other (%)	1	0	0	5	5	0	0

TABLE 3.

Classifier performance for identifying three distinct groups: suicidality, *non-suicidal psychiatric disorders,* and *no psychiatric disorders.* The number of features used for classification was a parameter tuned separately from the C and γ SVM kernel hyperparameters. The AROC is the Area under the Receiver Operating Characteristic curve. The *P*-values represent the significance of the AROCs compared to a completely random classifier. “First group vs second group” describes, for example, “Suicidality vs No disorder.” “LIWC” indicates features that represent LIWC word categories. (a) Leave-one-out (LOO) cross-validation performance using the first interviews with all participants. The feature selection for the cross-validation was performed within each fold; however, the “top features” presented here use the same method applied to the entire dataset. (b) Training sets constructed from stratified random samples of 50% of first interviews and 50% of second interviews with performance tested against all remaining interviews

(a)				
	Suicidality (n = 30) vs No disorder (n = 75)	**Suicidality (n = 27) vs Non-suicidal disorder (n = 16)**	**Suicidality (n = 30) vs No disorder + Non-suicidal disorder (n = 92)**	**Non-suicidal disorder (n = 17) vs No disorder (n = 75)**
AROC (95% CI)	72 (61-83)%	71 (55-89)%	57 (44-71)%	73 (62-85)%
P-value	.00018	.015	.22	<.0001
Total No. Features	16	256	16	64
Top 5 features for first group (SVM weights)	out, other, its, how, in a	some, them, from, in a, things	try to, sometimes, things, I do not really, my	or, can, you know, do not i, if
Top 5 features for second group (SVM weights)	no, things, people, my, it	okay, is, had, i guess, guess	I have, in, try, assent (LIWC), inhibit (LIWC)	i guess, guess, better, i do, though
(b)				
	Suicidality (n = 25) vs No disorder (n = 73)	**Suicidality (n = 26) vs Non-suicidal disorder (n = 14)**	**Suicidality (n = 26) vs No disorder + Non-suicidal disorder (n = 88)**	**Non-suicidal disorder (n = 15) vs No disorder (n = 73)**
AROC (95% CI)	71 (59-83)%	67 (48-87)%	71 (59-83)%	64 (48-79)%
P-value	.003	.09	.0011	.09
Total No. Features	512	128	64	512
Top 5 features for first group (SVM weights)	little, for, have a, how, now	been, its, think of, fear, i guess	like a, fear, job, things, always	you know, you, thats, probably, them
Top 5 features for second group (SVM weights)	little, for, have a, how, now	sometimes, things, you know, you, they are	wrong, a lot, makes me, feel like, really	better, it like, being angry, right, i would

TABLE 4.

Classifier performance for identifying individual disorders. The number of features used for classification was a parameter tuned separately from the other hyperparameters. The AROC is the Area under the Receiver Operating Characteristic curve. The *P*-values represent the significance of the AROCs compared to a completely random classifier. (a) Leave-one-out (LOO) cross-validation performance using the first interviews with all participants. The feature selection for the cross-validation was performed within each fold; however, the "top features" presented here use the same method applied to the entire dataset. (b) Training sets constructed from stratified random samples of 50% of first interviews and 50% of second interviews with performance tested against all remaining interviews

(a)				
	Suicidality (n = 30/122)	**Despressive Disorders s (n = 29/122)**	**Anxiety Disorders (n = 15/122)**	**Bipolar Disorders (n = 6/122)**
AROC (95% CI)	57 (44-71)%	65 (54-76)%	66 (52-80)%	78 (64-92)%
P-value	.22	.011	.04	.003
Total No. Features	16	512	32	64
Top 5 features for presence of comorbidity (SVM weights),	try to, sometimes, things, I do not really, my	its, is, yes, like i, hope	I feel, yes, myself, my, try to	it is, be, weird, now, things
Top 5 features for absence of comorbidity (SVM weights)	I have, in, try, assent (LIWC), inhibit (LIWC)	fears, if i, on, get, know	no, and I, to, that I, feel	get, angry, for, think, so
(b)				
	Suicidality (n = 26/114)	**Depressive Disorders (n = 25/113)**	**Anxiety Disorders (n = 12/112)**	**Bipolar Disorders (n = 4/114)**
AROC (95% CI)	71 (59-83)%	69 (58-81)%	64 (46-81)%	66 (34-99)%
P-value	.0011	.001	.12	.3
Total No. Features	64	128	256	32
Top 5 features for presence of comorbidity (SVM weights)	like a, fear, job, things, always	uh, like that, in, that i, was	kinda, going, no, more, you	time, i just, or, pretty, thats
Top 5 features for absence of comorbidity (SVM weights)	wrong, a lot, makes me, feel like, really	i am, she, at, talk to, come	myself, most, lot of, a lot of, of	secrets, give, anything to, to explain, thats really