BRIEF RESEARCH REPORT
published: 02 February 2022
doi: 10.3389/fdgth.2022.818705

Integration and Validation of a Natural Language Processing Machine Learning Suicide Risk Prediction Model Based on Open-Ended Interview Language in the Emergency Department

Joshua Cohen[1]*, *Jennifer Wright-Berryman*[2], *Lesley Rohlfs*[1], *Douglas Trocinski*[3], *LaMonica Daniel*[4] and *Thomas W. Klatt*[5]

[1] Clarigent Health, Mason, OH, United States, [2] Department of Social Work, College of Allied Health Sciences, University of Cincinnati, Cincinnati, OH, United States, [3] WPP Emergency Services, Raleigh, NC, United States, [4] WPP Clinical Research, Raleigh, NC, United States, [5] Behavioral Health Network, Raleigh, NC, United States

OPEN ACCESS

Edited by:
Lasse B. Sander,
University of Freiburg, Germany

Reviewed by:
Yannik Terhorst,
University of Ulm, Germany
Mike Conway,
The University of Utah, United States

***Correspondence:**
Joshua Cohen
jcohen@clarigenthealth.com

Specialty section:
This article was submitted to
Digital Mental Health,
a section of the journal
Frontiers in Digital Health

Received: 19 November 2021
Accepted: 10 January 2022
Published: 02 February 2022

Citation:
Cohen J, Wright-Berryman J, Rohlfs L, Trocinski D, Daniel L and Klatt TW (2022) Integration and Validation of a Natural Language Processing Machine Learning Suicide Risk Prediction Model Based on Open-Ended Interview Language in the Emergency Department. Front. Digit. Health 4:818705. doi: 10.3389/fdgth.2022.818705

Background: Emergency departments (ED) are an important intercept point for identifying suicide risk and connecting patients to care, however, more innovative, person-centered screening tools are needed. Natural language processing (NLP) -based machine learning (ML) techniques have shown promise to assess suicide risk, although whether NLP models perform well in differing geographic regions, at different time periods, or after large-scale events such as the COVID-19 pandemic is unknown.

Objective: To evaluate the performance of an NLP/ML suicide risk prediction model on newly collected language from the Southeastern United States using models previously tested on language collected in the Midwestern US.

Method: 37 Suicidal and 33 non-suicidal patients from two EDs were interviewed to test a previously developed suicide risk prediction NLP/ML model. Model performance was evaluated with the area under the receiver operating characteristic curve (AUC) and Brier scores.

Results: NLP/ML models performed with an AUC of 0.81 (95% CI: 0.71–0.91) and Brier score of 0.23.

Conclusion: The language-based suicide risk model performed with good discrimination when identifying the language of suicidal patients from a different part of the US and at a later time period than when the model was originally developed and trained.

Keywords: suicide, machine learning, natural language processing, emergency department (ED), risk assessment, mental health, validation, feasibility & acceptability

INTRODUCTION

Suicide remains the 10th leading cause of death in the United States overall (1). Although recent data suggest that suicide rates from 2018 to 2019 decreased 2.1%, 12 million adults reported suicidal thoughts, 3.5 million reported a suicide plan, and 1.4 million reported a suicide attempt (1). A report of the Joint Commission of the Accreditation of Healthcare Organizations (2) revealed that suicide screening in emergency departments (ED) may be an important intercept point for identifying those at risk and connecting them to care, however, current screening tools may not provide person-centered risk identification with consistent sensitivity (3). Therefore, innovative and person-centered methods of identifying suicide risk are a critical need.

Despite decades of research, suicide rates have steadily risen, while detection of who is at highest risk for death has not improved (4, 5). A recent meta-analysis suggests that current methods of predicting risk for suicide death are no better than 50% or random chance (6, 7). Although theories exist about why people die by suicide (8–11) they have not aided the development of adequate predictive models for reducing death rates, nor have they resulted in screening instruments that have sufficient predictive value (7, 12).

Natural language processing (NLP) and machine learning (ML) have shown promise in identifying suicide risk (13). However, many of these NLP methods have been applied to analyzing social media texts or in non-clinical settings (14, 15). Additionally, NLP methods that have been used with clinically significant language data have been gathered from existing records (16, 17) instead of using interviews at point-of-contact. A recent meta-analysis suggests that NLP holds promise for accurate detection of suicide risk when deployed within the assessment process (18), as most studies showed an AUC >0.90. However, the authors assert more work should be done using real-time language data collection for translation into clinical practice. A study by Chakravarthula et al. used NLP with military couples to identify suicide risk by analyzing couples' conversations (19). The investigators found that the model predicted risk better than random chance in all risk categories. A 2021 study using smartphones to collect language data from 588 veterans also successfully determined suicide risk with 0.86 sensitivity, 0.70 specificity and an AUC >0.80 (20).

Previous research of the technology used in this study was conducted primarily in the Midwestern United States from 2013 to 2015 by Pestian et al., and used language collected through interviews of suicidal and non-suicidal adults and adolescents to "train" an NLP/ML model to identify language features of suicidal individuals (21, 22). In these studies, called the Adolescent Controlled Trial (ACT) and Suicide Thought Markers (STM) study, a total of 160 suicidal participants, 126 non-suicidal participants with mental illness, and 153 non-suicidal participants without mental illness were enrolled across three hospitals' EDs and psychiatric units. Support vector machine (SVM) models were trained to identify suicidal participants (case) vs. non-suicidal participants with and without mental illness (control) with an area under the receiver operating characteristic curve (AUC) of 0.69–0.93 depending on the cross-validation approach used (21, 22). While these models have performed well-across a series of studies, many factors could influence language potentially affecting model generalizability and validation when applied at later time periods or in separate geographic regions.

Validation is one of the most important steps during model development and provides evidence a model will perform as expected with a new dataset and in new settings. Broadly, internal validation tests a model with the development data through different protocols such as cross-validation, where data are segmented into groups or "folds" and a model is iteratively trained on all but one fold, with performance examined on the hold-out. Internal validation procedures may produce accurate models and demonstrate a proof of concept, however, if a model is to be used in more critical settings such as medical or psychiatric settings, additional testing should be done. External validation tests a model with data collected separate from the data used to develop the model, where one or more variables (e.g., time point or location) are changed to ensure model performance remains acceptable (23, 24). Additionally, clincal NLP models pose unique challenges to algorithm portability – the ability to apply a model in diverse settings – due to the need to assemble clinical corpora, site-specific reporting structures, and the idiosyncrasies of language use (25, 26), underscoring the need for external validation.

The purpose of this pilot study was to (1) determine if the interview process to collect language for an NLP/ML model could be integrated into two EDs in the Southeastern United States, and (2) evaluate model performance on language from persons in a different geographic region than where the original model was developed. Notably, this study was conducted during the COVID-19 pandemic, which has led to increased stress and isolation and has likely impacted language use (27, 28). External validation is especially important during COVID-19 for a spoken language-based model because the data used to develop our model was collected prior to the pandemic, meaning the model cannot recognize the pandemic as a factor impacting suicidal risk. We do not know when the pandemic will end and therefore, must understand any limitations of these methods in this setting.

METHODS

The protocol was approved by the health system's institutional review board which oversaw the study conducted at the two EDs. All subjects gave informed consent in accordance with the Declaration of Helsinki before they participated in the study.

Study Staff and Participants

Study Staff

The study staff was composed of four Behavioral Health Licensed Clinicians (BHLC) for case participants and two clinical research coordinators (CRC) for control participants. Study staff completed online training to learn study procedures, principles of human subject protection, and good clinical practice.

Participants and Case-Control Definitions

Criteria for participant recruitment were: (1) a patient receiving ED services, (2) age 18–65, (3) able to provide

Abbreviations: ACT, Adolescent Controlled Trial (22); AUC, area under receiver operating characteristic curve; BHLC, Behavioral Health Licensed Clinician; C-SSRS, Columbia - Suicide Severity Rating Scale; CI, Confidence Interval; CRC, Clinical Research Coordinator; ED, Emergency Department; MHSAFE, Mental Health Hopes Secrets Anger Fear and Emotional Pain; ML, Machine Learning; NLP, Natural Language Processing; STM, Suicidal Thought Markers study (21); SVM, Support Vector Machine.

informed consent, and (4) English as a primary language. Case participants presented to an ED with suicidal ideation or a suicide attempt within the last 30 h. Control participants presented to an ED for any non-suicide related condition and had no lifetime history of suicide risk or mental health diagnosis.

Study Design

This case-control study sought to externally validate an NLP/ML model trained on data collected from the ACT and STM studies, and therefore aimed to keep procedures as consistent as possible between those studies. Because there is no generally accepted approach to estimate sample sizes for development and validation of risk prediction models (23), sample size was determined from previous studies and feasibility considerations. Previous studies enrolled 30–44 case and control participants per site, which allowed performance estimates with acceptable precision (21, 22).

Upon admission to the ED, patients were invited to participate in the study and the C-SSRS Screener was administered (see **Figure 1**). The C-SSRS Screener is a structured interview based on the full-length version (29), and is designed to measure suicidal ideation and suicidal behaviors on an ordinal scale. In a multi-site emergency department study, the C-SSRS demonstrated high sensitivity and specificity for classifying suicidal behavior, as well as strong internal consistency and predictive validity (11). The C-SSRS Screener is composed of two initial questions and four follow up questions if the participant answers "yes" to question 2. The first five questions relate to the severity of suicidal ideation and ask about the past month. These questions include: (1) "Have you wished you were dead or wished you could go to sleep and not wake up?" (2) "Have you actually had any thoughts of killing yourself?" (3) "Have you been thinking about how you might do this?" (4) "Have you had these thoughts and had some intention of acting on them?" and (5) "Have you started to work out or worked out the details of how to kill yourself? Do you intend to carry out this plan?" The final question measures the presence of suicidal behavior over the lifetime and the past 3 months: (6) "Have you ever done anything, started to do anything, or prepared to do anything to end your life?"

The study staff then conducted a semi-structured interview comprised of pre-established probes (renamed MHSAFE - hope, secrets, anger, fear, and emotional pain - from the "Ubiquitous Questionnaire"), and used an application installed on a tablet to record the interview. The original Ubiquitous Questionnaire was developed with experts in suicide research in order to elicit emotional language related to salient variables of hope, secrets, anger, fear, and emotional pain (30). Study staff were instructed to ask questions about participants' current feelings on all five areas, for example, "tell me about your hope?" or "are there secrets in your life?" The interviews in previous research lasted 8.1 ± 4.5 min (21). A HIPAA- compliant service was used to manually transcribe the recorded interviews. Scores from the suicide risk model were not returned to clinicians in this study because they were generated asynchronously from manual transcripts after care determinations had been made, and because the purpose of this study was to evaluate model validity.

Data Analysis

All analysis was performed using the Python programming language [version 3.7.5; (31)]. The open-source Python libraries Pandas [version 1.1.2; (32, 33)], Numpy [version 1.18.5; (34, 35)], scikit-learn [version 0.23.2; (36)], Matplotlib [version 3.7.5; (37)], SciPy [version 1.5.2; (38)], and NLTK [version 3.2.2; (39)] were used for data analysis and all NLP/ML model building. Student's *t*-tests were performed with SciPy's ttest_ind function. The study adhered to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement for reporting (23). **Figure 1** shows model development and validation procedures.

The NLP/ML pipeline used in this study followed similar techniques used by Pestian et al., focused on the term frequency of n-grams (contiguous sequence of n number of words) and SVMs (21, 22, 40–42). The Porter Stemmer algorithm was applied to participant language to normalize morphologically related terms (43). Language was tokenized by splitting on white spaces, and 1–3 continuous sequences of words were used as features. Scikit-learn's SelectKBest function was used to identify features with the highest ANOVA F-value, with the number of features selected as a tunable hyperparameter to optimize model performance, including 1024, 2048, 4096, and "all" features. Selected features were then scaled to unit variance. During SVM tuning, hyperparameters considered include: the regularization parameter (C), the kernel (radial basis function and linear kernels), the kernel coefficient (gamma, if applicable), and the class weight. Additional details on NLP/ML methods may be found in previous work (42). We have evaluated the performance of Logistic Regression and XGBoost models in previous work and found comparable performance across models (42), and therefore decided to continue with the SVM used previously (21). A comparison of different models is beyond the scope of this paper.

During model training using ACT and STM data, the only input was the participant's language, labeled as case or control. During model testing using data from this study, a participant's language was fed into the model, and a probability for belonging to the case group was returned. Model performance was then evaluated by comparing model predictions to the participant's labeled group (case or control) and calculating the AUC and Brier score, two recommended measures to provide a more complete picture of model performance (23). AUC values range from 0.5 (random chance) to 1.0 (perfect model). The Brier score is a measure of model calibration and ranges from 0 to 1, where low scores indicate less discrepancy between predicted probabilities and outcomes. If a model is calibrated, then its output probabilities convey meaningful information. For example, a calibrated model that returns a 30% chance of having a disease means of all tests that received a score of about 30%, 30% of them had the disease (44). Thus, the AUC is a preferred metric for model evaluation because it gives an overall measure of model discrimination without imposing

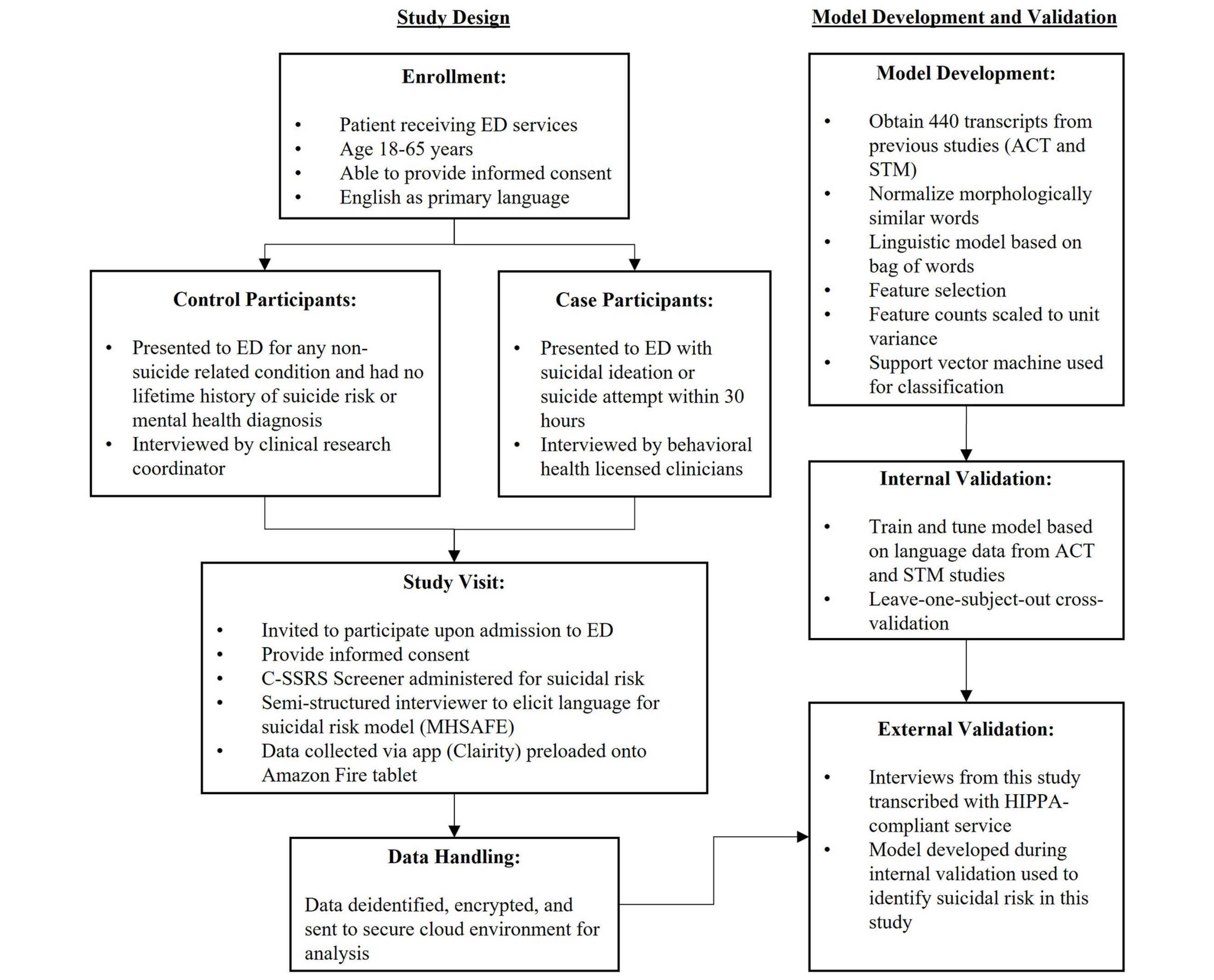
Study Design
Model Development and Validation
Enrollment:
• Patient receiving ED services
• Age 18-65 years
• Able to provide informed consent
• English as primary language
Control Participants:
• Presented to ED for any non-suicide related condition and had no lifetime history of suicide risk or mental health diagnosis
• Interviewed by clinical research coordinator
Case Participants:
• Presented to ED with suicidal ideation or suicide attempt within 30 hours
• Interviewed by behavioral health licensed clinicians
Study Visit:
• Invited to participate upon admission to ED
• Provide informed consent
• C-SSRS Screener administered for suicidal risk
• Semi-structured interviewer to elicit language for suicidal risk model (MHSAFE)
• Data collected via app (Clairity) preloaded onto Amazon Fire tablet
Data Handling:
Data deidentified, encrypted, and sent to secure cloud environment for analysis
Model Development:
• Obtain 440 transcripts from previous studies (ACT and STM)
• Normalize morphologically similar words
• Linguistic model based on bag of words
• Feature selection
• Feature counts scaled to unit variance
• Support vector machine used for classification
Internal Validation:
• Train and tune model based on language data from ACT and STM studies
• Leave-one-subject-out cross-validation
External Validation:
• Interviews from this study transcribed with HIPPA-compliant service
• Model developed during internal validation used to identify suicidal risk in this study

FIGURE 1 | Schematic of study and modeling procedures.

probability thresholds for classification that can result in a loss of information (23, 45).

Additional classification metrics were calculated for different probability thresholds to classify a new interview as positive (suicidal). Metrics considered include sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Sensitivity and specificity measure how well a test identifies true positives and true negatives, respectively. PPV measures the proportion of true positives out of all who test positive, and NPV measures the proportion of true negatives out of all who test negative.

Internal Validation

Data from the ACT and STM studies have been internally validated in separate publications and report AUCS from 0.69 to 0.93 depending on the features (acoustic or linguistic), participants included (controls, those with mental illness not suicidal, and suicidal), and cross-validation technique (leave-one-subject-out and leave-one-site-out) (21, 22). For internal validation in this study, we performed a leave-one-subject-out cross-validation for the combined ACT and STM dataset comprised of 440 participant interviews using only linguistic data. This cross-validation technique provides the most optimistic internal validation performance by iteratively training a model on all but one interview and predicting to which class the hold-out interview belongs. We focused exclusively on linguistic features because previous work compared linguistic and acoustic (e.g., fundamental frequency) features and found acoustic features did not improve predictive value for this dataset (21).

External Validation

For external validation, a model was trained and tuned on the complete ACT and STM dataset (controls, suicidal, and mentally ill), and then used to predict suicidal risk from the language samples collected in this study. The suicidal risk for model performance was determined by the study arm to which the participant was assigned. For case participants, this suicidal risk can be characterized by recent suicide death-related ideations or behaviors that lead to their admission to the ED. Language samples were collected within 30 h of admission.

RESULTS

Population and Data Collection

Between September and December 2020, 70 participants were enrolled, each providing one session recording. The control arm of the study interviewed 33 individuals and the case arm interviewed 37. **Table 1** shows a summary of participant descriptive statistics. A significant difference in participant word count between case and control interviews is shown in **Table 1** (t-test, $p = 0.02$), although this difference is not present in our training data ($p = 0.77$).

All control participants responded negatively to all questions on the C-SSRS Screener and confirmed no history of suicidality or a mental illness diagnosis. **Table 2** describes how participants in the case group answered the C-SSRS Screener. One case participant answered negatively to all questions on the C-SSRS Screener. This participant was still considered a case in this study because they were admitted to the ED for suicidal ideation with a diagnosis for suicidal ideation.

Internal Validation

The NLP/ML model reached an AUC of 0.82 (95% CI = 0.78–0.86) and a Brier score of 0.16 using a leave-one-subject-out cross-validation technique for the 440 subject interviews in the ACT and STM training dataset. This model performed optimally with 4096 n-gram features and a linear kernel. **Table 3** shows classification performance metrics for different probability risk thresholds for a positive (suicidal) prediction. The top ten model features for case and control predictions are available in **Supplementary Table 1**.

External Validation

The NLP/ML model trained and tuned on the ACT and STM dataset reached an AUC of 0.81 (95% CI = 0.71–0.91) and a Brier score of 0.23 when predicting suicidal risk on the 70 patient interviews collected in this study. **Table 3** shows classification performance metrics for different probability risk thresholds for a positive (suicidal) prediction. Receiver operating characteristic curves for internal and external validation are available as **Supplementary Figure 1**.

DISCUSSION

In this study, we found it feasible to integrate technology and procedures to collect language for a suicide risk prediction model into the ED workflow. Additionally, a follow-up study using thematic analysis of BHLC semi-structured interviews (46) revealed that minor issues such as logging into the system were easily overcome, and the use of the app to consent the participant and record the session were well-managed in the workflow. The clinicians also reported that in comparison to standardized scales (screening as usual), use of the probes did not impede the ED process, gleaned more information about the person's mental state, and the probes were reported as a more person-centered approach to screening for suicide risk. Although a five-to-ten-minute interview may take longer than a self-report brief scale, more usable data for clinical decision-making were reportedly obtained via the probes (46).

The model, trained on language from the ACT and STM studies, performed with good discrimination when identifying the language of suicidal vs. non-suicidal participants in this study. As a variable, language is influenced by time period, location, and large-scale events, such as the COVID-19 pandemic (28). Therefore, we were uncertain how well a language-based model trained on data collected in the Midwest from 2013 to 2015 would perform on the language collected in this study. Interestingly, despite these factors, model performance was similar to previous studies [AUC range 0.69–0.93; (21, 22)], supporting geographic and temporal validation.

TABLE 1 | Participant descriptive statistics.

Variable	Control	Case	Both
Enrolled	33	37	70
Average age (SD)	41.2 (12.5)	41.1 (12.8)	40.1 (12.5)
Gender			
Male (%)	18 (54.50%)	20 (54.10%)	38 (54.30%)
Female (%)	15 (45.50%)	16 (43.20%)	31 (44.30%)
Transgender (%)	0 (0%)	1 (2.70%)	1 (1.40%)
Race			
White or Caucasian (%)	15 (45.50%)	20 (54.10%)	35 (50.00%)
Black or African American (%)	15 (45.50%)	17 (45.90%)	32 (45.70%)
Other (%)	3 (9.10%)	0 (0%)	3 (4.30%)
Average interview length (min) (SD)	7.8 (3.1)	7.1 (3.1)	7.4 (3.1)
Average participant word count (SD)	723 (401)	485 (432)	593 (431)

TABLE 2 | Summary of case participant answers to the C-SSRS screener.

C-SSRS question	Question topic	N	%
Past month			
-	No suicidal ideation (SI)	1	3%
1	Wish to be dead	35	95%
2	Non-specific active SI	32	86%
3	SI with methods	26	70%
4	Suicidal intent	24	65%
5	Suicidal intent with plan	24	65%
Lifetime			
6a	Suicidal behavior	26	70%
Past 3 months			
6b	Suicidal behavior	15	41%

TABLE 3 | Internal and external validation classification performance at different risk thresholds.

Risk threshold[a]	Sensitivity[b] (95% CI)	Specificity[c] (95% CI)	PPV[d] (95% CI)	NPV[e] (95% CI)
Internal validation				
≥10%	0.93 (0.88–0.96)	0.39 (0.33–0.45)	0.47 (0.41–0.52)	0.91 (0.84–0.95)
≥20%	0.85 (0.79–0.90)	0.57 (0.51–0.62)	0.53 (0.47–0.59)	0.87 (0.81–0.91)
≥35%	0.73 (0.66–0.80)	0.72 (0.67–0.77)	0.60 (0.53–0.67)	0.82 (077–0.87)
≥50%	0.63 (0.55–0.70)	0.84 (0.79–0.87)	0.69 (0.61–0.76)	0.80 (0.75–0.84)
External validation				
≥10%	0.81 (0.66–0.91)	0.55 (0.38–0.70)	0.67 (0.52–0.79)	0.72 (0.52–0.86)
≥20%	0.73 (0.57–0.85)	0.76 (0.59–0.87)	0.77 (0.61–0.88)	0.71 (0.55–0.84)
≥35%	0.65 (0.49–0.78)	0.88 (0.73–0.95)	0.86 (0.69–0.94)	0.69 (0.54–0.81)
≥50%	0.54 (0.38–0.69)	0.94 (0.80–0.98)	0.91 (0.72–0.97)	0.65 (0.50–0.77)

[a]Model scores equal to or above this value are classified as suicidal. [b]Sensitivity = true positives divided by sum of true positives and false negatives. [c]Specificity = true negatives divided by sum of true negatives and false positives. [d]Positive predictive value = true positives divided by sum of true positives and false positives. [e]Negative predictive value = true negatives divided by sum of true negatives and false negatives.

As a performance metric, AUC may be interpreted as the probability randomly selected case (suicidal) participants will receive a higher probability score from the model than randomly selected control (non-suicidal) participants (47). The classification metrics in **Table 3** require defining a threshold for classification, although as mentioned, setting a risk threshold for classification may result in the loss of clinically relevant information if a model is calibrated. Brier scores were used to measure calibration in this study, and we found the Brier score increased 0.07 during external validation, indicating the model may be less calibrated in this setting. Indeed, during internal validation, sensitivity and specificity are roughly equal at a risk threshold of 35% but are equal at a lower risk threshold of 21% for external validation, indicating interviews in this study receive a lower probability of being suicidal, overall. Thus, while model performance as measured by the AUC is similar to the internal validation, it may be that models require additional calibration when applied in new settings.

While a change in geography likely impacts the need for additional calibration, we expect temporal effects play a larger role on the model's output. Data drift- the degradation of model performance over time, is a common concern when developing models, and is due in part to a change in relation between features (e.g., language) and the outcome of interest (e.g., suicidal risk). Broadly, the language features in the model may be categorized as content words (i.e., what someone talks about) and function words (i.e., how someone speaks). While the underlying concept behind the MHAFE interview is to elicit language about universally relatable topics (30), and some content and function word usage might be time-agnostic, participants in this study discussed topics that did not exist in 2015, such as the COVID-19 pandemic. Therefore, the model was unaware of additional factors that could contribute to an individual's suicidal risk, potentially lowering the model's score.

More advanced natural language processing techniques may offer a solution to generalize over broader geographical and temporal settings. Word vectors encode words into a high dimensional space (50–300 dimensions) that retain semantic meaning and have demonstrated state-of-the-art performance on many language tasks (48–50). However, the semantic meaning encoded in word vectors is derived from specific corpora (e.g., all of Wikipedia) and in many cases has been found to also retain biases (50–52). Additionally, models using word vectors may struggle to explain the reason for a specific prediction, which is becoming required for clinical decision support systems (53, 54). For these reasons, we have focused primarily on more traditional NLP/ML techniques but will explore more advanced techniques as more tools for explainable AI and AI safety become available.

Nevertheless, as with all screening and diagnostic tools, there will always be variables that can impact results, and how a tool was developed should be considered when interpreting results. For example, there have been gender imbalances in clinical research (55, 56), so caution must be taken when generalizing results across the underrepresented gender. Similarly, clinicians using a language-based tool to identify suicide risk should be trained to understand its limitations, such as this model's inability to recognize language related to COVID-19. We have been collecting virtual interviews throughout the pandemic with participants from every United States geographical region to update our model and better understand other potential limitations related to geo-temporal effects.

Traditional theories to explain suicide follow an ideation-to-action framework, where there is a linear progression from suicidal ideation to suicidal behavior. Recent work, however, has highlighted these theories of suicide are incomplete (7, 57, 58). A meta-analysis of 71 studies examining the relation between suicide ideation and later suicide found ∼60% of individuals who either engaged in suicidal behaviors or later died by suicide did not express suicidal ideation (58). These findings have led to the suggestion of multiple pathways that can lead to suicidal behavior, one of which includes a subgroup of individuals who do not first experience active suicidal ideation before suicidal behavior (57). Therefore, suicide screening tools built upon ideation-to-action theories, such as the C-SSRS Screener used in this study, likely miss a significant portion of individuals who may later die by suicide. Longitudinal studies could help us understand if patient language may help in identifying these false negatives of traditional scales.

Clinically, the results of this study suggest that the MHSAFE process for screening suicide risk may add to the limited options available to providers. In this ED study, the BHLC reported ease of integration of the technology, and that this qualitative, person-centered, interview approach to suicide risk screening has added clinical benefit, such as allowing the patient to open up and provide more information and nuance (46). Future versions of the tool will provide a return of suicide risk result that can be used in clinical decisions (further assessment, safety planning).

Limitations and Future Directions

Although these findings align with previous studies, some limitations should be noted. First, the use of BHLCs for case interviews and CRCs for control interviews could have influenced the model's ability to classify participant interviews. All study staff received the same training on the MHSAFE interview, but individual interviewer styles, previous experience, and being unblinded to outcomes could affect interviews. This study sought to externally validate a model developed on data from the ACT and STM studies, where CRCs interviewed both case and control participants. The consistency of model performance across the internal and external validation sets suggests the interviewer does not play a significant role in the model's ability to identify the language of suicidal individuals. Notably, the present approach is more reflective of potential real-world applications, supporting ecological validity.

Additional limitations include small sample size, which limited representation of gender and race, which may impact generalizability. Lastly, the use of volunteer participants could have biased clinician experience and model performance.

Future studies will analyze how a return of model results to clinicians impacts clinical decision making and risk-treatment alignment. We are also investigating techniques to autonomously conduct the MHSAFE interview using an automated voice response system. Lastly, this model was developed on US English, and future studies will focus on the model's validity with different English dialects as well as the development of models for different languages.

CONCLUSIONS

A brief interview can be successfully implemented into two EDs and NLP/ML models can predict suicide risk from patient language with good discrimination. A strength of this study is that it was conducted in a separate geographic region and at a later time period, supporting generalizability. Screening for suicide can be an open-ended and dynamic process, and these findings have highlighted an opportunity for identifying suicide risk using a person's language.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the dataset contains confidential health-related data that cannot be shared. These data will be made available for research purposes only to any researcher who meet criteria for access to confidential data based on relevant Institutional Review Boards. Requests to access the datasets should be directed to research@clarigenthealth.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by WakeMed Health & Hospitals' Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JC and JW-B wrote the manuscript. JC performed statistical analysis on the corpora. DT is the principal investigator of the WakeMed Study of Suicidal Language in the ED-Pilot. LR, LD, and TK supervised the clinical research. LR provided project administration. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by Clarigent Health.

ACKNOWLEDGMENTS

The authors would like to thank the Behavioral Health Licensed Clincians, Clinical Research Coordinators, and medical staff at WakeMed Health & Hospitals where this study was conducted for their support, Marci Campbell for her support with study development and implementation, and the Community of Scientists and Researchers that provide open-source machine learning tools and education resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.818705/full#supplementary-material

Supplementary Figure 1 | Receiver operating characteristic curve for internal and external validation.

Supplementary Table 1 | Top 10 model features by feature weight. Feature importance was determined from the coefficients of a linear SVM kernel, trained on the ACT and STM dataset. Case features influence the model toward a case prediction, while control features influence the model toward a control prediction. The features are n-grams, and morphologically similar terms have been normalized using the Porter Stemmer algorithm (e.g., "alone" to "alon"). This model has a total of 4,096 features.

REFERENCES

1. Stone DM, Jones CM, Mack KA. Changes in suicide rates — United States, 2018–2019. *MMWR Morb Mortal Wkly Rep*. (2021) 70:261–8. doi: 10.15585/mmwr.mm7008a1
2. Roaten K, Johnson C, Genzel R, Khan F, North CS. Development and implementation of a universal suicide risk screening program in a safety-net hospital system. *Jt Comm J Qual Patient Saf.* (2018) 44:4–11. doi: 10.1016/j.jcjq.2017.07.006
3. Simpson SA, Goans C, Loh R, Ryall K, Middleton MCA, Dalton A. Suicidal ideation is insensitive to suicide risk after emergency department discharge: performance characteristics of the columbia-suicide severity rating scale screener. *Acad Emerg Med*. (2020) 28:621–9. doi: 10.1111/acem.14198
4. WISQARS (Web-based Injury Statistics Query and Reporting System), *National Center for Injury Prevention and Control, Centers for Disease Control and Prevention*. Available online at: https://www.cdc.gov/injury/wisqars/index.html [accessed December 28, 2021].
5. Nock MK, Borges G, Bromet EJ, Cha CB, Kessler RC, Lee S. Suicide and suicidal behavior. *Epidemiol Rev.* (2008) 30:133–54. doi: 10.1093/epirev/mxn002
6. Bentley KH, Franklin JC, Ribeiro JD, Kleiman EM, Fox KR, Nock MK. Anxiety and its disorders as risk factors for suicidal thoughts and behaviors: a meta-analytic review. *Clin Psychol Rev.* (2016) 43:30–46. doi: 10.1016/j.cpr.2015.11.008
7. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull.* (2017) 143:187–232. doi: 10.1037/bul0000084
8. Durkheim E. *Le suicide; étude de sociologie*. Paris: F Alcan (1897).
9. Joiner TE. *Why People Die by Suicide*. Cambridge, MA: Harvard University Press (2007).
10. Schneidman ES. *Suicide as Psychache: A Clinical Approach to Self-Destructive Behavior*. Available online at: https://psycnet.apa.org/record/1993-98267-000 [accessed December 28, 2021].
11. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TE. The interpersonal theory of suicide. *Psychol Rev.* (2010) 117:575. doi: 10.1037/a0018697
12. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: applications and ethics. *Behav Sci Law.* (2019) 37:214–22. doi: 10.1002/bsl.2392
13. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: A scoping review of methods and applications. *Psychol Med.* (2019) 49:1426–48. doi: 10.1017/S0033291719000151
14. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. *Nat Lang Eng.* (2017) 23:649–85. doi: 10.1017/S1351324916000383
15. Burnap P, Colombo G, Scourfield J. Machine classification and analysis of suicide-related communication on Twitter. In; *HT 2015 - Proceedings of the 26th ACM Conference on Hypertext and Social Media* (Association for Computing Machinery, Inc) (2015).
16. Oseguera O, Rinaldi A, Tuazon J, Cruz AC. Automatic quantification of the veracity of suicidal ideation in counseling transcripts. *Commun Comput Inf Sci.* (2017) 713:473–9. doi: 10.1007/978-3-319-58750-9_66
17. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry.* (2017) 174:154–62. doi: 10.1176/appi.ajp.2016.16010077
18. Bernert RA, Hilberg AM, Melia R, Kim JP, Shah NH, Abnousi F. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *Int J Environ Res Public Health.* (2020) 17:5929. doi: 10.3390/ijerph17165929
19. Chakravarthula SN, Nasir M, Tseng SY, Li H, Park TJ, Baucom B, et al. Automatic prediction of suicidal risk in military couples using multimodal interaction cues from couples conversations. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE (2019), 6539–43. doi: 10.1109/ICASSP40776.2020.9053246
20. Belouali A, Gupta S, Sourirajan V, Yu J, Allen N, Alaoui A, et al. Acoustic and language analysis of speech for suicidal ideation among US veterans. *BioData Min.* (2021) 14:1–17. doi: 10.1186/s13040-021-00245-y
21. Pestian JP, Sorter M, Connolly B, Bretonnel Cohen K, McCullumsmith C, Gee JT, et al. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. *Suicide Life Threat Behav.* (2017) 47:112–21. doi: 10.1111/sltb.12312
22. Pestian JP, Grupp-Phelan J, Bretonnel Cohen K, Meyers G, Richey LA, Matykiewicz P, et al. Controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. *Suicide Life Threat Behav.* (2016) 46:154–9. doi: 10.1111/sltb.12180
23. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* (2015) 162:1–73. doi: 10.7326/M14-0698
24. Cabitza F, Campagner A, Soares F, García De Guadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed*. (2021) 208:106288. doi: 10.1016/j.cmpb.2021.106288
25. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc.* (2017) 24:986. doi: 10.1093/jamia/ocx039
26. Adekkanattu P, Jiang G, Luo Y, Kingsbury PR, Xu Z, Rasmussen LV, et al. Evaluating the portability of an NLP system for processing echocardiograms: a retrospective, multi-site observational study. *AMIA Annu Symp Proc*. (2019) 2019:190.
27. Rothman S, Gunturu S, Korenis P. The mental health impact of the COVID-19 epidemic on immigrants and racial and ethnic minorities. *QJM.* (2020) 113:779–82. doi: 10.1093/qjmed/hcaa203
28. Pennebaker JW. The secret life of pronouns. *New Sci.* (2011)211:42–5. doi: 10.1016/S0262-4079(11)62167-2
29. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am J Psychiatry.* (2011) 168:1266–77. doi: 10.1176/appi.ajp.2011.10111704
30. Pestian J. A conversation with edwin shneidman. *Suicide Life Threat Behav.* (2010) 40:516–23. doi: 10.1521/suli.2010.40.5.516
31. van Rossum G. *Python Tutorial*. Amsterdam, Netherlands: Stichting Mathematisch Centrum (1995). Available online at: http://oai.cwi.nl/oai/asset/5007/05007D.pdf (accessed December 28, 2021).
32. The Pandas Development Team Pandas. Dev/Pandas: Pandas 1.1.2. (2020). Available online at: https://zenodo.org/record/4019559/export/xm#.X6O3z1BS9PY (accessed December 28, 2021).
33. McKinney W. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. Austin, TX: SCIPY (2010), 56–61.
34. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng*. (2011) 13:22–30. doi: 10.1109/MCSE.2011.37
35. Oliphant TE. Python for scientific computing. *Comput Sci Eng.* (2007) 9:10–20. doi: 10.1109/MCSE.2007.58
36. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. *Scikit-learn: Machine Learning in Python*. (2011). Available online at: http://scikit-learn.sourceforge.net [accessed May 21, 2020].
37. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng.* (2007) 9:99–104. doi: 10.1109/MCSE.2007.55
38. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 10: fundamental algorithms for scientific computing in Python. *Nat Methods.* (2020) 17:261–72. doi: 10.1038/s41592-020-0772-5
39. Bird S, Klein E, Loper E. NLTK Book. *O'Reilly.* (2016). Available online at: http://www.nltk.org/book_1ed/ [accessed September 14, 2020].
40. Pestian J, Santel D, Sorter M, Bayram U, Connolly B, Glauser T, et al. A Machine learning approach to identifying changes in suicidal language. *Suicide Life Threat Behav*. (2020) 50:sltb.12642. doi: 10.1111/sltb.12642
41. Glauser T, Santel D, DelBello M, Faist R, Toon T, Clark P, et al. Identifying epilepsy psychiatric comorbidities with machine learning. *Acta Neurol Scand.* (2020) 141:388–96. doi: 10.1111/ane.13216

42. Cohen J, Wright-Berryman J, Rohlfs L, Wright D, Campbell M, Gingrich D, et al. feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. *Int J Environ Res Public Health.* (2020) 17:1–17. doi: 10.3390/ijerph17218187
43. Porter MF. An algorithm for suffix stripping. *Program.* (1980) 14:130–7. doi: 10.1108/eb046814
44. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci.* (2017) 5:457–69. doi: 10.1177/2167702617691560
45. Venkatesh KK, Strauss RA, Grotegut CA, Heine RP, Chescheir NC, Stringer JSA, et al. Machine learning and statistical models to predict postpartum hemorrhage. *Obstet Gynecol.* (2020) 135:935–44. doi: 10.1097/AOG.0000000000003759
46. Pease J, Thompson D, Wright-Berryman J, Campbell M. User feedback on the use of a natural language processing application to identify suicide risk in the emergency department. (submitted).
47. Youngstrom EA. A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: we are ready to ROC. *J Pediatr Psychol.* (2014) 39:204–21. doi: 10.1093/jpepsy/jst062
48. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In; *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* (International Conference on Learning Representations, ICLR).
49. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014). Available online at: https://aclanthology.org/D14-1162/
50. Bhardwaj R, Majumder N, Poria S. Investigating gender bias in BERT. *Cogn Comput.* (2021) 13:1008–18. doi: 10.1007/s12559-021-09881-2
51. Straw I, Callison-Burch C. Artificial Intelligence in mental health and the biases of language based models. *PLoS ONE.* (2020) 15:e0240376. doi: 10.1371/journal.pone.0240376
52. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science.* (2017) 356:183–6. doi: 10.1126/science.aal4230
53. Jovic A, Stancin I, Friganovic K, Cifrek M. Clinical decision support systems in practice: current status and challenges, In: *020 43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020 - Proceedings* (Institute of Electrical and Electronics Engineers Inc.) (2020). 355–360. doi: 10.23919/MIPRO48935.2020.9245283
54. Schoonderwoerd TAJ, Jorritsma W, Neerincx MA, van den Bosch K. Human-centered XAI: developing design patterns for explanations of clinical decision support systems. *Int J Hum Comput Stud.* (2021) 154:102684. doi: 10.1016/j.ijhcs.2021.102684
55. Carcel C, Harris K, Peters SAE, Sandset EC, Balicki G, Bushnell CD, et al. Representation of women in stroke clinical trials. *Neurology.* (2021) 97:e1768–74. doi: 10.1212/WNL.0000000000012767
56. Melloni C, Berger JS, Wang TY, Gunes F, Stebbins A, Pieper KS, et al. Representation of women in randomized clinical trials of cardiovascular disease prevention. *Circ Cardiovasc Qual Outcomes.* (2010) 3:135–42. doi: 10.1161/CIRCOUTCOMES.110.868307
57. Bryan CJ, Butner JE, May AM, Rugo KF, Harris JA, Oakey DN, et al. Nonlinear change processes and the emergence of suicidal behavior: a conceptual model based on the fluid vulnerability theory of suicide. *New Ideas Psychol.* (2020) 57:100758. doi: 10.1016/j.newideapsych.2019.100758
58. McHugh CM, Corderoy A, Ryan CJ, Hickie IB, Large MM. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych Open.* (2019) 5:e18. doi: 10.1192/bjo.2018.88

Conflict of Interest: JC and LR are employed full-time by Clarigent Health. JW-B is a part-time consultant and is sponsored by a grant from Clarigent Health. DT, LD, and TK are employed full-time by WakeMed Health & Hospitals. The above interests do not alter our adherence to Frontiers Media's policies. Clarigent Health and WakeMed Health & Hospitals did not influence or restrict the submission of this publication.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.