Investigating the Differential Impact of Psychosocial Factors by Patient Characteristics and Demographics on Veteran Suicide Risk Through Machine Learning Extraction of Cross-Modal Interactions*

Joshua Levy†
Department of Computational Biomedicine, Cedars Sinai Medical Center
Los Angeles, CA USA
Email: joshua.levy@cshs.org

Monica Dimambro
White River Junction VA Medical Center
White River Junction, VT USA
Email: monica.dimambro@va.gov

Alos Diallo, Jiang Gui
Dartmouth College Geisel School of Medicine
Hanover, NH USA
Email: alos.b.diallo.gr@dartmouth.edu, jiang.gui@dartmouth.edu

Brian Shiner, Maxwell Levis
White River Junction VA Medical Center
White River Junction, VT USA
Email: brian.shiner@va.gov, maxwelle.levis@va.gov

Accurate prediction of suicide risk is crucial for identifying patients with elevated risk burden, helping ensure these patients receive targeted care. The US Department of Veteran Affairs' suicide prediction model primarily leverages structured electronic health records (EHR) data. This approach largely overlooks unstructured EHR, a data format that could be utilized to enhance predictive accuracy. This study aims to enhance suicide risk models' predictive accuracy by developing a model that incorporates both structured EHR predictors and semantic NLP-derived variables from unstructured EHR. XGBoost models were fit to predict suicide risk– the interactions identified by the model were extracted using SHAP, validated using logistic regression models, added to a ridge regression model, which was subsequently compared to a ridge regression approach without the use of interactions. By introducing a selection parameter, α, to balance the influence of structured (α=1) and unstructured (α=0) data, we found that intermediate α values achieved optimal performance across various risk strata, improved model performance of the ridge regression approach and uncovered significant cross-modal interactions between psychosocial constructs and patient characteristics. These interactions highlight how psychosocial risk factors are influenced by individual patient contexts, potentially informing improved risk prediction methods and personalized interventions. Our findings underscore the importance of incorporating nuanced narrative data into

* This work is supported by Department of Defense grant PR220927 to JL, ML, JG, BS, NIH P30CA023108 support for JL, and by VA Clinical Science Research and Development Career Development Award (CX002630) to ML.
† To whom correspondence should be addressed.

predictive models and set the stage for future research that will expand the use of advanced machine learning techniques, including deep learning, to further refine suicide risk prediction methods.

Keywords: machine learning, suicide risk, clinical notes, electronic health records, Veterans

1. Introduction

Veterans are at an elevated risk of suicide, underscoring the critical need for advanced risk stratification methods within the US Department of Veterans Affairs. The primary tool currently in use is Recovery Engagement and Coordination for Health – Veterans Enhanced Treatment (REACH-VET), an AI-driven model that utilizes structured data to assess and categorize suicide risk [1,2]. This model plays a crucial role in pinpointing Veterans who are at the highest risk and disproportionately contribute to the annual suicide statistics [3].

Recent research efforts have focused on augmenting the REACH-VET model by integrating unstructured data sources, such as clinical notes, to uncover additional predictors of risk [4,5]. Our previous studies have aimed at identifying specific risk groups and stratifying the impact of various textual suicide risk factors within these groups [6–10]. By using REACH-VET to establish baseline risk, we have developed NLP models that employ semantic databases and textual analysis to track risk factors across different risk tiers and determine optimal intervention timing.

Our prior research effectively identified novel NLP-derived variables that complement traditional demographic and structured risk predictors. By matching cases and controls based on their risk percentiles as measured through structured predictors, we sought to control for potential confounding factors. Controlling for confounding factors, however, does not address the possibility of effect modification, an area that remains relatively underexplored in this context. Understanding how risk factors differ by Veteran subgroups is crucial not only for improving predictive accuracy but also for enhancing the explainability of how psychosocial factors relate to suicide risk across diverse groups. Exploring these interactions could lead to interventions that are more tailored and effective, underscoring the importance of this research for future clinical applications.

Classification and regression trees (CART) are particularly useful for examining effect modifiers among predictors through conditional decision splits. Previous studies have demonstrated their utility in revealing complex statistical interactions. Despite the effectiveness of CART in managing interactions, a significant challenge persists due to the overwhelming number of NLP variables compared to the relatively fewer structured predictors and patient-level clinical factors. This imbalance complicates their effective integration into the predictive model. This disparity necessitates innovative approaches to manage and interpret the extensive data generated by NLP techniques within our predictive models in the context of these patient factors.

This manuscript describes our methodology for refining risk prediction models by integrating both structured and unstructured data within a risk-matched Veteran population, aiming to deliver a more intricate comprehension of suicide risk. Our approach not only seeks to provide more pertinent risk assessments tailored to specific subpopulations but also aims to demonstrate how machine learning models can effectively identify effect modifiers of crucial psychosocial variables based on patient characteristics. These modifiers, once validated through conventional statistical regression

methods, have the potential to significantly improve the interpretability and precision of existing models for assessing suicide risk.

2. Methods

2.1. *Patient Selection*

To establish our study group, we integrated data from the VA Corporate Data Warehouse (CDW) Electronic Health Records (EHR) with mortality information from the VA-Department of Defense Mortality Data Repository [11]. This allowed us to pinpoint Veterans who died by suicide and interacted with VHA healthcare services in 2017 or 2018, totaling 2,842 cases. Following established recommended guidance for matched case-control studies that focus on infrequent events, we matched each suicide case with five controls. Assistance from the VA Office of Mental Health and Suicide Prevention was crucial in selecting control subjects who were treated at the same VHA facility and during the same period as the cases. Controls were chosen to match the deceased cases' REACH-VET risk percentile and were alive at the time the cases died (totaling 14,042 controls) [12]. Controls were unique and non-overlapping, such that no cases could share the same controls. We validated the effectiveness of our matching approach by evaluating the standardized mean differences in various demographic and clinical parameters between cases and controls (**Table 1**). In a prior study that analyzed risk trends in a national sample of recent VA suicide deaths [12], we found that patients at varying suicide risk tiers (high, moderate/med, and low), have very different diagnostic, service usage, and demographic patterns. To best develop targeted risk models, we stratified the present study's sample using these risk tiers.

Table 1: Patient Characteristics/Demographics

	Case (N=2842)	Control (N=14042)	p-value
Demographics			
Female	119 (4.2%)	1079 (7.7%)	0.149
Non married	1688 (59.4%)	7861 (56.1%)	0.068
Married	1154 (40.6%)	6163 (43.9%)	0.038
Homeless_prior24m	212 (7.5%)	1189 (8.5%)	
Veteran	2834 (99.7%)	13971 (99.6%)	0.017
Rural	635 (22.3%)	3215 (22.9%)	0.013
Risk Tier			0.007
High	389 (13.7%)	1940 (13.8%)	
Moderate	1436 (50.5%)	7040 (50.2%)	
Low	1017 (35.8%)	5044 (36.0%)	
Race			0.273
Am. Ind. or Asian Pacific	61 (2.1%)	308 (2.2%)	
Black	154 (5.4%)	1638 (11.7%)	
Hispanic	124 (4.4%)	875 (6.2%)	
Unknown	129 (4.5%)	306 (2.2%)	
White	2374 (83.5%)	10897 (77.7%)	
Age			0.008
Mean (SD)	60.5 (18.0)	60.4 (15.7)	
Deployment			
Vietnam	1100 (38.7%)	5862 (41.8%)	0.066
Afghanistan or Iraq	957 (33.7%)	4761 (33.9%)	0.017
Mental Health Diagnosis/ Risk Flag			
Anxiety	1341 (47.2%)	6686 (47.7%)	0.009
Bipolar	545 (19.2%)	2238 (16.0%)	0.085
Conduct	56 (2.0%)	316 (2.3%)	0.02
Depression	1876 (66.0%)	9137 (65.2%)	0.02
Neurocognitive	316 (11.1%)	1671 (11.9%)	0.025
OCD	80 (2.8%)	325 (2.3%)	0.032
PTSD	1060 (37.3%)	5273 (37.6%)	0.005
Personality	389 (13.7%)	1599 (11.4%)	0.070
Sleeping	1331 (46.8%)	7270 (51.8%)	0.100
Substance	1249 (43.9%)	5401 (38.5%)	0.112
Trauma	1442 (50.7%)	7235 (51.6%)	0.016
Combat	731 (25.7%)	2680 (19.1%)	0.159
Military Sexual Trauma	126 (4.4%)	875 (6.2%)	0.080
Number of Inpatient Mental Health Days within 1 Year of Death			
Mean (SD)	17.2 (66.1)	15.6 (64.6)	0.024
Prescriptions			
Opioid Rx_prior12	885 (31.1%)	4338 (30.9%)	0.004
Opioid Rx_prior24	1104 (38.8%)	5686 (40.5%)	0.035
Mood Stabilizer Rx_prior12	1017 (35.8%)	4718 (33.6%)	0.045
Mood Stabilizer Rx_prior24	1178 (41.4%)	5455 (38.9%)	0.052
Antipsychotic Rx_prior12	616 (21.7%)	2364 (16.9%)	0.122
Antipsychotic Rx_prior24	708 (24.9%)	2791 (19.9%)	0.120
Antidepressant Rx_prior12	1573 (55.3%)	7661 (54.6%)	0.014
Antidepressant Rx_prior24	1733 (61.0%)	8401 (59.9%)	0.022

2.2. *Data Collection and Partitioning*

2.2.1. *Clinical Note Retrieval*

We retrieved unstructured EHR notes from the CDW that were recorded within 30 days before each case's death. This timeframe was chosen based on earlier research that highlighted the significance of clinical notes during the period immediately leading up to death by suicide. To prevent the

influence of potential data leakage / endogeneity, we excluded notes from the final two days before death and any notes that referenced death or a high likelihood of death within the five days prior to the suicide. Additionally, we removed patients from our analysis if their records contained more than six times the average number of notes, thus preventing a disproportionate focus on individuals with higher healthcare engagement. This resulted in a dataset of 92,399 notes from 389 cases and 1,940 controls at high risk, 107,532 notes from 1,436 cases and 7,040 controls at moderate risk, and 44,613 notes from 1,017 cases and 5,044 controls at low risk. Model training and interpretation was conducted on the note-level, whereas performance was reported on the patient level (**see 2.4**).

2.3. *Data Preparation*

2.3.1. *Derivation of NLP Variables*

To capture word counts, we first converted all our text to lowercase, removed stop words like "his/hers", "were/would", "and/with", etc. and tokenized our data set into unigrams or bigrams. We used Sentiment Analysis and Cognition Engine (SÉANCE) to analyze sentiment from these tokens, transforming our corpus into 516 semantic variables. SÉANCE is a Python-based software package that is accessible on VA servers and has been found to be comparable to the commonly used Linguistic Inquiry and Word Count (LIWC) software [13,14]. SÉANCE utilizes a variety of established linguistic databases, including SemanticNet [15,16], General Inquirer Database (GID) [17], EmoLex [18,19], Lasswell [20], Valence Aware Dictionary and sEntiment Reasoner (VADER) [21], Hu–Liu [22,23], Harvard IV-4 [17], and the Geneva Affect Label Coder (GALC) [24]. Each database consists of expert-derived dictionary lists and rule-based systems [25], comprising over 250 unique variables, which can be assessed in positive and negative iterations, leading to 516 SÉANCE variables.

2.3.2. *Extraction of Patient Characteristics*

Using data from the Corporate Data Warehouse (CDW), we extracted a comprehensive array of information encompassing demographics, social determinants of health, patterns of service usage, prescription histories, and diagnostic details. This data set included key demographic variables such as age, gender, marital status, and race. Social determinants like homelessness and military service were also considered, providing context to the healthcare challenges these Veterans may face. The service usage patterns captured included the number and types of visits to emergency departments and mental health services, which are critical indicators of health engagement and potential crisis points. Prescription data detailed the use of critical medications such as opioids and antipsychotics, while diagnostic information covered a wide range of mental health conditions from anxiety and depression to PTSD and substance abuse disorders (**Table 1**). We observed a significant disparity in the number of traditional patient characteristics available (n=66) compared to the number of NLP-derived variables (n=516), which include terms and their negations extracted from clinical notes. A complete list of variables included in the model can be found in the Supplementary Material, available at the following URL: https://github.com/jlevy44/NLP_Demographics_VA/tree/main/Data_Dictionaries .

2.3.3. *Training, Validation and Test Patient Cohorts*

For each risk tier, patients were stratified into training, validation, and test sets using an 80%, 10%, and 10% split, respectively. We utilized the *GroupShuffleSplit* function from the scikit-learn

package (Python v3.8) to ensure that all notes and patient characteristics from the same individual were grouped into the same set [26]. This approach prevents any notes from a single patient from being distributed across different sets, thereby avoiding data leakage and ensuring the integrity of test set statistics. Variables were standardized via scaling parameters estimated from the training set.

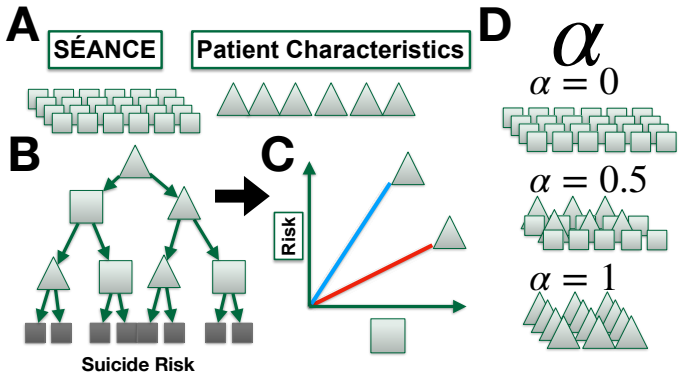

Figure 1: Workflow overview: A) The number of SÉANCE variables dwarfs the number of patient characteristics / demographics, **B)** Cross-modal interactions between SÉANCE and patient characteristics can be identified using a CART approach (e.g., XGBoost) through conditional decision splits between the different sets of variables; **C)** Shows how the relationships between NLP variables (squares) and suicide risk vary across different demographic subgroups (triangles). The lines represent these varying associations, providing simplified interpretations based on the GLM approach. **D)** Selection of SÉANCE and patient characteristics variables controlled through α, intermediate values reflect selection of both variable types, increasing likelihood of detecting cross-modal interactions

2.4. *Selected Machine Learning Models*

All 582 patient characteristics / demographic and SÉANCE variables were modeled simultaneously to predict whether a clinical note corresponded to a patient who had died by suicide. Note-level predicted probabilities ($p = f(\vec{x})$) were averaged across the notes within each patient into a final patient-level score ($\bar{p}$) reflecting the risk of suicide used as the final comparison. We evaluated model performance on both the validation and test sets by calculating the patient-level area under the receiver operating characteristic curve (AUROC) which compared $\bar{p}$ to whether the patient died by suicide. To ensure robustness, we employed a 1000-sample non-parametric bootstrapping to compute 95% confidence intervals for AUROC estimates.

We aimed to evaluate the performance of two machine learning models: 1) Penalized high-dimensional generalized linear models, exemplified by ridge logistic regression [27], which apply an L2-norm penalty to shrink model coefficients (set to 2.5e5 after a coarse hyperparameter search). This method reduces model complexity and prevents overfitting by addressing multicollinearity. 2) Classification and regression trees (CART) [28], as implemented by Extreme Gradient Boosting (XGBoost). XGBoost is an advanced form of gradient boosting that incrementally refines decision trees by concentrating on errors from previous trees [29]. It uses a gradient descent algorithm to meticulously adjust tree parameters, optimizing them based on the error gradient relative to earlier predictions. While the ridge regression model serves as a baseline for performance comparison, the XGBoost model is expected to enhance performance by capturing statistical interactions within and across modalities—specifically among patient characteristics / demographic and SÉANCE variables, as well as interactions between these modalities. XGBoost was specifically chosen for its capability to assign weights to features, directly influencing their selection probabilities during model training. This feature is crucial for effectively balancing the influence of different predictors

(see **2.5**). While LightGBM and BART (Bayesian Additive Regression Trees) offer similar functionalities, they were not selected for specific reasons [30,31]. LightGBM, for instance, only reweights the feature split gain after their initial selection, without altering initial selection probabilities. BART, on the other hand, allows for assignment of priors for variable selection, but computational demands are significantly higher, making it less suitable for our current scope but a potential candidate for future exploration. Ridge regression was selected as representative of generalized linear modeling approaches after initial comparisons to LASSO and ElasticNet [32].

The primary objective of this study is not simply to compare Ridge regression with XGBoost. Rather, our aim is to show that the interactions identified by XGBoost can offer additional valuable information, enhancing the predictive accuracy of Ridge regression and other generalized linear models that are known for their parsimonious interpretations. Consequently, we expect that incorporating these interactions as predictors—a method we have named *Ridge-Int*—will significantly improve the performance of Ridge regression, bridging the gap between complex machine learning and traditional statistical models [33].

2.5. *Key Contribution: Weighting the selection of NLP variables and patient characteristics*

In tree-based models, the selection of variables for inclusion at various levels or nodes typically occurs with uniform probability. This approach can inadvertently lower the probability of selecting variables from smaller sets of variables, such as patient characteristics, compared to larger sets like those from NLP-derived variables. Consequently, this bias in variable selection could hinder the identification of meaningful interactions between patient characteristics / demographic and NLP variables, as the former are less likely to be chosen as nodes or leaves in the model.

To address this imbalance, we hypothesize that strategically weighting the selection of variables from these two distinct sets—patient characteristics / demographic and NLP variables—could be crucial for uncovering optimal interactions between them. In this study, we conduct a sensitivity analysis to explore the impact of different weighting strategies on the detection of interactions (**Figure 1**). Specifically, we investigate three scenarios:

1. **Upweighting Patient Characteristics / Demographics:** We hypothesize that increasing the selection probability of patient characteristics (including demographics) could enhance the identification of interactions within these features.
2. **Upweighting NLP Features:** Conversely, increasing the weight of NLP features is expected to surface more interactions within the NLP data.
3. **Balanced Weighting:** Applying an equal weighting strategy, adjusted for the numerical disparity between the sets (upweighting patient characteristics / demographics proportionally to the number of NLP features), is hypothesized to facilitate the detection of cross-modal interactions, balancing the trade-offs between the two.

To test these hypotheses, we introduce a selection hyperparameter, $\alpha \in [0,1]$, which determines the extent to which one set of predictors is favored over the other. The weighting formula for individual patient characteristics / demograpgics is defined as $\alpha * \frac{n_{SEANCE}}{n_{demographics}} + \epsilon$, where $\frac{n_{SEANCE}}{n_{demographics}}$ represents the ratio of the number of NLP variables to patient characteristics, adjusting for their discrepancy. Conversely, the weight for selecting NLP variables is set as $1-\alpha+ \epsilon$. Thus, an α value of 0 would give priority to NLP variables, highlighting interactions within the NLP data, whereas an α of 1 would prioritize patient characteristics / demographics, enhancing the identification of interactions solely between structured patient characteristics. Here, ϵ is a small constant (ϵ = 1e-7) introduced to

ensure that the probability of selecting variables from either predictor set never reaches zero as required by the XGBoost package. This minimal adjustment allows for the rare but possible selection of variables from the non-prioritized set.

Table 2: XGBoost hyperparameter search grid

Hyperparameter	Values
colsample_bynode	0.25, 0.5, 0.75, 1
colsample_bylevel	0.25, 0.5, 0.75, 1
colsample_bytree	0.5, 0.75, 1
subsample	0.6, 0.8, 1
min_child_weight	1, 3, 5, 7
max_depth	3, 4, 5, 6
gamma	0, 1, 5, 10
reg_alpha	0, 0.1, 1, 10
reg_lambda	0, 0.1, 1, 10
Number of Trees	25, 50, 100

2.6. *Model Fitting, Interaction Extraction and Validation for Experimental Comparisons*

We trained XGBoost models using various values for α, including 0 (favoring SÉANCE variables), 0.1, 0.3, 0.5, 0.7, 0.9, and 1 (favoring patient characteristics). The model fitting process involved a 50-iteration randomized search for optimal hyperparameters (**Table 2**), with early stopping for tree boosting based on validation set performance. This procedure was repeated for all α and risk tiers.

For each value of α, we used the *interactiontransformer* package [33] to select candidate interactions for further analysis via the tree explainer. This assigns each interaction a global SHAP score, which represents the average influence of the interaction across all notes and patients, reflecting its overall contribution to the model's performance [34]. SHAP interaction scores were computed separately for the validation and test sets. To validate candidate interactions, we examined the top 1000 interactions identified by SHAP. For each interaction, we fit unpenalized generalized linear models (GLM, logistic regression) incorporating the interaction term (**Figure 1C**):

$$logit(p_{suicide}) = \beta_0 + \beta_1 feature_1 + \beta_2 feature_2 + \beta_3 feature_1 * feature_2$$

A candidate interaction was confirmed as validated if the p-value for the coefficient of the interaction term, β_3, was less than 0.05 divided by 1000. This stringent criterion reflects the Bonferroni adjustment applied to account for multiple comparisons (1000 candidate interactions), ensuring the robustness of our findings against Type I errors.

To evaluate the effectiveness of SHAP values in identifying and prioritizing key interactions, we used Fisher's exact test to compare the likelihood of GLM-validated variables appearing in the top 100 versus the top 1000 SHAP-ranked interactions. By calculating an odds ratio (OR) and a corresponding p-value as a measure of enrichment in the top 100 set, we quantified the degree to which SHAP values not only identify but also accurately prioritize the most impactful interactions.

We hypothesized that validated interactions would be predominantly found among the highest-ranked interactions by SHAP, indicating the effectiveness of SHAP in identifying the most influential interactions in terms of their contribution to the model's predictive accuracy. This step serves not only to validate the interactions but also to verify the reliability of SHAP's ranking mechanism in prioritizing the most statistically significant and predictive interactions [a].

For the validated interactions, we categorized the nature of each interaction based on its modality: either within modality interactions (such as demographic-demographic or SÉANCE-SÉANCE) or cross-modality (demographic-SÉANCE) interactions. We quantified these categories by calculating their proportions within the overall set of validated interactions, providing insight into the patterns of relationships that significantly contribute to the model.

After categorizing the interactions, we incorporated them into the predictive model. Specifically, we enhanced the Ridge regression model by adding either the validated interactions or the top 50

[a] Further clarification on these calculations can be found in Supplementary materials: "Clarification on Role of Algorithms and Methods", at: https://github.com/jlevy44/NLP_Demographics_VA/blob/main/suppl_material.docx

interactions ranked by p-value—whichever count was greater. This enhanced model, referred to as *Ridge-Int*, was designed to assess the impact of including significant interaction terms on the predictive accuracy. The performance of the *Ridge-Int* model was compared against the baseline Ridge model, which did not include interaction terms. We evaluated the models' effectiveness on both the validation and test sets using the AUROC, with 95% confidence intervals calculated using the previously described bootstrapping method.

Following the validation and integration of interactions into our models, we plotted the AUROCs for XGBoost, Ridge, and *Ridge-Int* against the hyperparameter α, along with the odds ratios for the significance of validated interactions and the proportion of validated interactions that were cross-modal within the validation and test set. We anticipated that the interactions identified by XGBoost would enhance the performance of Ridge regression, and that the performance metrics for both XGBoost and *Ridge-Int* would likely reach a plateau at an intermediate α value between 0 and 1. Similarly, we expected the enrichment of validated interactions and the proportion of cross-modal interactions to saturate at a midpoint α, demonstrating the relevance of cross-modal interactions for enhancing predictiveness. This analysis was stratified and performed across each risk tier, allowing for a nuanced evaluation of how the inclusion of interactions influences model performance within distinct suicide risk tiers.

2.7. *Interpretation of Randomly Selected Interactions*

To deepen our understanding of the interactions between different modalities, we analyzed the statistical interaction models fitted to the data, employing estimated marginal means as a post hoc comparison to elucidate the effects of various psychosocial constructs obtained through NLP on suicide risk [35,36]. These effects were specifically examined as conditioned by patient characteristics, and similarly, how patient characteristics / demographics influence the impact of psychosocial factors on suicide risk (**Figure 1C**). For illustrative clarity, effect estimates for randomly selected interactions were presented in detailed tables and supportive visualizations demonstrating the conditional/stratified effects of these psychosocial constructs by patient characteristics.

3. Results

3.1. *Affirming the Relevance of Cross-modal Interactions*

In our study, we adjusted the selection probability between two predictor sets to investigate their potential trade-offs in influencing model performance. By altering the hyperparameter α, we modulated the selection bias towards either patient characteristics / demographics or SÉANCE variables within the XGBoost model, and then examined the nature and impact of the interactions identified by SHAP. The interactions that were extracted and validated using unpenalized GLMs– prior to their inclusion in *Ridge-Int*– with interaction terms demonstrated statistical validity. Importantly, these interactions, once confirmed through statistical modeling, showed a high enrichment within the top 100 SHAP-ranked interactions, evidenced by significant odds ratios. This result supports the effectiveness of XGBoost and SHAP in pinpointing genuine interactions.

Our analysis revealed that interactions were more prevalently validated at intermediate values of α, suggesting an optimal balance at these levels for extracting meaningful interactions between different types of data (**Figure 2B, Table 3**). The proportion of cross-modal interactions that were validated peaked at these intermediate α values (**Figure 2C, Table 3**). This finding corroborates the

hypothesis that adjusting α allows for fine-tuning of the XGBoost model to effectively balance the contribution of both predictor sets, enhancing the model's capacity to uncover and utilize significant interactions between different data types. It should be noted that the enrichment of validated interactions was especially pertinent for low-risk tier patients, though a lower proportion of these interactions were cross-modal in nature.

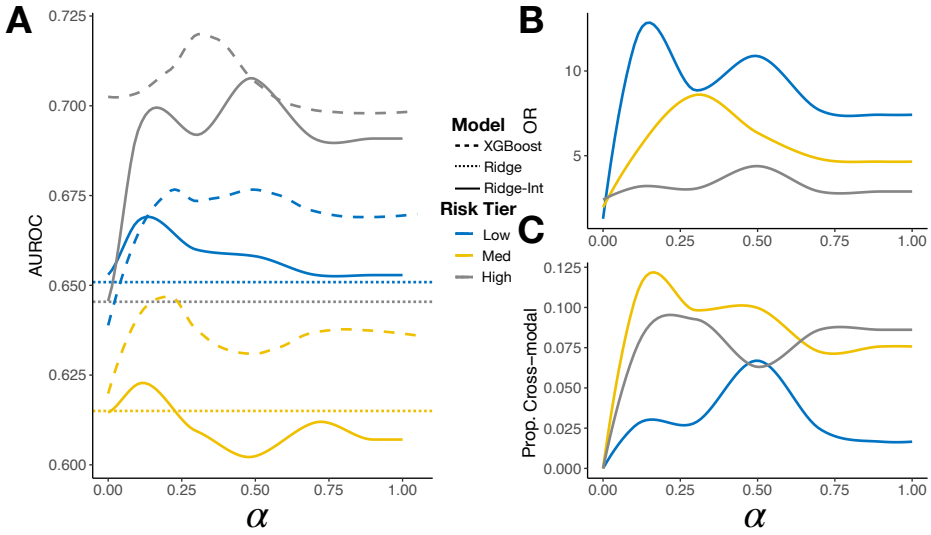

Figure 2: Model Comparison and Interaction Validation/Delineation: A) Test set model performance reported via the AUROC on the patient-level, aggregated across notes, for each model type and risk tier. **B)** Odds Ratio (OR) versus α. OR reflects the enrichment of validated interactions among the top-ranked interactions identified through SHAP, serving as a measure of how well the statistical model validates the interactions identified by SHAP. **C)** The proportion of validated interactions that were identified as cross-modal as a function of α.

3.2. *Model Performance Comparisons*

In our study, we hypothesized that cross-modal interactions would significantly enhance the predictive performance in suicide risk assessment, particularly when analyzing aggregated data across patient notes. Our results confirmed this hypothesis, demonstrating the critical role of these interactions in improving model accuracy. Notably, the XGBoost model, which explicitly accounts for statistical interactions through conditional decision splits, consistently outperformed the traditional Ridge regression model, which does not inherently consider interactions (**Figure 2A, Table 4).** Upon integrating these extracted interactions from the XGBoost model into the Ridge regression frameworks (*Ridge-Int*), we observed marked performance improvement for the low and high risk patients and modest improvement in the moderate risk patients for α=0.1 (**Figure 2A, Table 4)**.

Interestingly, the most pronounced gains were observed when the selection parameter α, which balances the influence of structured patient characteristics / demographics versus SÉANCE features, was set to intermediate values. This suggests that neither purely patient characteristics nor purely SÉANCE features are sufficient on their own; rather, it is their combination and the interactions between them that drive the predictive accuracy of the models. This phenomenon was corroborated by the relative alignment of these optimal α values with where the highest number of crossmodal interactions were identified and incorporated into the statistical modeling of the Ridge regression (**Figure 2A,C, Tables 2,3**).

Table 3: Validation and Analysis of XGBoost-Derived Interactions via SHAP and Subsequent Logistic Regression Modeling. OR indicate the degree to which validated interactions as confirmed through logistic regression modeling are enriched among the top SHAP-ranked interactions, reflecting the effectiveness of the XGBoost in identifying relevant interactions. Additionally, the table lists the proportion of these validated interactions that were characterized as cross-modal, highlighting their potential for bridging distinct data modalities.

Risk	α	OR	p	% Cross-Modal	Risk	α	OR	p	% Cross-Modal	Risk	α	OR	p	% Cross-Modal
Low	0	1.3	0.672	0.0%	**Med**	0	2.0	0.082	0.0%	**High**	0	2.4	0.017	0.0%
	0.1	**11.5**	**<0.001**	2.6%		**0.1**	5.0	<0.001	**10.3%**		0.1	3.1	<0.001	7.1%
	0.3	8.9	<0.001	2.9%		**0.3**	**8.6**	**<0.001**	9.8%		**0.3**	3.1	<0.001	**9.3%**
	0.5	10.9	<0.001	**6.7%**		**0.5**	6.3	<0.001	**10.0%**		**0.5**	**4.4**	**<0.001**	6.3%
	0.7	7.7	<0.001	2.5%		0.7	4.8	<0.001	7.3%		0.7	2.9	<0.001	8.6%
	0.9	7.4	<0.001	1.7%		0.9	4.6	<0.001	7.6%		0.9	2.9	<0.001	8.6%
	1	7.4	<0.001	1.7%		1	4.6	<0.001	7.6%		1	2.9	<0.001	8.6%

Table 4: Test Set Model Performance for XGBoost and Ridge Regression models, comparing performance across low, medium, and high-risk tiers. The AUROC values are presented alongside 95% CIs calculated through 1000-sample non-parametric bootstrapping. For Ridge Regression, 'n/a' indicates the performance of the model without the inclusion of interactions derived from XGBoost, serving as a baseline comparison. The variations in AUROC values across different α (ranging from 0 to 1) illustrate the impact of emphasizing either patient characteristics / demographics or SÉANCE features, or a balanced consideration of both, in predicting suicide risk.

XGBoost

Risk	α	AUROC	2.5%CI	97.5%CI	Risk	α	AUROC	2.5%CI	97.5%CI	Risk	α	AUROC	2.5%CI	97.5%CI
Low	0	0.622	0.565	0.681	**Med**	0	0.602	0.555	0.646	**High**	0	0.715	0.642	0.788
	0.1	**0.678**	**0.624**	**0.735**		**0.1**	**0.658**	**0.61**	**0.705**		0.1	0.687	0.61	0.767
	0.3	0.672	0.618	0.721		0.3	0.637	0.59	0.681		**0.3**	**0.726**	**0.65**	**0.799**
	0.5	**0.679**	**0.622**	**0.734**		0.5	0.629	0.579	0.681		0.5	0.705	0.627	0.781
	0.7	0.669	0.611	0.723		0.7	0.638	0.589	0.686		0.7	0.698	0.618	0.782
	0.9	0.669	0.611	0.723		0.9	0.637	0.587	0.683		0.9	0.698	0.618	0.782
	1	0.669	0.611	0.723		1	0.637	0.587	0.683		1	0.698	0.618	0.782

Ridge Regression

Risk	α	AUROC	2.5%CI	97.5%CI	Risk	α	AUROC	2.5%CI	97.5%CI	Risk	α	AUROC	2.5%CI	97.5%CI
Low	n/a	0.651	0.593	0.704	**Med**	n/a	0.615	0.566	0.666	**High**	n/a	0.645	0.561	0.729
	0	0.653	0.596	0.706		0	0.615	0.565	0.665		0	0.646	0.562	0.729
	0.1	**0.668**	**0.61**	**0.72**		**0.1**	**0.622**	**0.574**	**0.672**		0.1	0.693	0.602	0.778
	0.3	0.66	0.6	0.714		0.3	0.609	0.559	0.659		0.3	0.692	0.603	0.773
	0.5	0.658	0.599	0.715		0.5	0.602	0.552	0.652		**0.5**	**0.707**	**0.621**	**0.784**
	0.7	0.653	0.594	0.708		0.7	0.612	0.562	0.66		0.7	0.691	0.602	0.775
	0.9	0.653	0.594	0.708		0.9	0.607	0.557	0.655		0.9	0.691	0.602	0.775
	1	0.653	0.594	0.708		1	0.607	0.557	0.655		1	0.691	0.602	0.775

Table 5: Randomly Selected Validated Cross-Modal Interactions Across All Risk Tiers. Showcases a sample of cross-modal interactions validated from logistic regression analyses on note-level across different risk tiers, with findings adjusted for 1000 multiple comparisons for each risk tier using Bonferroni correction. Each column corresponds to different α levels (0.3, 0.5, and 0.7 as representative), demonstrating the variability in interaction significance and strength (log(OR)) with changing α values. See supplementary materials for term descriptions.

	α=0.3			α=0.5			α=0.7		
Risk	**Interaction**	**log(OR)**		**Interaction**	**log(OR)**		**Interaction**	**log(OR)**	**p-adj**
High	days_inpatMH_prior_12mo:Notlw_Lasswell	-0.02	7.4e-07	Sleeping:vader_neutral	-0.98	2.6e-02	Calc_age:Know_GI	0.16	1.0e-04
	days_inpatMH_prior_12mo:Coll_GI_neg_3	-0.03	4.0e-14	antipsy_prior24:vader_compound	0.19	1.0e-05	age_55_74:Hu_GI_neg_3	3.59	3.4e-10
	Substance:negative_adjectives_component	0.12	4.3e-02	age_55_74:Posaff_Lasswell	-4.22	1.6e-03	days_inpatMH_prior_12mo:Tool_GI_neg_3	0.02	7.4e-05
	Calc_age:negative_adjectives_component	0.01	7.0e-08	Substance:Work_GI	-4	1.5e-03	age_55_74:Socrel_GI_neg_3	2.74	4.8e-04
	age_55_74:Coll_GI_neg_3	10.47	2.0e-22	Calc_age:Endslw_Lasswell	0.1	7.9e-03	elix_cat:Know_GI	3.24	1.2e-06
Med	elix_cat:fear_and_digust_component	0.74	3.2e-03	elix_cat:Male_GI	-4.36	9.5e-04	elix_cat:Male_GI	-4.36	9.5e-04
	elix_cat:Fear_EmoLex	3.26	6.1e-03	elix_cat:fear_and_digust_component	0.74	3.2e-03	moodst_prior24:Abs_GI	4.22	8.4e-03
	Substance:Powcoop_Lasswell_neg_3	-12.25	2.4e-05	opioid_prior12:Submit_GI_neg_3	-4.52	7.0e-05	di_cat:hu_liu_pos_perc_neg_3	-0.19	1.9e-02
	Nonmarried:hu_liu_pos_perc_neg_3	0.41	7.4e-03	MH_cat:hu_liu_pos_nwords	3.1	2.5e-02	Trauma:vader_positive	2.24	4.6e-04
	Unknown:Pleasur_GI_neg_3	8.63	3.4e-02	MH_cat:Tranlw_Lasswell	2.85	1.8e-03	Unknown:Pleasur_GI	9.05	1.8e-02
Low	elix_cat:fear_and_digust_component	0.74	3.2e-03	elix_cat:Male_GI	-4.36	9.5e-04	elix_cat:Male_GI	-4.36	9.5e-04
	elix_cat:Fear_EmoLex	3.26	6.1e-03	elix_cat:fear_and_digust_component	0.74	3.2e-03	moodst_prior24:Abs_GI	4.22	8.4e-03
	Substance:Powcoop_Lasswell_neg_3	-12.25	2.4e-05	opioid_prior12:Submit_GI_neg_3	-4.52	7.0e-05	di_cat:hu_liu_pos_perc_neg_3	-0.19	1.9e-02
	Nonmarried:hu_liu_pos_perc_neg_3	0.41	7.4e-03	MH_cat:hu_liu_pos_nwords	3.1	2.5e-02	Trauma:vader_positive	2.24	4.6e-04
	Unknown:Pleasur_GI_neg_3	8.63	3.4e-02	MH_cat:Tranlw_Lasswell	2.85	1.8e-03	Unknown:Pleasur_GI	9.05	1.8e-02

3.3. *Select Interpretation of Findings from Cross-modal Workflow*

The XGBoost model successfully identified numerous cross-modal interactions, specifically at intermediate α, of which we selectively analyzed a few at random to elucidate their implications for decision-making and potential therapeutic advancements. For example, in **Table 5**, the results from GLM of validated interaction terms are presented. **Table 6** and **Figure 3** provide detailed breakdowns of four key interactions with further interpretation presented in the Discussion.

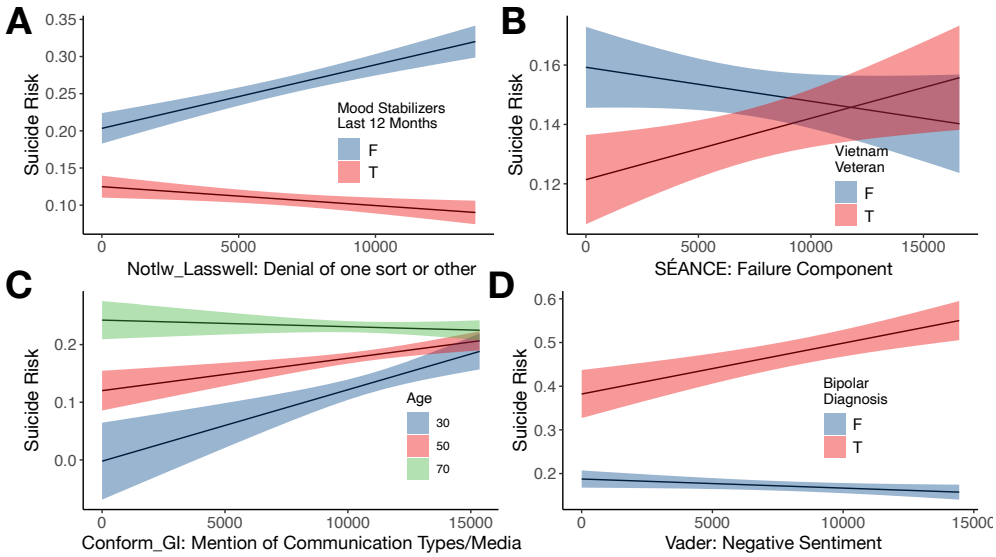

Figure 3: Interpretation of Conditional Effects of Psychosocial Constructs Across Patient Subgroups for Randomly Selected Validated Interactions. A) Suicide risk associated with mood stabilizer use fluctuates based on mention of denial in clinical notes. **B)** Intensified impact of failure mentions in notes on suicide risk among Vietnam Veterans compared to other Veterans. **C)** Varying effects of mentioning communication forms, such as mentions of social media, on suicide risk across age groups, with younger Veterans showing heightened sensitivity. **D)** Increased suicide risk due to negative sentiments among patients with bipolar disorder. Note that interpretations are on the note-level. Risk scale is expressed using the inverse logit link function.

Table 6: Select Interactions and Conditional Effects from Logistic Regression Analysis on Randomly Selected Validated Interactions. Interaction terms are denoted using ":", followed by a conditional effect, denoted by "|", representing the evaluation of the variable's effect under specific conditions set by the modifying variable on the right. Conditional effects are derived using estimated marginal means.

Risk	Term	log(OR)	p-value
High	moodst_prior12:Notlw_Lasswell	-1.11e-05	8.95e-12
	Notlw_Lasswell|moodst_prior12	8.55e-06	7.40e-11
	Notlw_Lasswell|No moodst_prior12	-2.54e-06	7.96e-03
	Vietnam:failure_component	3.21e-06	5.25e-03
	failure_component|Vietnam	-1.15e-06	1.45e-01
	failure_component|Not Vietnam	2.07e-06	1.40e-02
Med	Calc_age:Comform_GI	-3.38e-07	1.87e-04
	Comform_GI|Age=30	1.24e-05	4.05e-05
	Comform_GI|Age=50	5.63e-06	3.25e-04
	Notlw_Lasswell|Age=70	-1.14e-06	4.56e-01
Low	Bipolar:vader_negative	1.37e-05	6.51e-06
	vader_negative|Bipolar	-2.08e-06	4.66e-02
	vader_negative|Not Bipolar	1.16e-05	4.61e-05

4. Discussion

Recent advancements have emphasized the critical role of machine learning and the analysis of unstructured clinical reports in augmenting suicide risk prediction models [37]. These developments

aim to complement existing models that leverage structured predictors already operational within the VA system, which have been further repurposed to categorize risks into defined tiers for population studies [12]. Despite these innovations, the dynamics between predictors derived from structured and unstructured data, and their combined potential to improve suicide risk prediction, remain largely unexplored.

In this study, we aimed to refine suicide risk predictive models to cater specifically to relevant subgroups. Our strategy involved developing models that balanced the inclusion of both structured and unstructured (NLP) predictors. This approach allowed us to delve into the trade-offs and synergies between these predictor types through traditional statistical modeling of the interactions extracted from them. We introduced a predictor set selection parameter, α, to regulate the extent to which predictors from semantic NLP variables (SÉANCE) and structured EHR were utilized.

Our findings revealed that this methodology not only enhanced the accuracy of suicide risk predictions but also illuminated how cross-modal interactions between NLP variables and structured predictors could demonstrate the altered risk associated with various psychosocial constructs based on patient characteristics / demographics and vice versa. The ability to discern these interactions underscores the pivotal role of cross-modal dynamics in improving model performance, validating their importance in complex predictive tasks such as suicide risk assessment.

The implications of our analytical approach are significant– we will now discuss key lessons and insights derived from interpreting the interaction terms (**Figure 3, Tables 5, 6**). A positive interaction effect estimate signifies an elevated suicide risk when one variable increases, conditional on the rise of another variable. Conversely, a negative interaction effect indicates reduced risk under the same conditions. For example, our analysis showed that patients with a substance use disorder who frequently use negative adjectives in their clinical notes are at an increased risk of suicide [38]. In contrast, the presence of negative words has a less pronounced effect on patients without such a disorder. Another notable observation (**Figure 3, Table 6**) is that negative sentiments significantly elevate suicide risk among bipolar patients compared to those who are not bipolar, consistent with prior literature [39]. These instances demonstrate how psychosocial constructs variably affect different patient groups, paving the way for future large-scale studies aimed at identifying novel intervention targets and enhancing preventive strategies in suicide risk management.

This study has several limitations that merit consideration and can inform future work. Firstly, while the predictive modeling results were aggregated across patient notes, the initial predictive modeling and interpretation were conducted at the individual note level. Surprisingly, models trained solely with patient characteristics / demographics (α=1) showed an AUROC greater than expected, given that they were matched based on REACH-VET percentile scores derived from these same characteristics. This outcome suggests two key insights: additional stratification of suicide risk within defined risk tiers can unearth predictive factors not captured by models trained exclusively on structured predictors across the entire population (i.e., effect modification by risk tier); furthermore, the design of XGBoost, which ensures non-zero selection probabilities for variables, allowed the inclusion of a small yet significant set of SÉANCE variables to bolster model predictiveness. We did not compare the usage of TreeSHAP to other interaction extraction approaches [40–42]. Another limitation is that while the structured variables span over a year or more, the NLP variables are derived from observations within the past 30 days. Despite these observed limitations, the fundamental principles and broader findings of our research remain sound and valid. Another limitation is the statistical power to detect interactions, which may have been constrained by the limited sample size of this study. Future work aims to extend this analysis across a broader

temporal and demographic scope at the national level, which should incorporate a more diverse array of characteristics and potentially yield more robust findings above and beyond current suicide risk prediction approaches. Further external validation is limited by the focus on US veterans, and results may not generalize to other populations [43–46]. It is common to train models on historical data and validate them on more recent data, which could have strengthened the validity of our findings.

It should also be noted that previous research has highlighted that increased care utilization significantly influences the REACH-VET scores used for matching and stratification by risk tier [12]. Generally, more comprehensive data on Veterans can contribute to higher inferred suicide risk, whereas patients with less comprehensive records are typically assigned a lower risk. This variability in data completeness across different subpopulations underscores the need for our models to identify associations within these groups, especially since they may be differentially impacted by the extent of their record completeness [47–50].

Looking ahead, we plan to develop machine learning models that are not solely dependent on structured predictors for matching (i.e., randomly matched). This approach will allow us to potentially identify patterns that were previously obscured due to the biases introduced by data completeness. This could lead to more nuanced and effective predictive models that better address the diverse needs of all subgroups within the Veteran population.

The interpretation of findings from SÉANCE terms should be approached with caution [51,52]. SÉANCE encompasses a diverse array of lexical variables, each with different standards and encompassing varying numbers of words. However, these terms face challenges in capturing the nuanced contexts in which these words are used, which can complicate the interpretation of these concepts beyond their mere mention. This limitation is akin to the current challenges faced in sentiment analysis, even with the incorporation of negation terms. Initially, we adopted a semantic database approach as a proof of concept for this method. While we plan to expand our analysis to include a “bag-of-words” approach that captures all words within the corpus, this method also has its limitations as it tends to disregard their context within sentences. Therefore, our future work will focus on employing deep learning techniques to mine for motifs and patterns that can capture more complex and nuanced narratives. This approach will allow us to better contextualize these constructs and understand their differential impacts, informing future interventions more effectively [53–58].

Furthermore, deep learning models offer the flexibility to weigh different forms of information—including social determinants of health—on a patient-by-patient basis. They can also help identify critical timepoints for collecting notes that are most relevant to assessing suicide risk and determining optimal times for intervention. Our earlier work relied on count-based approaches, partly due to the limitations of computing resources available within the VA VINCI computing system. However, as advanced graphics processing units (GPU) systems become more accessible at the VA, we anticipate a shift towards more sophisticated deep learning approaches.

5. Conclusion

In conclusion, this study demonstrates the potential of integrating structured and unstructured data sources to enhance the predictiveness of suicide risk models for Veterans. The nuanced insights gained from cross-modal interactions identified through this comprehensive approach can better appreciate the dynamic interplay between numerical data from electronic health records and rich, psychosocial constructures available in clinical notes. As we move forward, the incorporation of more advanced machine learning techniques, particularly deep learning, promises to further refine our predictive capabilities and offer more targeted, effective interventions and risk prioritization.

References

1. McCarthy JF, Cooper SA, Dent KR, Eagan AE, Matarazzo BB, Hannemann CM, Reger MA, Landes SJ, Trafton JA, Schoenbaum M. Evaluation of the recovery engagement and coordination for health–veterans enhanced treatment suicide risk modeling clinical program in the veterans health administration. JAMA network open. American Medical Association; 2021;4(10):e2129900–e2129900.
2. Kessler RC, Hwang I, Hoffmire CA, McCarthy JF, Petukhova MV, Rosellini AJ, Sampson NA, Schneider AL, Bradley PA, Katz IR, Thompson C, Bossarte RM. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. Int J Methods Psych Res. 2017 Sep;26(3):e1575.
3. McCarthy JF, Bossarte RM, Katz IR, Thompson C, Kemp J, Hannemann CM, Nielson C, Schoenbaum M. Predictive Modeling and Concentration of the Risk of Suicide: Implications for Preventive Interventions in the US Department of Veterans Affairs. Am J Public Health. American Public Health Association; 2015 Sep;105(9):1935–1942.
4. Leonard Westgate C, Shiner B, Thompson P, Watts BV. Evaluation of Veterans' Suicide Risk With the Use of Linguistic Detection Methods. PS. 2015 Oct 1;66(10):1051–1056.
5. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, Watts B, Flashman L, McAllister T. Predicting the risk of suicide by analyzing the text of clinical notes. PloS one. Public Library of Science San Francisco, USA; 2014;9(1):e85733.
6. Levis M, Westgate CL, Gui J, Watts BV, Shiner B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychological medicine. Cambridge University Press; 2021;51(8):1382–1391.
7. Levis M, Levy J, Dufort V, Russ CJ, Shiner B. Dynamic suicide topic modelling: Deriving population-specific, psychosocial and time-sensitive suicide risk variables from Electronic Health Record psychotherapy notes. Clin Psychology and Psychoth. 2023 Jul;30(4):795–810.
8. Levis M, Levy J, Dufort V, Gobbel GT, Watts BV, Shiner B. Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. Psychiatry research. Elsevier; 2022;315:114703.
9. Levis M, Levy J, Dimambro M, Dufort V, Ludmer DJ, Goldberg M, Shiner B. Using natural language processing to evaluate temporal patterns in suicide risk variation among high-risk Veterans. Psychiatry Research. Elsevier; 2024;116097.
10. Levis M, Levy J, Dent KR, Dufort V, Gobbel GT, Watts BV, Shiner B. Leveraging natural language processing to improve electronic health record suicide risk prediction for Veterans Health Administration users. The Journal of clinical psychiatry. Physicians Postgraduate Press, Inc.; 2023;84(4):47557.
11. Department of Veterans Affairs, Department of Defense. Joint Department of Veterans Affairs (VA) and Department of Defense (DoD) Mortality Data Repository – National Death Index (NDI) [Internet]. 2017. Available from: https://www.mirecc.va.gov/suicideprevention/documents/VA_DoD-MDR_Flyer.pdf
12. Levis M, Dimambro M, Levy J, Dufort V, Fraade A, Winer M, Shiner B. Characterizing Veteran suicide decedents that were not classified as high-suicide-risk. Psychological Medicine. Cambridge University Press; 2024;1:1–10.
13. Crossley SA, Kyle K, McNamara DS. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. Behav Res. 2017 Jun;49(3):803–821.

14. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annu Rev Psychol. 2003 Feb;54(1):547–577.
15. Angioni M, Demonits R, Deriu M, Tuveri F. Semanticnet: a WordNetbased Tool for the Navigation of Semantic Information. Proceedings, GWC. 2008;21–34.
16. Das A, Bandyopadhyay S. Semanticnet-perception of human pragmatics. Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon [Internet]. 2010 [cited 2024 Jul 31]. p. 2–11. Available from: https://aclanthology.org/W10-3402.pdf
17. Stone PJ, Dunphy DC, Smith MS. The general inquirer: A computer approach to content analysis. MIT press; 1966 [cited 2024 Jul 31]; Available from: https://psycnet.apa.org/record/1967-04539-000
18. Mohammad SM, Turney PD. CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. Computational Intelligence. 2013 Aug;29(3):436–465.
19. Mohammad S, Turney P. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text [Internet]. 2010 [cited 2024 Jul 31]. p. 26–34. Available from: https://aclanthology.org/W10-0204.pdf
20. Lasswell HD, Namenwirth JZ. The Lasswell value dictionary. New Haven. 1969;
21. Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the international AAAI conference on web and social media [Internet]. 2014 [cited 2024 Jul 31]. p. 216–225. Available from: https://ojs.aaai.org/index.php/icwsm/article/view/14550
22. Hu M, Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. Seattle WA USA: ACM; 2004 [cited 2024 Jul 31]. p. 168–177. Available from: https://dl.acm.org/doi/10.1145/1014052.1014073
23. Hu M, Liu B. Mining opinion features in customer reviews. AAAI [Internet]. 2004 [cited 2024 Jul 31]. p. 755–760. Available from: https://cdn.aaai.org/AAAI/2004/AAAI04-119.pdf
24. Scherer KR. What are emotions? And how can they be measured? Social Science Information. SAGE Publications Ltd; 2005 Dec 1;44(4):695–729.
25. Urbanowicz RJ, Moore JH. Learning Classifier Systems: A Complete Introduction, Review, and Roadmap. Journal of Artificial Evolution and Applications. 2009 Sep 22;2009:1–25.
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825–2830. PMID: 34682092
27. Hoerl AE, Kennard RW. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics. 1970 Feb;12(1):69–82.
28. Loh WY. Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011 Jan 1;1:14–23.
29. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016 [cited 2019 Nov 26]. p. 785–794. Available from: http://doi.acm.org/10.1145/2939672.2939785

30. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems [Internet]. 2017 [cited 2024 Jul 31];30. Available from: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
31. Tan YV, Roy J. Bayesian additive regression trees and the General BART model. Statistics in Medicine. 2019;38(25):5048–5069. PMID: 31460678
32. Ranstam J, Cook JA. LASSO regression. British Journal of Surgery. 2018 Sep 1;105(10):1348.
33. Levy JJ, O'Malley AJ. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. BMC Med Res Methodol. 2020 Jun 29;20(1):171. PMCID: PMC7325087
34. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020 Jan;2(1):56–67. PMID: 32607472
35. Lenth RV, Buerkner P, Giné-Vázquez I, Herve M, Jung M, Love J, Miguez F, Riebl H, Singmann H. emmeans: Estimated Marginal Means, aka Least-Squares Means [Internet]. 2023 [cited 2023 Mar 1]. Available from: https://CRAN.R-project.org/package=emmeans
36. Searle SR, Speed FM, Milliken GA. Population marginal means in the linear model: an alternative to least squares means. The American Statistician. Taylor & Francis; 1980;34(4):216–221.
37. Riblet NB, Matsunaga S, Lee Y, Young-Xu Y, Shiner B, Schnurr PP, Levis M, Watts BV. Tools to detect risk of death by suicide: A systematic review and meta-analysis. The Journal of clinical psychiatry. Physicians Postgraduate Press, Inc.; 2022;84(1):43891.
38. Harlow LL, Newcomb MD, Bentler PM. Depression, self-derogation, substance use, and suicide ideation: Lack of purpose in life as a mediational factor. Journal of Clinical Psychology. 1986;42(1):5–21.
39. Stange JP, Hamilton JL, Burke TA, Kleiman EM, O'Garro-Moore JK, Seligman ND, Abramson LY, Alloy LB. Negative cognitive styles synergistically predict suicidal ideation in bipolar spectrum disorders: A 3-year prospective study. Psychiatry Research. 2015 Mar 30;226(1):162–168.
40. Agrawal R, Trippe B, Huggins J, Broderick T. The Kernel Interaction Trick: Fast Bayesian Discovery of Pairwise Interactions in High Dimensions. Proceedings of the 36th International Conference on Machine Learning [Internet]. 2019 [cited 2024 Sep 29]. Available from: https://proceedings.mlr.press/v97/agrawal19a.html
41. Gui J, Andrew AS, Andrews P, Nelson HM, Kelsey KT, Karagas MR, Moore JH. A Robust Multifactor Dimensionality Reduction Method for Detecting Gene–Gene Interactions with Application to the Genetic Analysis of Bladder Cancer Susceptibility. Annals of Human Genetics. 2011;75(1):20–28.
42. Tsang M, Cheng D, Liu Y. Detecting Statistical Interactions from Neural Network Weights. ArXiv [Internet]. 2017 May 14 [cited 2024 Sep 29]; Available from: https://www.semanticscholar.org/paper/Detecting-Statistical-Interactions-from-Neural-Tsang-Cheng/5f85a8eaa7a1a1686f5a2bf721c63e337f03d8eb
43. Nock MK, Millner AJ, Ross EL, Kennedy CJ, Al-Suwaidi M, Barak-Corren Y, Castro VM, Castro-Ramirez F, Lauricella T, Murman N. Prediction of suicide attempts using clinician

assessment, patient self-report, and electronic health records. JAMA network open. American Medical Association; 2022;5(1):e2144373–e2144373.
44. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Scientific reports. Nature Publishing Group UK London; 2018;8(1):7426.
45. Atmakuru A, Shahini A, Chakraborty S, Seoni S, Salvi M, Hafeez-Baig A, Rashid S, San Tan R, Barua PD, Molinari F. Artificial Intelligence-based Suicide Prevention and Prediction: A Systematic Review (2019-2023). Information Fusion. Elsevier; 2024;102673.
46. Bayramli I, Castro V, Barak-Corren Y, Madsen EM, Nock MK, Smoller JW, Reis BY. Temporally informed random forests for suicide risk prediction. Journal of the American Medical Informatics Association. Oxford University Press; 2022;29(1):62–71.
47. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, Nock MK, Smoller JW, Reis BY. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. AJP. 2017 Feb 1;174(2):154–162.
48. Bostwick JM, Pabbati C, Geske JR, McKean AJ. Suicide Attempt as a Risk Factor for Completed Suicide: Even More Lethal Than We Knew. AJP. 2016 Nov 1;173(11):1094–1100.
49. Tanguturi Y, Bodic M, Taub A, Homel P, Jacob T. Suicide risk assessment by residents: Deficiencies of documentation. Academic Psychiatry. Springer; 2017;41:513–519.
50. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, Walsh CG, Brent DA. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA open. Oxford University Press; 2021;4(1):ooab011.
51. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence. Nature Publishing Group UK London; 2019;1(5):206–215.
52. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019 Dec;17(1):195.
53. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521(7553):436–444. PMID: 36774395
54. Sawhney R, Joshi H, Gandhi S, Shah R. A time-aware transformer based model for suicide ideation detection on social media. Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) [Internet]. 2020 [cited 2024 Jul 31]. p. 7685–7697. Available from: https://aclanthology.org/2020.emnlp-main.619/
55. Hsieh TY, Wang S, Sun Y, Honavar V. Explainable Multivariate Time Series Classification: A Deep Neural Network Which Learns to Attend to Important Variables As Well As Time Intervals. Proceedings of the 14th ACM International Conference on Web Search and Data Mining [Internet]. Virtual Event Israel: ACM; 2021 [cited 2024 Jul 31]. p. 607–615. Available from: https://dl.acm.org/doi/10.1145/3437963.3441815
56. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S, Reblitz-Richardson O. Captum: A unified and generic model interpretability library for PyTorch. arXiv:200907896 [cs, stat] [Internet]. 2020 Sep 16 [cited 2021 Feb 11]; Available from: http://arxiv.org/abs/2009.07896
57. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances

in Neural Information Processing Systems 30 [Internet]. Curran Associates, Inc.; 2017 [cited 2019 Jun 9]. p. 4765–4774. Available from: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf PMID: 31050537
58. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [Internet]. arXiv; 2022 [cited 2024 Jul 31]. Available from: http://arxiv.org/abs/2203.05794