Published in final edited form as:
J Clin Psychiatry. ; 84(4): . doi:10.4088/JCP.22m14568.

Leveraging Natural Language Processing to Improve Electronic Health Record Suicide Risk Prediction for Veterans Health Administration Users

Maxwell Levis, PhD,
Joshua Levy, PhD,
Kallisse R. Dent, MPH,
Vincent Dufort, PhD,
Glenn T. Gobbel, PhD, DVM, MS,
Bradley V. Watts, MD, MPH,
Brian Shiner, MD, MPH

VAMC White River Junction, White River Junction, Vermont (Levis, Dufort, Watts, Shiner); Department of Psychiatry, Geisel School of Medicine, Hanover, New Hampshire (Levis, Watts, Shiner); Departments of Pathology and Laboratory Medicine, Geisel School of Medicine, Hanover, New Hampshire (Levy); *VA Serious* Mental Illness Treatment Resource and Evaluation Center, Ann Arbor, Michigan (Dent); Department of Biomedical Informatics, Nashville, Tennessee (Gobbel); VA Office of Systems Redesign and Improvement, White River Junction, Vermont (Watts); National Center for PTSD, White River Junction, Vermont (Shiner).

Abstract

Background: Suicide risk prediction models frequently rely on structured electronic health record (EHR) data, including patient demographics and health care usage variables. Unstructured EHR data, such as clinical notes, may improve predictive accuracy by allowing access to detailed information that does not exist in structured data fields. To assess comparative benefits of including unstructured data, we developed a large case-control dataset matched on a state-of-the-art structured EHR suicide risk algorithm, utilized natural language processing (NLP) to derive a clinical note predictive model, and evaluated to what extent this model provided predictive accuracy over and above existing predictive thresholds.

Methods: We developed a matched case-control sample of Veterans Health Administration (VHA) patients in 2017 and 2018. Each case (all patients that died by suicide in that interval, n = 4,584) was matched with 5 controls (patients who remained alive during treatment year) who shared the same suicide risk percentile. All sample EHR notes were selected and abstracted using

Corresponding Author: Maxwell Levis, PhD, White River Junction VA Medical Center, 163 Veterans Dr, White River Junction, VT 05009 (maxwelle.levis@va.gov).

Ethical Standards:
The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Relevant Financial Relationships: None.

Supplementary Material: Available at Psychiatrist.com.

NLP methods. We applied machine-learning classification algorithms to NLP output to develop predictive models. We calculated area under the curve (AUC) and suicide risk concentration to evaluate predictive accuracy overall and for high-risk patients.

Results: The best performing NLP-derived models provided 19% overall additional predictive accuracy (AUC = 0.69; 95% CI, 0.67, 0.72) and 6-fold additional risk concentration for patients at the highest risk tier (top 0.1%), relative to the structured EHR model.

Conclusions: The NLP-supplemented predictive models provided considerable benefit when compared to conventional structured EHR models. Results support future structured and unstructured EHR risk model integrations.

Suicide is a leading cause of death in the United States, ranking as the second most common cause of death among individuals 10 to 34 years old and fourth among individuals 35 to 44 years old.[1] Nationally, suicide rates have risen from 10.5 per 100,000 in 1999 to 13.5 per 100,000 in 2020.[2] Suicide rates are particularly elevated among Veterans.[3,4] Responding to this concern, the Veterans Health Administration (VHA) has substantially invested in suicide prevention, including establishing the Veterans Crisis Line, staffing designated suicide prevention specialists at each medical center, and establishing suicide prediction and surveillance metrics, helping ensure that individuals receive targeted preventative services.[3,5]

One of the VHA's high-profile contributions toward suicide prevention has been the development of Recovery Engagement and Coordination for Health—Veterans Enhanced Treatment (REACH-VET)[6] program. REACH-VET utilizes a machine-learning–based suicide prediction algorithm to identify and provide outreach to patients at the highest 0.1% risk for suicide in the subsequent month. REACH-VET's algorithm systematically analyzes structured electronic health record (EHR) variables associated with risk for death by suicide including health service use, psychotropic medication, diagnoses, socio-demographics, and the interaction of demographics and diagnoses over a range of time intervals.

Although REACH-VET's algorithm offers an effective model for identifying high-risk patients (eg, REACH-VET's top 0.1% risk tier, above which is considered high-risk, accounts for 2.8% of VHA patient suicides),[7] the majority of VHA patients that die by suicide do not fit within this high-risk tier. As such, REACH-VET fails to detect risk among the preponderance of patients who go on to die by suicide.[8] As a mechanism of expanding predictive accuracy, literature suggests integrating supplementary data formats in addition to structured EHR variables.[9] Prior work evidences the utility of leveraging natural language processing (NLP), a subfield of artificial intelligence that evaluates textual patterns, to develop analyzable variables from unstructured clinical EHR notes.[10–12] Within a prior investigation, using a convenience sample of VHA patients starting PTSD treatment, we found this method allowed access to personalized psychosocial content, including information about patients' interpersonal dynamics and relationships, and offered small predictive benefits over REACH-VET's algorithm.[13]

Although related EHR note text research has increased rapidly,[10,14] few studies have evaluated comparative benefits of including this method alongside existing predictive methods. The present study specifically targets this goal by a sample that was matched on REACH-VET's risk algorithm, allowing analysis of the impact of including EHR note-derived risk variables over and above the REACH-VET's suicide risk prediction method. This study relies on a recent representative sample of Veterans engaged in VHA care who were matched on REACH-VET suicide risk scores, including all patients that died by suicide in 2017 and 2018.

METHODS

Sample Selection

To develop the study sample, we linked VA Corporate Data Warehouse (CDW) EHR with cause of death data from the VA-Department of Defense Mortality Data Repository (MDR)[15] to identify all patients who died by suicide that had at least 1 VHA health care encounter in either 2017 or 2018 (cases = 4,584).

REACH-VET's algorithm automatically evaluates 61 EHR suicide associated structured variables (Supplementary Table 1). REACH-VET's interactive dashboard alerts program coordinators about patients whose suicide risk is within the top 0.1% of risk within the patient's administrative parent facility. Following guidance about rare event matched case-control methods,[16] we matched each case with 5 controls. With support from the VA Office of Mental Health and Suicide Prevention, we identified controls who received care at the same VHA facility during the same interval, shared the same REACH-VET risk percentile at the time of the case's death, and were alive at the time of the case's death (controls = 22,657). For descriptive purposes, we assessed demographic characteristics for the REACH-VET matched sample, including age, race and ethnicity, marital status, military service era, and level of VHA service-connected disability from the month before the matched cases' death date, and calculated standardized mean differences to assess case and control differences.

Corpus Development

We extracted all medical encounter EHR notes from CDW in the year prior to cases' date of death for both cases and matched controls. We excluded notes within 5 days before death as VHA EHR often documents calls to or from families following a death by suicide, and dates of death can sometimes be incorrect by several days. We excluded patients who had more than 6-fold the mean number of notes from the dataset to avoid overweighting patients who had more frequent visits. 2,296,938 notes were selected for analysis. As evidence suggests that suicide risk fluctuates over time,[17] we developed distinct models for different duration intervals in the year before suicide. We accordingly evaluated notes within 5 duration intervals: 30 days before suicide (ie, from 30 days until 5 days before death by suicide), 60 days before death, 90 days before death, 120 days before death, and 1 year before death.

NLP techniques.—We analyzed corpus using Term Frequency–Inverse Document Frequency (TFIDF), an NLP method that measures term importance by calculating their frequency within each individual document within the context of the broader document corpus.[18,19] In TFIDF, term values are weighted proportionally vis-à-vis the amount of times a given term appears in a document and inversely by the total number of documents in the broader corpus that contain the specific term. By addressing total number of documents, TFIDF accounts for terms being more common, reducing weight of very common terms within the corpus and increasing weight of rarer terms specific to a potentially relevant corpus subset. In preparation for TFIDF analysis, notes were tokenized (process of breaking unstructured text into discrete units) and lemmatized (process of grouping different forms of same term so that term can be analyzed as a single entity), and stop-words (terms that are non-impactful, like “a” or “the”) were removed using the NLTK package (Version 3.5).[20] Lemmatization relied on NLTK’s WordNet Lemmatizer.[20] Analysis evaluated up to 3 consecutive terms (n-grams) to better include words indicative of negation (like “non” or “not”).[21] We selected to use TFIDF, as opposed to count matrices, because count data are bounded, which could impact model structure.[22] In contrast, TFIDF, which is normalized through using inverse document frequency, does not have this concern. Additionally, we completed initial analysis using count matrix models, which had consistently lower sensitivity than TFIDF models (Supplementary Table 2). We therefore did not include count matrix methods in subsequent analyses.

We primarily utilized ensemble decision tree algorithms, including classification and regression tree (CART)[23] methodologies, a bagging decision tree approach (Random Forest),[24] and a gradient boosting library (XGBoost)[25] to analyze TFIDF output. CART models learn a series of conditional decision splits based on stochastic selection of predictors and splitting values to form decision trees. Each split further partitions observations into bins where observations demonstrate maximal similarity (eg, cases and controls separately cluster together based on Gini scoring metrics).[26] Random Forest develops multiple decision tree classifiers on bootstrapped dataset subsamples and then averages predictions across trees, each of which cover a biased subset of predictors. The “bagging” of decision tree outputs increases predictive accuracy and reduces overfitting over the whole dataset by maximizing coverage across all predictors and reducing the bias of any given decision tree. XGBoost iteratively morphs subsequent decision trees to account for previous trees’ potential errors, with new models learning from prior models’ errors. XGBoost uses gradient descent to construct new trees based on the residual prediction from the sum of previous trees and the outcome (in the form of the negative binomial likelihood). In CART methods, predictors are premised to interact based on the conditional dependency between subsequent decision splits, and important predictors recur frequently across trees while simultaneously demonstrating the capacity to optimally partition the data. As a comparison, we also utilized Naive Bayes,[27] a comparatively simple probabilistic classifier that premises predictor independence, and Logistic Regression,[28] a widely used classification method that relies on logistic functions to transform linear combinations of independent predictors to a probability between 0 and 1. We utilized class balancing techniques that undersample the predominant class during model training (eg, Random Forest) or reweight the model objective (eg, XGBoost, Logistic Regression).[29] We also ran Brier statistics on all models with and without calibration statistics using isotonic regression. Results were consistent across all methods (Brier score = 0.14). Given this consistency, we did not include the Brier score or calibration in our reporting.

Model development.—For each model, we randomly divided notes into training ($\frac{2}{3}$ of sample) and testing ($\frac{1}{3}$ of sample) sets. We made sure to identically partition each model by setting the random seed to ensure datasets were preserved across algorithms and matching was maintained across partitions. To prevent leakage of information between training and testing data, notes belonging to the same patient were allocated to the same partition. We implemented machine learning models on the training set to optimize model parameters, which were in turn utilized in the testing set to estimate prediction scores. Within the training set, we subjected initial models to randomized search cross-validation scans to refine parameter tunings (hyperparameter tunings are presented in Supplementary Table 3). For cross validation, we performed a group shuffle split (ie, patients isolated to specific training/validation folds) with 5 folds (cv = 5), randomly selecting up to 100 random hyperparameter configurations (n_iter = 100). Cross-validation further subpartitions the training set into multiple training and validation sets (split while accounting for grouping of notes on the patient level) to estimate the overall predictive performance on validation data not used to update model parameters. This approach helps indicate a set of hyperparameters or modeling method that may perform favorably on a held-out test set. Patient-level probabilities for the final test-set were obtained by averaging note-level probabilities within selected time intervals stratified based on group. Predictors were ranked based on feature importance to identify corresponding corpus terms.[30,31] We anticipated normalization and standard scaling would have minimal impact as Random Forest models are relatively invariant to the scale of the features. To evaluate utilization of standard scaling, we ran a sequence of analyses that show that this approach did not offer additional value (Supplementary Table 4).

Model evaluation.—Following prior REACH-VET publications,[8] we calculated the probability of suicide for each patient and assessed suicide risk concentration within our derived models' top 0.1%, 1.0%, and 5.0% predicted probabilities. We also investigated the top 10% of predictive probabilities to better appreciate risk concentration across a broader patient population. Following REACH-VET,[8] we defined the risk concentration as ratio of observed cases to expected distribution of cases, assuming cases have uniform distribution across all REACH-VET risk tiers after matching. This analysis estimates, for example, ratio of cases within the models' highest 10% of predicted probability to the expected number of cases in the top 10%, with the assumption being that the top 10% would contain 10% of cases. As sample was matched on REACH-VET risk, any risk concentration increase above 1 was premised to be indicative of improvement over REACH-VET's algorithm. As in related studies,[7] analyses did not focus on specificity, as that rate remained very close to 1 across sample.

As a measure of overall performance, we calculated area under the receiver operating characteristic curve (AUC) to estimate average sensitivity across a range of predicted probability cutoff points. AUC values range from 0 to 1, where 0.5 indicates no discriminative ability (similar to chance) and 1 indicates perfect predictive accuracy. As sample was matched on REACH-VET risk, any improvement over an AUC of 0.5 was indicative of increased predictive accuracy over REACH-VET's algorithm. To assess statistical significance for AUC statistics, 1,000-sample nonparametric bootstrapping was used to estimate 95% confidence intervals (CIs). Model features were derived by ranking predictor importance and then selecting the top 12 features. Analysis utilized Python (Version 3.8.3) and Scikit-learn (Version 0.23.1)[32] and XGBoost (Version 1.3.3)[25] libraries. A checklist for transparent model reporting and a methods overview diagram are included (Supplementary Figures 1 and 2).

Ethical Standards

All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. Site institutional review board determined that informed consent was not needed, given the study's reliance on retrospective EHR data.

RESULTS

Sample demographics are presented in Table 1. As cases and controls were matched on REACH-VET's scores, a metric based on demographics and service usage, among other risk variables, we expected cases and controls would share very similar demographics. As anticipated, groups were very similar, evidenced by consistently low standard mean difference (SMD) values. Controls included very slightly larger numbers of patients that were 55–74 years old, were Black, were married, and had no service-connected disability. Sample sizes and note counts for all duration intervals are presented in Table 2. The 5 different duration intervals contained subsets of the total sample. Longer duration intervals contained more patients and more notes, relative to shorter duration intervals; for instance, the full year interval contained 4,567 cases with 405,969 notes and 22,616 controls with 1,890,969 notes, while the 30-day interval contained 2,688 cases with 39,893 notes and 13,339 controls with 169,278 notes.

Leveraging note-derived NLP models improved REACH-VET's risk concentration and predictive accuracy in all duration intervals. Full models and evaluation statistics are presented in Table 3. Although all classification algorithms demonstrated benefits, Random Forest offered the most consistent benefit for risk concentration and predictive accuracy metrics. Shorter duration interval models were more predictive than longer duration interval models, with the 30 days back model offering the most added benefit.

The Random Forest model that exclusively evaluated notes 30 days back from date of suicide offered the highest AUC, accounting for 19% overall improvement over REACH-VET's algorithm (0.69 AUC [95% CI, 0.67–0.72]). At the top 0.1%, 1%, 5%, and 10% tiers of highest predicted risk, this model accounted for 0.4%, 3%, 13%, and 24% of VHA patient suicides, respectively; patients who scored in this model's top 10% model accounted for 24% of all suicides, offering 4-fold, 3-fold, 2.6-fold, and 2.4-fold improvement in risk identification over REACH-VET's algorithm at these respective tiers. The Naive Bayes model from this interval offered even higher risk concentration improvement (6-fold), even though its AUC scores were somewhat lower.

Derived text features varied considerably between classification algorithms and between duration intervals, as presented in Table 4. "Suicide" or "suicidal" was identified within all models at each duration interval except for Naive Bayes. Prominent terms associated with known suicide factors were also identified.

DISCUSSION

This study evaluated the added predictive benefits of NLP-derived unstructured EHR suicide risk models over and above REACH-VET's algorithm, a widely used structured EHR-prediction model. The study relied on a REACH-VET risk matched sample, such that any additional predictive accuracy was associated with improvement over and above REACH-VET. Our best models accounted for 6-fold risk concentration improvement for patients in the highest 0.1% risk tier and 19% predictive accuracy sample-wide improvement.

In contrast to prior findings,[33] all classification algorithms had comparative predictive utility as measured by AUC and risk tier statistics. While Naive Bayes' performance had somewhat lower AUC than the other methods, it offered greater risk concentration improvement at the 0.1% risk tier. As a computationally simpler algorithm that processes all terms as opposed to decision tree selections,[27] Naive Bayes ran much more quickly and required less analytic resources. Naive Bayes' output, however, selected somewhat less clinically actionable terms, failing to capture "suicide" within its top 12 features.

NLP-derived models highlighted a variety of themes including suicidality (identified by words like "suicidal," "suicide attempt," and "self-harm"), psychiatric diagnoses (identified by words like "bipolar," "delusional disorder," and "borderline"), mental health services (identified by words like "electroconvulsive," "inpatient psychiatric unit," and "chlorpromazine"), medical issues (identified by words like "prostate," "trach," and "mesothelioma"), interpersonal connections (identified by words like "wife," "brother," and "divorce"), and high-risk behaviors (identified by words like "gun," "alcohol dependence," and "cocaine"). Many of these derived themes have close relevance to known suicide risk factors, including prior suicide attempts,[34] psychiatric diagnoses,[35] mental health service usage,[36] medical diagnoses,[37] interpersonal connections,[38] gun ownership,[39] and alcohol and drug dependence.[40,41] Notably, "electroconvulsive" frequently emerged as a classifier in high-performing models. Though electroconvulsive therapy (ECT) is rarely used and does not appear to prevent suicide in contemporary VA practice,[42,43] it tends to be reserved for the highest-risk patients and even mentioning consideration of ECT in a clinical note appears to be a marker of increased risk.

Developing and implementing suicide risk screening can be arduous and beset with practical challenges.[44] Many psychosocial risk factors have not been developed into structured variables, constraining potential predictive ability of models like REACH-VET. NLP-derived risk modeling presents a pragmatic method to systematically extract and evaluate relevant terms associated with domains where structured variables have not been developed or are not in usage. NLP-derived risk modeling may lessen concerns about patient disclosure and stigma,[45,46] and avoid adding clinical time or cost burden.[47,48]

When comparing this study's population-specific method with our previous more general NLP investigation,[13] the current method offered considerable improvement. Differences may stem from the current study's ability to develop population-specific linguistic references rather than rely on nonclinical semantic resources. This finding accords with related research suggesting that personalized analysis offers increased predictive benefit.[48] Differences between study results could also be associated with respective sample dissimilarities; whereas our prior study only included VHA patients with PTSD diagnoses, the current study included all recent patients, a much larger and more representative population with a much more diverse note corpus.

Whereas our prior studies suggested that samples with longer treatment durations and more notes offered increased predictive accuracy,[12,13] our findings indicate that, when accounting for REACH-VET suicide risk, the opposite was true. We similarly detected a higher proportion of terms directly associated with suicidality in the shorter duration intervals, relative to the full year interval. This may stem from models' difficulty accounting for corpus size and breadth of note noise; whereas shorter interval durations contained fewer notes, longer interval durations contained many more notes. Differences across duration may be indicative of REACH-VET's comparative predictive strength at earlier timepoints in the treatment year relative to the NLP-derived model. This could make sense given that REACH-VET's algorithm incorporates demographic variables that are relatively static and service usage variables that stretch back up to 2 years.

Limitations

We used TFIDF to evaluate text patterns and several leading machine learning classification algorithms to develop predictive models. Alternative analytic and sample weighting methods may have led to contrasting results. Future investigations should develop more nuanced appraisals of change over time. To best replicate prior REACH-VET studies, we abstained from filtering notes by medical encounter type. By not filtering notes, however, our dataset was compromised by a high degree of noise, information that was not associated with suicidality. Filtering strategies could better remove this content and utilize it more meaningfully. Evaluations of risk concentration at the highest risk tier (0.1%) may have been impacted by sample size, a concern that could similarly be levied at prior REACH-VET studies.[8]

Although our NLP-supplemented method provided additional predictive accuracy over and above REACH-VET's algorithm, it is important to reiterate that the current REACH-VET continues to work well and make an impactful contribution toward predicting patients' suicide risk.[49] Moreover, the VHA is engaged in the process of further enhancing subsequent REACH-VET rollouts. As our sample was matched on REACH-VET risk, those designated in the highest risk tier may have benefited from associated suicide prevention services. It is difficult to evaluate to what extent these services impacted sample suicide rates. As such, it is difficult to authoritatively ascertain our predictive model's added impact. Our results suggest that leveraging NLP-derived risk variables could provide substantial benefit for a future REACH-VET rollout.

CONCLUSIONS

Findings suggest unstructured data can aid established structured data-based predictive models. Future studies will evaluate incorporating both methods concurrently to establish whether integrating models achieves further accuracy improvement. A future study could also focus on applying Explainable artificial intelligence (XAI) techniques[50] as well as utilization of a deep learning pipeline such as BERT.[51] Findings support continued NLP investigations to enhance suicide prevention.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

The authors thank the VA Office of Mental Health and Suicide Prevention, Serious Mental Illness Treatment Resource and Evaluation Center, and John F. McCarthy, PhD, MPH, for facilitating the REACH-VET risk matching process. Dr McCarthy has no conflicts of interest to declare.

Funding/Support:

This work was funded by the VISN 1 Career Development Award (Levis). Dr Levy was supported under National Institutes of Health grant subawards P20GM104416 and P20GM130454.

Role of the Funders/Sponsors:

The supporters had no role in the design, analysis, interpretation, or publication of this study.

REFERENCES

1. NIMH. Suicide is a Leading Cause of Death in the United States. https://www.nimh.nih.gov/health/statistics/suicide. Published March 2022.
2. Hedegaard H, Curtin S, Warner M. Increase in Suicide Mortality in the United States, 1999–2018. National Center for Health Statistics (US). https://stacks.cdc.gov/view/cdc/86670. 2020.
3. Office of Mental Health and Suicide Prevention. 2021 National Veteran Suicide Prevention Annual Report. https://www.mentalhealth.va.gov/docs/data-sheets/2021/2021-National-Veteran-Suicide-Prevention-Annual-Report-FINAL-9-8-21.pdf. 2021.
4. Rubin R Task Force to Prevent Veteran Suicides. JAMA. 2019;322(4):295.
5. Carroll D, Kearney LK, Miller MA. Addressing suicide in the veteran population: engaging a public health approach. Front Psychiatry. 2020;11:569069. [PubMed: 33329108]
6. VA Health Care. REACH VET, Predictive Analytics for Suicide Prevention. https://www.dspo.mil/Portals/113/Documents/2017%20Conference/Presentations/REACH%20VET%20Predictive%20Modeling.pdf?ver=2017-08-10-132615-843. 2017.
7. Kessler RC, Hwang I, Hoffmire CA, et al. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans Health Administration. Int J Methods Psychiatr Res. 2017;26(3):e1575. [PubMed: 28675617]
8. McCarthy JF, Bossarte RM, Katz IR, et al. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the US Department of Veterans Affairs. Am J Public Health. 2015;105(9):1935–1942. [PubMed: 26066914]
9. Kessler RC. Clinical epidemiological research on suicide-related behaviors-where we are and where we need to go. JAMA Psychiatry. 2019;76(8):777–778. [PubMed: 31188420]

10. Cook BL, Progovac AM, Chen P, et al. Novel use of Natural Language Processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. Comput Math Methods Med. 2016:8708434.

11. Poulin C, Shiner B, Thompson P, et al. Predicting the risk of suicide by analyzing the text of clinical notes. PLoS One. 2014;9(1):e85733. [PubMed: 24489669]

12. Levis M, Levy J, Dufort V, et al. Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. Psychiatry Res. 2022;315:114703. [PubMed: 35841702]

13. Levis M, Leonard Westgate C, Gui J, et al. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychol Med. 2021;51(8):1382–1391. [PubMed: 32063248]

14. Tsui FR, Shi L, Ruiz V, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open. 2021;4(1):b011.

15. VA DoD. Center of Excellence for Suicide Prevention. Joint Department of Veterans Affairs (VA) and Department of Defense (DoD) Mortality Data Repository - National Death Index (NDI). MIRECC website. https://www.mirecc.va.gov/suicideprevention/documents/VA_DoD-MDR_Flyer.pdf. Accessed December 31, 2020.

16. Lacy MG. Efficiently studying rare events: case-control methods for sociologists. Sociol Perspect. 1997;40(1):129–154.

17. Torous J, Larsen ME, Depp C, et al. Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. Curr Psychiatry Rep. 2018;20(7):51. [PubMed: 29956120]

18. Beel J, Gipp B, Langer S, et al. Research-paper recommender systems: a literature survey. Int J Digit Libr. 2016;17(4):305–338.

19. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manage. 1988;24(5):513–523.

20. Sun W, Cai Z, Li Y, et al. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng. 2018:4302425.

21. Pimpalkar AP, Retna Raj RJ. Influence of pre-processing strategies on the performance of ml classifiers exploiting TF-IDF and BOW features. ADCAIJ Adv Distrib Comput Artif Intell J. 2020;9(2):49–68.

22. Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing [review article]. IEEE Comput Intell Mag. 2018;13(3):55–75.

23. Lemon SC, Roy J, Clark MA, et al. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med. 2003;26(3):172–181. [PubMed: 14644693]

24. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. 1995;1:278–282.

25. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016:785–794.

26. Strobl C, Boulesteix AL, Augustin T. Unbiased split selection for classification trees based on the Gini Index. Comput Stat Data Anal. 2007;52(1):483–501.

27. Zhang H Exploring conditions for the optimality of Naive Bayes. Int J Pattern Recognit Artif Intell. 2005;19(02):183–198.

28. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 1st ed. Wiley; 2013.

29. Susan S, Kumar A. The balancing trick: optimized sampling of imbalanced datasets: a brief survey of the recent state of the art. Eng Rep. 2021;3(4).

30. Qi Z The text classification of theft crime based on TF-IDF and XGBoost model. In: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE; 2020:1241–1246.

31. Wang Y, Wang X-J. A new approach to feature selection in text classification. In: 2005 International Conference on Machine Learning and Cybernetics. IEEE; 2005;6:3814–3819.

32. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/. 2011;12:2825–2830.
33. Kanakaraj M, Guddeti RMR. Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015). IEEE; 2015:169–170.
34. Bostwick JM, Pabbati C, Geske JR, et al. Suicide attempt as a risk factor for completed suicide: even more lethal than we knew. Am J Psychiatry. 2016;173(11):1094–1100. [PubMed: 27523496]
35. Costa L da S, Alencar ÁP, Nascimento Neto PJ, et al. Risk factors for suicide in bipolar disorder: a systematic review. J Affect Disord. 2015;170:237–254. [PubMed: 25261630]
36. Stene-Larsen K, Reneflot A. Contact with primary and mental health care prior to suicide: a systematic review of the literature from 2000 to 2017. Scand J Public Health. 2019;47(1):9–17. [PubMed: 29207932]
37. Henson KE, Brock R, Charnock J, et al. Risk of suicide after cancer diagnosis in England. JAMA Psychiatry. 2019;76(1):51–60. [PubMed: 30476945]
38. Van Orden KA, Cukrowicz KC, Witte TK, et al. Thwarted belongingness and perceived burdensomeness: construct validity and psychometric properties of the Interpersonal Needs Questionnaire. Psychol Assess. 2012;24(1):197–215. [PubMed: 21928908]
39. Anestis MD, Houtsma C. The association between gun ownership and statewide overall suicide rates. Suicide Life Threat Behav. 2018;48(2):204–217. [PubMed: 28294383]
40. Borges G, Bagge CL, Cherpitel CJ, et al. A meta-analysis of acute use of alcohol and the risk of suicide attempt. Psychol Med. 2017;47(5):949–957. [PubMed: 27928972]
41. Pavarin RM, Sanchini S, Tadonio L, et al. Suicide mortality risk in a cohort of individuals treated for alcohol, heroin or cocaine abuse: results of a follow-up study. Psychiatry Res. 2021;296:113639. [PubMed: 33352416]
42. Peltzman T, Gottlieb DJ, Shiner B, et al. Electroconvulsive therapy in Veterans Health Administration hospitals: prevalence, patterns of use, and patient characteristics. J ECT. 2020;36(2):130–136. [PubMed: 31913928]
43. Watts BV, Peltzman T, Shiner B. Electroconvulsive therapy and death by suicide. J Clin Psychiatry. 2022;83(3):21m13886.
44. Kessler RC, Bossarte RM, Luedtke A, et al. Suicide prediction models: a critical review of recent research with recommendations for the way forward. Mol Psychiatry. 2020;25(1):168–179. [PubMed: 31570777]
45. Ganzini L, Denneson LM, Press N, et al. Trust is the basis for effective suicide risk screening and assessment in veterans. J Gen Intern Med. 2013;28(9):1215–1221. [PubMed: 23580131]
46. Husky MM, Zablith I, Alvarez Fernandez V, et al. Factors associated with suicidal ideation disclosure: results from a large population-based study. J Affect Disord. 2016;205:36–43. [PubMed: 27400193]
47. Kleiman EM, Nock MK. New directions for improving the prediction, prevention, and treatment of suicidal thoughts and behaviors among hospital patients. Gen Hosp Psychiatry. 2020;63:1–4. [PubMed: 31229288]
48. Kessler RC, Bernecker SL, Bossarte RM, et al. The role of big data analytics in predicting suicide. In: Passos IC, Mwangi B, Kapczinski F, eds. Personalized Psychiatry. Springer International Publishing; 2019:77–98.
49. McCarthy JF, Cooper SA, Dent KR, et al. Evaluation of the Recovery Engagement and Coordination for Health-Veterans Enhanced Treatment suicide risk modeling clinical program in the Veterans Health Administration. JAMA Netw Open. 2021;4(10):e2129900.
50. Loh HW, Ooi CP, Seoni S, et al. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). Comput Methods Programs Biomed. 2022;226:107161. [PubMed: 36228495]
51. Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv181004805 Cs. arXiv website. https://arxiv.org/abs/1810.04805. Published online May 24, 2019. Accessed February 14, 2022.

52. Andrade C Mean difference, standardized mean difference (SMD), and their use in meta-analysis: as simple as it gets. J Clin Psychiatry. 2020;81(5):20f13681.

Clinical Points

- Although suicide remains a leading cause of death, predicting suicide risk remains challenging. Leveraging electronic health record (EHR) data via natural language processing may offer enhanced accuracy for predicting suicide risk.
- This study illustrates how using unstructured EHR data adds predictive accuracy to the Veterans Health Administration (VHA)'s leading suicide prediction model.
- Derived suicide prediction model offered 19% overall additional predictive accuracy and 6-fold additional risk concentration for users classified as being at the highest risk for suicide using the VHA's model.

Table 1.

Sample Characteristics[a]

	Cases (n = 4,584)	Controls (n = 22,657)	Standardized mean difference
Age			
Mean (SD), y	61	64	
18–34 y, n (%)	560 (12)	2,187 (10)	0.082
35–54 y, n (%)	1,005 (22)	5,065 (22)	0.010
55–74 y, n (%)	1,903 (42)	11,298 (50)	0.168
75+ y, n (%)	1,116 (24)	4,107 (18)	0.152
Sex, n (%) male	4,421 (96)	21,092 (93)	0.151
Race, n (%)			
Pacific Islander	65 (1)	291(1)	0.012
American Indian	40 (1)	219 (1)	0.010
Black—non-Hispanic	244 (5)	2,411 (11)	0.197
White—non-Hispanic	3,761(82)	17,604 (78)	0.109
Hispanic	181 (4)	1,368 (6)	0.096
Marital status, n (%)			
Divorced	1,302 (28)	6,377 (28)	0.006
Married	1,686 (37)	9,938 (44)	0.145
Single	750 (16)	3,286 (15)	0.051
Separated	159 (4)	892 (4)	0.025
Widowed	238 (5)	1,240 (5)	0.013
Service era, n (%)			
Vietnam	1,660 (36)	9,096 (40)	0.081
OEF/OIF/OND	1,522 (33)	7,832 (35)	0.029
Service-connected disability, n (%)			
None	2,006 (44)	12,034 (53)	0.188
0%–60%	1,356 (30)	5,930 (26)	0.076
60%–100%	1,222 (27)	4,693 (21)	0.140
Burden of mental illness, n (%)			
Low: 0 conditions	1,881 (41)	8,352 (37)	0.086
Medium: 1–2 conditions	1,648 (36)	8,449 (37)	0.028
High: 3+ conditions	981 (21)	5,701 (25)	0.089
Burden of physical illness			
Low: 0 conditions	1,572 (34)	6,384 (28)	0.132
Medium: 1–2 conditions	1,760 (38)	9,236 (41)	0.048
High: 3+ conditions	1,064 (23)	6,359 (28)	0.111
Mental health comorbidities			
Depression only	574 (13)	3,867 (17)	0.128
Substance use only	242 (5)	1,001 (4)	0.040
Depression + substance use	403 (9)	2,230 (10)	0.036
Neither	3,365 (73)	15,559 (69)	0.105

[a] Descriptive characteristics of Veterans Health Administration (VHA) patients that died by suicide during 2017 or 2018 (cases) and Recovery Engagement and Coordination for Health—Veterans Enhanced Treatment (REACH-VET)–matched VHA patients that did not die during those intervals (controls). We considered standardized mean difference of 0.2–0.5 as small, values of 0.5–0.8 as medium, and values > 0.8 as large.[52] Following this metric, differences between cases and controls were very small, a finding that makes sense given that cases and controls were matched on REACH-VET suicide risk percentile.

Abbreviation: OEF/OIF/OND = Operation Enduring Freedom/Operation Iraqi Freedom/Operation New Dawn.

Table 2.

Sample Sizes and Note Counts for Duration Intervals[a]

	Cases				Controls			
Days back	N	No. of case notes	Mean (SD) case notes	Median (IQR) case notes	N	No. of control notes	Mean (SD) control notes	Median (IQR) control notes
30	2,688	39,893	13 (32)	5 (275)	13,339	169,278	13 (36)	5 (410)
60	3,384	80,034	24 (48)	9 (294)	16,848	365,262	22 (57)	8 (420)
90	3,707	117,651	32 (60)	12 (322)	18,557	540,250	30 (71)	11 (428)
120	3,911	154,166	39 (76)	15 (356)	19,658	707,881	36 (83)	14 (423)
360	4,567	405,969	89 (162)	37 (394)	22,616	1,890,969	84 (157)	36 (424)

[a] To evaluate changes over time, we developed subsamples based on length of time before cases’ death by suicide. Controls who did not die but shared the same predicted suicide risk percentile and treatment facility as cases were evaluated for the same time duration as matched cases. Our analysis includes 5 different duration intervals: 30 days until 5 days before death, 60 days until 5 days before death, 90 days until 5 days before death, 120 days until 5 days before death, and 360 days until 5 days before death. This table presents the mean, standard deviation (SD), median, and interquartile range (IQR) of the number of patient notes.

Table 3.

Natural Language Processing–Derived Risk Models[a]

	RF	Risk concentration at each risk tier				XG	Risk concentration at each risk tier				LR	Risk concentration at each risk tier				NB	Risk concentration at each risk tier			
Days back	**AUC (95% CI)**	**Top 10%**	**Top 5%**	**Top 1%**	**Top .1%**	**AUhC (95% CI)**	**Top 10%**	**Top 5%**	**Top 1%**	**Top .1%**	**AUC (95% CI)**	**Top 10%**	**Top 5%**	**Top 1%**	**Top .1%**	**AUC (95% CI)**	**Top 10%**	**Top 5%**	**Top 1%**	**Top .1%**
30	0.69 (0.67–0.72)	2.4	2.6	3.0	4.0	0.67 (0.64–0.69)	2.2	2.6	2.0	4.0	0.67 (0.64–0.69)	2.1	2.6	3.0	4.0	0.67 (0.64–0.69)	2.3	2.6	2.0	6.0
60	0.65 (0.62–0.67)	2.0	2.2	2.0	4.0	0.64 (0.62–0.66)	2.0	2.2	3.0	3.0	0.64 (0.62–0.68)	2.1	2.2	3.0	4.0	0.64 (0.61–0.65)	1.9	2.2	2.0	5.0
90	0.64 (0.62–0.66)	2.0	2.0	3.0	4.0	0.64 (0.61–0.66)	2.0	2.4	2.0	3.0	0.63 (0.61–0.65)	1.9	2.0	3.0	4.0	0.63 (0.61–0.65)	1.8	2.0	3.0	5.0
120	0.64 (0.62–0.66	2.0	2.0	2.0	3.0	0.64 (0.62–0.66)	2.0	1.0	2.0	5.0	0.63 (0.61–0.65)	1.9	2.0	2.0	2.0	0.62 (0.60–0.64)	1.7	1.6	2.0	3.0
360	0.62 (0.60–0.64)	1.9	2.2	2.0	1.0	0.62 (0.60–0.64)	1.7	1.0	2.0	1.0	0.62 (0.60–0.64)	1.9	2.0	2.0	1.0	0.61 (0.59–0.63)	1.6	1.6	2.0	2.0

[a] Table presents TFIDF[19] output analyzed by Random Forest (RF),[24] XGBoost (XG),[25] Logistic Regression (LR),[28] and Naïve Bayes (NB)[27] classification models. Each model evaluates notes from different time intervals back from date of death by suicide for cases or matched time points for controls. Overall predictive accuracy is estimated via AUC. Risk concentration for Veterans with the highest predicted risk (10%, 5%, 1%, 0.1%) is also estimated. Following Recovery Engagement and Coordination for Health—Veterans Enhanced Treatment studies, to evaluate risk concertation, we gauged the proportion of death by suicide to the expected proportion of death by suicide assuming uniform sample distribution; ie, among Veterans Health Administration patients who scored within the highest 10% of our model, 24% died by suicide.

Abbreviations: AUC = area under the curve, TFIDF = Term Frequency–Inverse Document Frequency.

Table 4.

Natural Language Processing–Derived Terms

Model	Top 12 term features
TFIDF 30	
RF	Suicide, attempt, suicidal, suicide prevention, hyponatremia, trach, hospice, self-harm, self-inflicted, pain, electroconvulsive, wife
XGBoost	Trach, suicide, suicidal, intensive care unit, psychiatry, suicide prevention, brother, chlorpromazine, suicidal ideation, discharge, suicide attempt, pain
LR	Hyponatremia, trach, electroconvulsive, death, constipation, paranoid, brother, atenolol, suicidal, pneumonia, disturbed, harm
NB	Active, pain, medication, his, time, mouth, group, plan, risk, resident, assessment, report
TFIDF 60	
RF	Suicide, suicidal, hospice, attempt, call, died, mesothelioma, suicide attempt, pain, self-inflicted, against medical advice, suicide prevention
XGBoost	Mobility, time, CIWAar score maximum, suicide, attempt, admitted, thought, diagnosis major depressive, personal hygiene, intensity, gun, cocaine crack
LR	Mesothelioma, suicidal, died, hyponatremia, nasogastric tube, death, gun, inpatient psychiatric unit, bipolar, disorder history, constipation, divorce
NB	Active, pain, medication, time, his, him, care, group, assessment, risk, report, treatment
TFIDF 90	
RF	Suicide, hospice, attempt, cocaine, suicide prevention, electroconvulsive, suicidal, chlorpromazine, call, gun, hyponatremia, suicide attempt
XGBoost	Patient family, mobility, sleep, psychiatric behavioral, fentanyl, harm idea, lack progress, listened passively, co-occurring psychiatric, gastrointestinal, cocaine, suicide
LR	Mesothelioma, hyponatremia, trach, died, electroconvulsive, alcohol dependence, CIWAar, co-occurring psychiatric, bipolar, self-inflicted, gun, suicidal
NB	Active, medication, pain, his, mouth, care, time, group, assessment, report, risk, assessment
TFIDF 120	
RF	Suicide, electroconvulsive, suicide prevention, attempt, bipolar, cocaine, mesothelioma, suicide attempt, suicidal, total parenteral nutrition, gun, killing
XGBoost	Killing self, reassessment, addiction management, psychiatric behavioral health, homelessness, suicide, pressure, electroconvulsive, within patient, post-therapy, plan delineated, verbal
LR	Total parenteral nutrition, electroconvulsive, mesothelioma, trach, fresh start, fluoxetine, hyponatremia, bipolar, co-occurring psychiatric, alcohol dependence, major depressive, borderline personality
NB	Active, medication, pain, his, mouth, care, time, group, risk, call, plan, tablet
TFIDF 360	
RF	Bipolar, suicide, suicide prevention, cocaine, self-harm, electroconvulsive, suicide attempt, incontinent, oncology, exercise, niacin, pain
LR	Safety substance, delusional disorder, divorce, prostate, transgender, declined treatment, self-harm, skin, hyponatremia, trach, risedronate, thymectomy
XGBoost	Addiction management, plan effectiveness, safe environment, lexapro, reassessment continue, restorative, rule compliance, self-care, aid discussion, interaction response, intensive psychotherapy, suicide prevention
NB	Active, medication, pain, you, his, mouth, care, time, group, tablet, plan, report

[a] Table presents selected terms from derived from Random Forest (RF),[24] XGBoost,[25] Logistic Regression (LR),[28] and Naïve Bayes (NB)[27] classification models. Models evaluated TFIDF[19] output from notes from 30 days, 60 days, 90 days, 120 days, and 1 year before death, not including the 5 days closest to death.

Abbreviations: CIWAar = Clinical Institute Withdrawal Assessment Alcohol Scale Revised, TFIDF = Term Frequency–Inverse Document Frequency.