

Research and Applications

Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts

Fuchiang R. Tsui ^{1,2,3,4}, Lingyun Shi^{1,3}, Victor Ruiz^{1,3}, Neal D. Ryan⁵, Candice Biernesser⁵, Satish Iyengar⁶, Colin G. Walsh⁷ and David A. Brent⁵

¹Tsui Laboratory, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, ²Department of Anesthesiology and Critical Care Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, ³Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, ⁴Department of Anesthesiology and Critical Care, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, ⁵Department of Psychiatry, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, ⁶Department of Statistics, School of Arts and Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA and ⁷Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, Tennessee, USA

Corresponding Author: Fuchiang R. Tsui, FAMIA, 2716 South St., Philadelphia, PA 19146, USA; tsuif@email.chop.edu

Received 22 December 2020; Revised 2 February 2021; Editorial Decision 5 February 2021; Accepted 10 February 2021

ABSTRACT

Objective: Limited research exists in predicting first-time suicide attempts that account for two-thirds of suicide decedents. We aimed to predict first-time suicide attempts using a large data-driven approach that applies natural language processing (NLP) and machine learning (ML) to unstructured (narrative) clinical notes and structured electronic health record (EHR) data.

Methods: This case-control study included patients aged 10–75 years who were seen between 2007 and 2016 from emergency departments and inpatient units. Cases were first-time suicide attempts from coded diagnosis; controls were randomly selected without suicide attempts regardless of demographics, following a ratio of nine controls per case. Four data-driven ML models were evaluated using 2-year historical EHR data prior to suicide attempt or control index visits, with prediction windows from 7 to 730 days. Patients without any historical notes were excluded. Model evaluation on accuracy and robustness was performed on a blind dataset (30% cohort).

Results: The study cohort included 45 238 patients (5099 cases, 40 139 controls) comprising 54 651 variables from 5.7 million structured records and 798 665 notes. Using both unstructured and structured data resulted in significantly greater accuracy compared to structured data alone (area-under-the-curve [AUC]: 0.932 vs. 0.901 $P < .001$). The best-predicting model utilized 1726 variables with AUC = 0.932 (95% CI, 0.922–0.941). The model was robust across multiple prediction windows and subgroups by demographics, points of historical most recent clinical contact, and depression diagnosis history.

Conclusions: Our large data-driven approach using both structured and unstructured EHR data demonstrated

accurate and robust first-time suicide attempt prediction, and has the potential to be deployed across various populations and clinical settings.

ABSTRACT

Suicide is a leading cause of death in the United States, and the suicide rate has been increasing in the United States in contrast with a declining rate in many other countries. Moreover, two-thirds of suicide deaths represent their first suicide attempt. To address this public health crisis, we aimed to assess first-time suicide attempt risk by developing machine learning (ML) and natural language processing (NLP) technologies applied to patients' electronic health record (EHR) data.

We proposed a large data-driven approach to aggregate multi-faceted patient information from EHR data, including narrative clinical notes, demographics, medications, diagnosis history, and healthcare-seeking habits. Our data-driven models captured thousands of patient-specific risk factors, including social determinants of health (SDOH), obtained from NLP of clinical notes that contain approximately 80% of EHR data. Our models using clinical notes and other EHR data significantly improved risk prediction performance compared to other models that did not leverage clinical notes.

Our large data-driven approach using ML and NLP of EHR data may be valuable for healthcare professionals to identify patients at risk of first-time suicide attempts accurately. Ultimately, timely risk assessment can facilitate prescribing interventions for reducing first-time suicide attempts, especially in nonmental healthcare settings where mental health services are not readily available.

Key words: suicide attempt, machine learning, natural language processing, electronic health records

INTRODUCTION

The suicide rate in the United States has increased over decades in stark contrast with a decline in many other countries.^{1,2} Accurate and early identification of patients at high suicidal risk is crucial to bolster interventions and evidence-based system changes in healthcare delivery that effectively decrease the rates of suicide attempts and suicide deaths.³ First-time suicide attempts account for two-thirds of suicide decedents, stressing the importance of predicting first-time suicide attempts.⁴⁻⁶

Three barriers have impeded accurate identification of suicidal risk. First, suicidal behavior is relatively rare and predictive models often require large population samples.⁷ Second, risk assessment relies heavily on patient self-report, yet patients may be motivated to conceal their suicidal intentions.^{8,9} Third, prior to suicide attempts, the last point of clinical contact of patients who die by suicide commonly involves providers with varying levels of suicidal-risk assessment training.^{10,11}

The increasing ubiquity of electronic health record (EHR) data and advances in the use of machine learning (ML) present an opportunity to improve the prediction of suicidal behavior.^{12,13} While EHRs contain extensive longitudinal data on a large number of individuals, ML approaches have shown success to handle the challenges inherent to EHR data, such as variable collinearity, high-dimensionality, nonlinear interactions, and missing data.^{14,15}

There have been promising ML applications to predict the risk of suicidal behavior using structured (tabular)^{8,9,12,16} or unstructured (free-text) data separately.¹⁷ Barak-Corren et al.¹² used a Naïve Bayes model to predict suicide attempts and suicide deaths within 3–5 years, achieving an area under the receiver operating characteristic curve (AUC) of 0.77. Walsh et al.^{9,16} used Random Forest models to accurately predict suicide attempts in adolescents and adults in a time window as short as 7 days (AUC \geq 0.83). While Walsh and colleagues' studies show impressive results, they relied on manual chart reviews, which is cost-intensive and unlikely to

scale in routine practice. Simon et al.⁸ reported on the use of penalized least absolute shrinking and selection operator (LASSO) regression in a prospective study of nearly 3 million patients, and achieved AUCs of 0.85 and 0.86 for the prediction of suicide attempts and death within 90 days after the historical most recent (last) point of clinical contact, respectively. The main limitation of this study was that it focused solely on patients with documented mental health diagnoses, which covers only half of all suicide decedents in the United States.¹⁰ Those studies focused on structured EHR data, leaving a gap for incorporating natural language processing (NLP) of unstructured narrative notes into predictive algorithms. The use of unstructured data can be of value to predict suicide risk,¹⁷ especially because approximately 80% of EHR data are locked in narrative form.¹⁸ NLP can identify suicide attempt predictors from clinical notes,^{19,20} such as clinician positive valence assessments²¹ and social determinants of health (SDOH).^{22,23} SDOH are nonmedical factors, such as housing, employment, and family support, which have profound influences on health outcomes.²⁴

This study complements the above-noted pioneering studies in four ways. First, we focused on first-time suicide attempt as a primary outcome. Second, we used both structured and unstructured data to test the extent to which NLP added to the predictive accuracy of algorithms that relied solely on structured data. Third, we developed a large data-driven approach to systematically assess tens of thousands of variables collected from EHR data. Last, we tested the robustness (bias) of our models stratified by patient demographics, last point of clinical contact, and depression diagnosis history.

MATERIALS AND METHODS

The Institutional Review Board at the University of Pittsburgh reviewed the study and determined this to be exempt research.

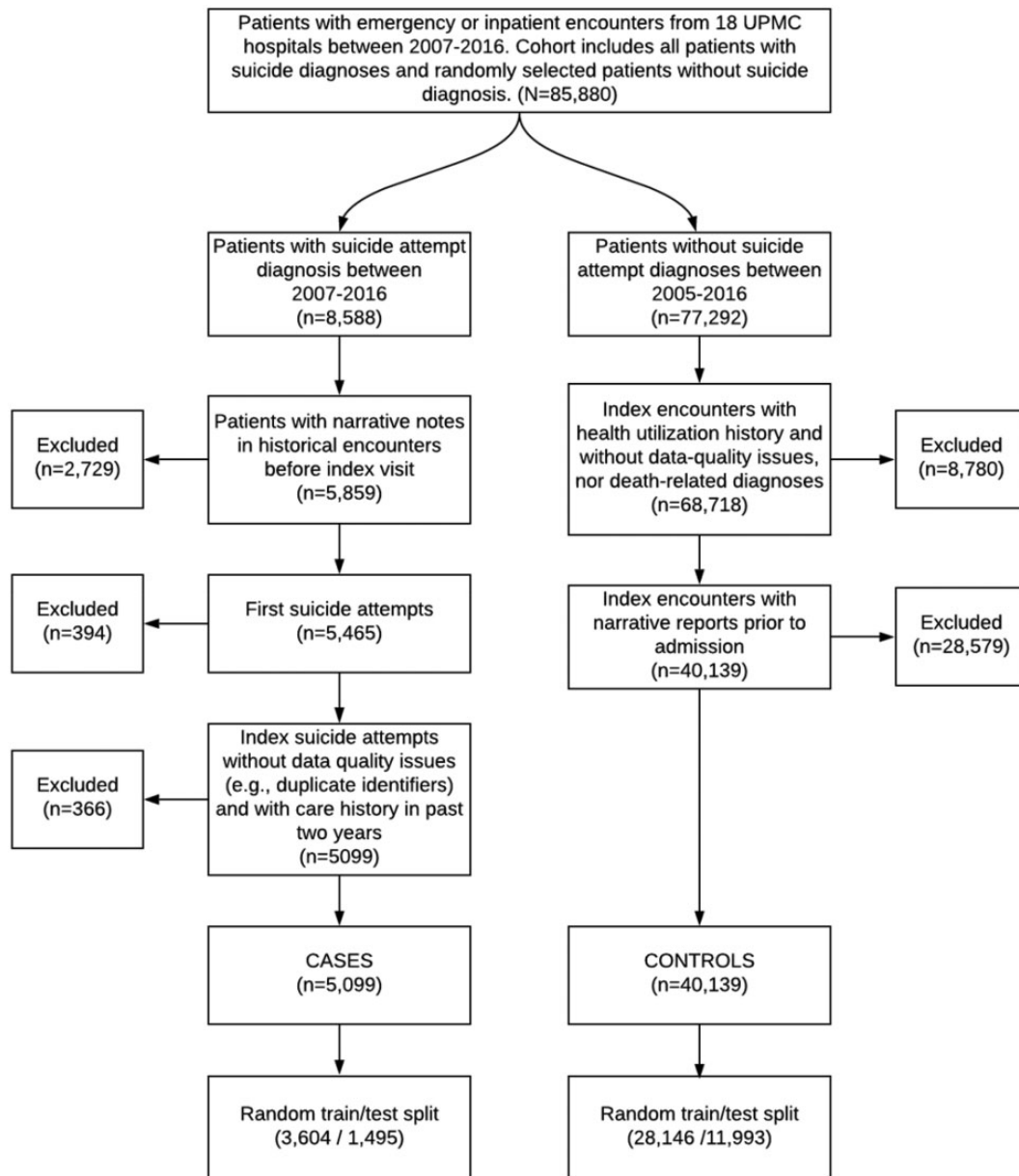


Figure 1. The diagram of cohort identification process. From all inpatient or emergency department visits between 2007 and 2016, our initial cohort comprised 8588 suicide attempt patients based on diagnoses and randomly selected 77 292 patients without any suicide attempt diagnoses. After applying the exclusion criteria, we had a final cohort with 5099 case patients and 40 139 control patients. The cohort was further divided into training and test datasets for model building and testing, respectively. Abbreviation: UPMC, University of Pittsburgh Medical Center.

Patients’ private information was removed by a certified honest broker (PRO17060116).

Dataset

In this case-control study, we identified a cohort from inpatients and emergency department patients aged 10–75 years at any of the 18 hospitals in the University of Pittsburgh Medical Center (UPMC) between January 1, 2007 and December 31, 2016. The dataset was retrieved from the UPMC’s Medical ARchival System.

Figure 1 summarizes the cohort identification process. Case index visits were defined as any emergency department or inpatient encounters with a coded suicide attempt diagnosis based on a list of International Classification of Diseases Ninth (ICD-9) and Tenth

(ICD-10) revision codes identified by Hedegaard et al.²⁵ First-time suicide attempt was chosen as a primary outcome because of its increased correlation with suicide death and future suicide attempts.²⁶ Thus, we excluded suicide attempts where patients had any previous suicide attempt diagnosis from historical encounters within the UPMC hospital network as far back as January 1, 2005 (i.e., 2 years before the start of the study period.) We recognize that this is a silver standard, since previous suicide attempts may not be documented at all, or they might be documented outside of the UPMC hospital network.

Control index visits were defined as emergency department or inpatient encounters without any known suicide attempt diagnosis.²⁵ These were randomly selected from all emergency department and inpatient visits across the 18 hospitals, regardless of demographics

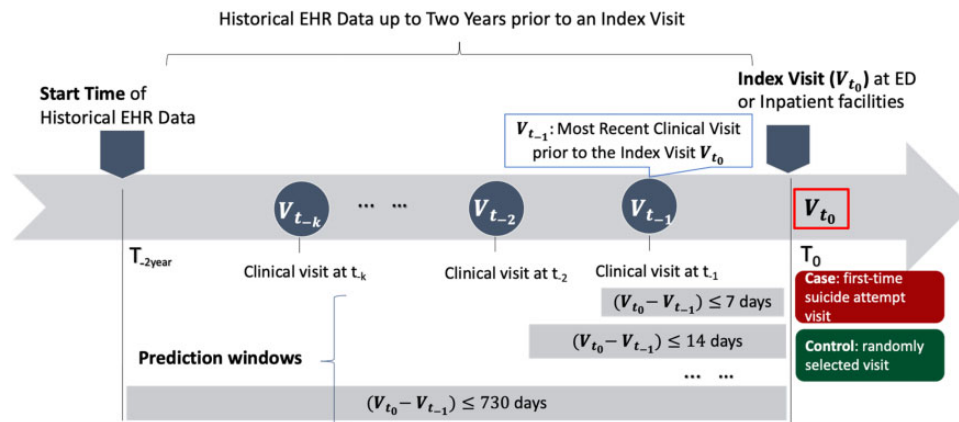


Figure 2. The temporal diagram showing historical electronic health record (EHR) data up to 2 years prior to an index visit at an emergency department or an inpatient facility. A case index visit represents a first-time suicide attempt visit and a control index visit represents a randomly selected visit from controls with longitudinal EHR data. A first-time suicide attempt visit (V_{t_0}) is defined as first known suicide attempt visit between 2005 and 2016; V_{t_0} : the index visit; $V_{t_{-1}}$: last point of clinical contact or last clinical encounter prior to the index visit (V_{t_0}). A prediction window is defined as the time interval between the index visit time (V_{t_0}) and the historical most recent clinical-visit time ($V_{t_{-1}}$) prior to the index visit.

or services provided; a ratio of 9 controls per case was selected given the balance between a sample size and prevalence of suicide attempts. suicide attempt prevalence in predictive modeling literature ranged from 36 to 62 900 per 100 000 individuals.⁷ Outpatient visits were excluded from case and control index visits because the vast majority (98%) of suicide attempts were found in either emergency department or inpatients settings in our database. Moreover, the literature shows that patients seen in these two settings have an increased risk for suicide and suicidal behavior.^{27,28}

We retrieved 2-year historical data prior to an individual index visit for risk prediction. Figure 2 shows a temporal diagram of a patient's historical clinical encounters related to the index visit. We excluded from analysis any cases and controls that had no historical clinical visits or without any clinical notes in the 2-years window; thus, each cohort patient had a minimum of one clinical encounter with at least one clinical note as shown in Figure 1.

To assess the validity of suicide-related ICD diagnoses across 77 ICD-9/10 codes, a total of 151 records were reviewed independently by three reviewers (NR, DB, and CB). Ambiguous cases were reviewed and consensus was reached to make a final determination of the presence or absence of an attempt. Suicide attempt was confirmed in 112 records, whereas in 39 records there was probable but not definitive evidence of a suicide attempt (e.g., self-inflicted gunshot wound that the patient claimed was an accident). Therefore, ICD suicide-related diagnoses were utilized as coded.

Data sources

The study comprised five data sources. The first data source was unstructured data (clinical notes) comprising history and physical examination, progress, and discharge summary notes. These notes were generated by clinicians, for example, attending physicians and residents during clinical encounters. The remaining four were structured data including demographics, diagnoses, healthcare utilization data (e.g., the number of previous inpatient visits), and medications.

Missing data and data imputation

We identified a large set of variables from the cohort and assigned observed values for the variables from EHRs. When observed values were missing, we added an “unknown” category to multi-category

variables such as race and insurance, and to all the variables from unstructured data to avoid imputation bias. For example, if a patient's race was not observed in the EHR data, we assigned the race variable an “unknown” category. However, for those binary variables without observed values from diagnoses and medications, we imputed them with “absent” (“no”) values; for example, a patient without a diagnosis ICD-10 code F32.9 would have a record of F32.9 with an “absent” value.²⁹

Natural language processing

We employed the cTAKES³⁰ 4.0.0 open-source NLP tool to process narrative notes without any preprocessing steps. This tool has been widely employed and rigorously evaluated in extracting clinical findings³¹ that can be used in ML models for prediction tasks (Figure 3).³² The cTAKES tool can process a variety of narrative notes, such as history and physical, progress, discharge summary, and radiology notes.^{31,33,34} This process consists of extracting clinical concepts that are then annotated with Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS). UMLS, developed by the National Library of Medicine, unifies standard vocabularies used in healthcare and biomedical sciences into a single comprehensive thesaurus and ontology system. For example, when cTAKES extracts *depressive disorder* in a note, it annotates that concept with its correspondent CUI C0011581 code and assigns a polarity to each annotation, that is, whether the identified concept is a positive (present) or negative (absent) finding. We updated cTAKES 4.0.0 with the 2017AA release of the UMLS Knowledge Sources.³⁵ Those cTAKES extracted CUIs from 2-year historical notes served as features (variables) for model construction. Whenever a CUI was extracted in multiple encounters in the patient's longitudinal EHR data, we chose the polarity (present/absent) of their most recent mention. We chose cTAKES over MetaMap, a popular open-source NLP tool from NLM, due to cTAKES' slightly better performance in literature.³⁶

Machine learning and feature engineering

We used a large data-driven approach, employing all features from EHR data, applied to four common ML algorithms in the primary analysis: Naive Bayes (NB), least absolute shrinkage and selection

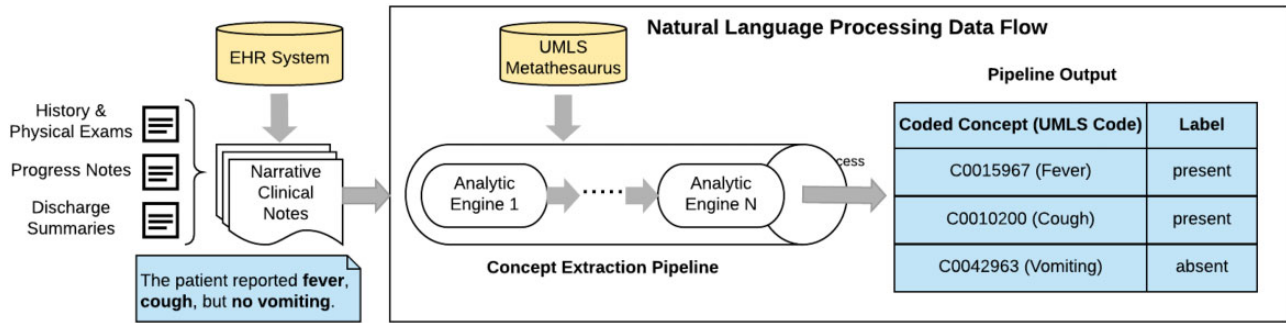


Figure 3. The process flow of a medical natural language processing (NLP) pipeline, which transforms a narrative sentence in a clinical note to structured outcomes. For example, the sentence has three symptoms (fever, cough, and vomiting) and vomiting concept is negated. Negated concepts are common in clinical notes.

operator (LASSO) regression, random forest (RF), and the ensemble of extreme gradient boosting (EXGB).^{15,37,38} All the models estimated the posterior probability of suicide attempt based on historical data extracted within 2 years prior to the index visit, that is, $P(\text{suicide attempt} | 2\text{-year EHR data prior to the index visit})$.

To effectively process a large number of features obtained from EHRs, we employed three feature engineering frameworks: filter, wrapper, and embedded frameworks. For NB, we filtered features based on information gain scores.³⁹ EXGB and RF used a wrapper framework, which merges feature search and modeling together, and Gini feature importance was used for feature search.⁴⁰ LASSO used embedded feature selection framework through L1 regularization.⁴¹

Primary analysis: Model training and testing

For the primary analysis that estimated each model’s capability for predicting first-time suicide attempts, we randomly sampled 70% of the cohort as a “training” dataset, while the remaining 30% served as a “test” (blind) dataset. All of the models were trained and tuned in the training dataset exclusively. *Model performance was reported exclusively on the test dataset.*

Prediction evaluation metrics

The predictive models were evaluated in the test dataset using three standard evaluation metrics: AUC, sensitivity, and specificity. Sensitivity and specificity were measured at different binary-classification thresholds that the authors deemed relevant for clinical practice. In specific, we fixed either sensitivity or specificity at 90% and 95% and measured the value of the metric that was not fixed. This results in four possible combinations (e.g., sensitivity = 90% with specificity = 81.24%; Table 1). The adjusted *P*-values for paired AUC comparisons were conducted using two-sided DeLong tests with Bonferroni multiple-hypotheses corrections.⁴²

Measure of NLP impact to prediction performance

To measure how NLP impacted the predictive performance, two full-feature models (e.g., EXGB and LASSO) trained from both structured and unstructured data were compared with two structured-feature-only models trained from structured data exclusively (i.e., S-EXGB and S-LASSO). The NB and RF were excluded in this comparative analysis due to their lesser performance in the primary analysis. All the three evaluation-metrics were employed to measure the NLP impact. Both S-EXGB and S-LASSO were built to best estimate suicide attempt risk for the scenario when no narrative

notes were available for analysis. All full-feature and structured-feature-only models were tested in the test dataset.

Prediction windows

A prediction window was defined as the interval between an index visit time and the time of last clinical contact prior to the index visit (last point of clinical contact). We selected four windows—7, 30, 90, and 730 days (Figure 2)—based on previous studies.^{9,16}

Analysis of model features

We evaluated each feature (variable) from the best scoring model by using three statistical metrics: feature scores (e.g., Gini feature importance),^{43,44} unadjusted odds ratios (ORs),⁴⁵ and adjusted odds ratios (aORs).⁴⁶ ORs were obtained by univariate logistic regression, whereas aORs controlled demographic confounders using multi-variate regression, adjusted for age, sex, race, and health insurance type.

Model robustness (Bias) evaluation

In addition to overall model performance, it becomes increasingly important to test whether ML models behave robustly across relevant subgroups.^{47,48} Robustness (or bias) was evaluated by comparing model AUCs in the test dataset across six axes (18 subgroups): age (<35, ≥35 years old), sex (male, female), race (White, Black, Other, Unknown), insurance type (commercial, Medicaid, Medicare, self-pay, others), last point of clinical contact prior to the index visit (outpatient, emergency department, inpatient), and depression diagnosis history (present, absent.)

Sensitivity analysis of NLP-based suicide attempt concept extraction

The accuracy and completeness of ICD codes may be limited.^{49,50} We performed a sensitivity analysis to explore the possibility that cases and controls may have been mislabeled by diagnosis ICD codes. First, we identified mentions of suicide attempts from unstructured notes based on extracted CUIs from NLP. We then removed patients with suicide mentions from the dataset, thus reducing the possibility of including suspected follow-up (not first) suicide attempts in the case group or patients with suicide history in the control group.

Table 1. Predictive performance^a across four predictive models using full-feature (structured and unstructured) data with 4 prediction windows (7, 30, 90, and 730 days)

Predictive model (number of features)	EXGB (<i>n</i> = 1726)	LASSO (<i>n</i> = 484)	NB (<i>n</i> = 2126)	RF (<i>n</i> = 1617)	
Prediction window: ≤ 7 days	Cases/Controls	273/3980			
	AUC (95% CI) ^b	0.9298 (0.9153 – 0.9445)	0.9042 (0.8841 – 0.9231)	0.7580 (0.7340 – 0.7808)	0.9055 (0.8882 – 0.9220)
	<i>P</i> -value	Ref	<.001	<.001	<.001
	4 sets of sensitivity/specificity (%)	90.00/79.92	90.00/71.66	90.00/46.63	90.00/72.19
		95.00/69.45	95.00/58.47	95.00/22.02	95.00/60.95
	75.09/90.00	71.43/90.00	25.79/90.00	69.60/90.00	
	64.84/95.00	59.71/95.00	12.90/95.00	57.51/95.00	
Prediction window: ≤ 30 days	Cases/Controls	625/7900			
	AUC (95% CI)	0.9320 (0.9222 – 0.9409)	0.9086 (0.8964 – 0.9205)	0.7663 (0.7500 – 0.7825)	0.9002 (0.8873 – 0.9123)
	<i>P</i> -value	Ref	<.001	<.001	<.001
	4 sets of sensitivity/specificity (%)	90.00/81.24	90.00/74.77	90.00/45.72	90.00/71.73
		95.00/70.49	95.00/58.59	95.00/22.94	95.00/59.08
	77.44/90.00	73.28/90.00	27.03/90.00	70.40/90.00	
	65.92/95.00	60.32/95.00	13.51/95.00	54.24/95.00	
Prediction window: ≤ 90 days	Cases/Controls	971/10108			
	AUC (95% CI)	0.9286 (0.9210 – 0.9361)	0.9031 (0.8928 – 0.9127)	0.7643 (0.7514 – 0.7767)	0.8848 (0.8745 – 0.8950)
	<i>P</i> -value	Ref	<.001	<.001	<.001
	4 sets of sensitivity/specificity (%)	90.00/80.05	90.00/73.16	90.00/49.69	90.00/66.56
		95.00/69.87	95.00/59.81	95.00/24.99	95.00/54.00
	76.73/90.00	71.37/90.00	25.88/90.00	65.29/90.00	
	64.06/95.00	57.16/95.00	12.94/95.00	49.43/95.00	
Prediction window: ≤ 730 days	Cases/Controls	1495/11993			
	AUC (95% CI)	0.9190 (0.9118 – 0.9255)	0.8926 (0.8844 – 0.9010)	0.7554 (0.7448 – 0.7653)	0.8645 (0.8550 – 0.8730)
	<i>P</i> -value	Ref	<.001	<.001	<.001
	4 sets of sensitivity/specificity (%)	90.00/76.88	90.00/70.15	90.00/51.10	90.00/62.51
		95.00/65.85	95.00/57.45	95.00/28.16	95.00/50.47
	73.78/90.00	67.56/90.00	23.84/90.00	58.80/90.00	
	60.13/95.00	51.97/95.00	11.92/95.00	43.34/95.00	

Note: EXGB, Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest (RF) applied further feature engineering frameworks (wrapper and embedded) to get a smaller number of features. The number listed in the parentheses associated with each model represents the final number of features used in the model. A boldfaced number represents the best AUC within each prediction window compared to other models.

EXGB: Ensemble of eXtreme Gradient Boosting; LASSO: Least Absolute Shrinkage and Selection Operator; NB: Naïve Bayes; RF: Random Forest.

^aAll the models were trained in a training dataset and tested in a test (blind or hold-out) dataset. The evaluation metrics include the area under the receiver operating characteristic curve (AUC) with 95% confidence interval, sensitivity (or recall) and specificity (or precision). Each *P*-values was tested with respect to the AUC of Ensemble of eXtreme Gradient Boosting (EXGB) in the same prediction window. For each model, we started a total of 2126 features (including 215 social features) after applying feature filter.

^bAll 95% confidence intervals were measured through 2000 stratified bootstrap replicates.

RESULTS

Cohort data

We initially identified 8588 case patients with 12 446 suicide attempt visits (some patients had multiple attempts) and randomly selected 77 292 control patients with 77 292 index visits (one index visit per control) from emergency department and inpatient settings during the study period (Figure 1). For those patients with multiple attempts, we only selected the first-attempt visit per patient in the dataset. After applying the aforementioned exclusion and inclusion criteria (e.g., index visit inpatient or emergency department, at least one previous visit with a clinical note within 2 years of the index visit), the final cohort consisted of 5099 (11.3%) cases and 40 139 (88.7%) controls. The test dataset had 1495 (29.3% of cohort cases) cases and 11 993 (29.9% of cohort controls) controls.

The majority of suicide attempts were patients aged 15–54 years (83.8%), an age group that had more females than males (59.5%) (Supplementary Figure S1). The majority (65.8%) of cohort cases

made a suicide attempt within 90 days of their historical most recent clinical visit prior to the index visit (Supplementary Figure S2).

The structured EHRs had 5 738 154 records that included medications, demographics, healthcare utilization data, and diagnoses. The unstructured EHRs had 798 665 narrative clinical notes. After we applied the cTAKES to all the clinical notes, there were a total of 54 651 features from both structured (13 873; 25.4%) and unstructured (40 778; 74.6%) EHR data.

Model prediction performance

All of the models were constructed using the training dataset with up to 2126 features from structured and unstructured data using the feature filtering method. The models were tested in the test dataset, and Figures 4 and 5 show the prediction performance: Receiver operating characteristic (ROC) curves and AUCs of the four predictive models in 30- and 730-day prediction windows, respectively. Detailed prediction performance across four prediction windows (7,

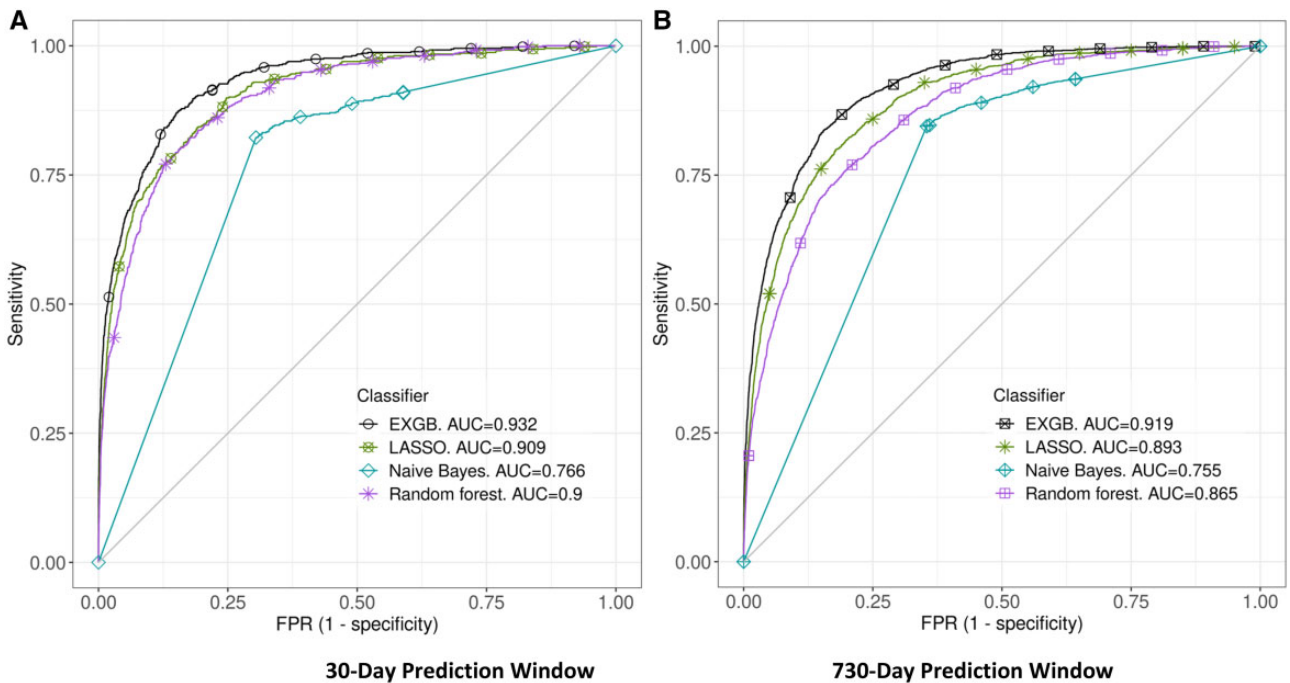


Figure 4. Receiver Operating Characteristic (ROC) curves of four ML models. Plots A and B show ROCs in 30- and 730-day prediction windows, respectively. Abbreviations: EXGB, Ensemble of eXtreme Gradient Boosting; LASSO, Least Absolute Shrinkage and Selection Operator.

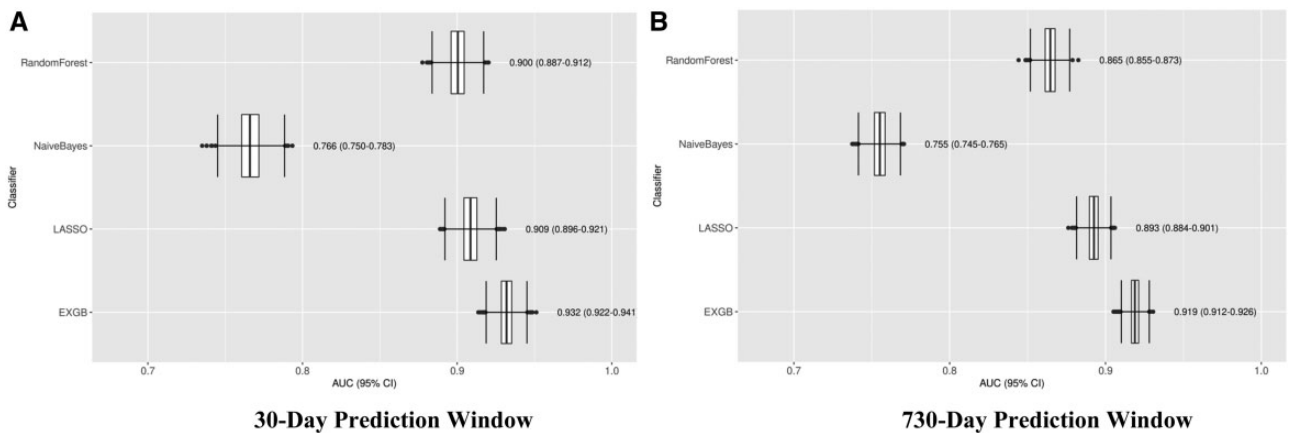


Figure 5. Plots of predictive model accuracy, measured by the area under a receiver operating characteristic curve (AUC), among 4 predictive models. Plot (A) shows model performance in 30-day prediction window. Plot (B) shows model performance in 730-day prediction window. Abbreviations: EXGB, Ensemble of eXtreme Gradient Boosting; LASSO, Least Absolute Shrinkage and Selection Operator.

30, 90, and 730 days) are in Table 1; among the four models, EXGB consistently had the best performance across all windows, with AUCs ranging from 0.919 (95% CI, 0.912–0.926, in 730-day window) to 0.932 (95% CI, 0.922–0.941, in 30-day window). In the 30-day prediction window, the model reached 90% sensitivity at 81% specificity and 77% sensitivity at 90% specificity. While LASSO regression model performed well, EXGB model outperformed it [30-day: AUC 0.909 (95% CI, 0.896–0.921) vs. 0.932 (95% CI, 0.922–0.941); 730-day: AUC 0.893 (95% CI, 0.884–0.901) vs. 0.919 (95% CI, 0.912–0.926)] (Table 1).

NLP impact to prediction

Predictive models using full-feature data (comprising both structured and unstructured data) significantly outperformed ($P < .001$)

the models using only structured features. Table 2 shows the detailed pairwise comparison among gradient boosting and regression models. In the 30-day prediction window, the two structure-data-only models (S-EXGB and S-LASSO) had lower AUCs with statistical significance compared to the full-feature-data models using full-feature data, for example, AUCs 0.901 (S-EXGB) vs. 0.932 (EXGB), $P < .001$. Moreover, EXGB using full-feature data had the best sensitivity-specificity performance compared to the other models (Table 2).

Model robustness (Bias) evaluation results

Overall, EXGB was robust and performed similarly across 18 subgroups stratified by demographics (age, sex, race, and health insurance type), last point of clinical contact, and depression diagnosis

Table 2. Impact of NLP on suicide attempt prediction^a

Model versus prediction window		Gradient boosting model		Regression model	
		Full-feature EXGB Model (1766 features)	Structured-feature-only S-EXGB Model (422 features)	Full-feature LASSO Model (484 features)	Structured-feature-only S-LASSO Model (192 features)
Prediction window ≤ 7 days	Cases/Controls, <i>n</i>	273/3980			
	AUC (95% CI) ^b	0.9298 (0.9153 – 0.9445)	0.9037 (0.8838 – 0.9220)	0.9042 (0.8841 – 0.9231)	0.8763 (0.8515 – 0.8989)
	<i>P</i> -value	Ref	<.001	Ref	<.001
	4 sets of sensitivity/specificity (%)	90.00/79.92	90.00/71.61	90.00/71.66	90.00/60.00
		95.00/69.45	95.00/53.72	95.00/58.47	95.00/36.03
	75.09/90.00	70.70/90.00	71.43/90.00	68.50/90.00	
	64.84/95.00	58.24/95.00	59.71/95.00	55.68/95.00	
Prediction window ≤ 30 days	Cases/Controls, <i>n</i>	625/7900			
	AUC (95% CI)	0.9320 (0.9222 – 0.9409)	0.9007 (0.8880 – 0.9129)	0.9086 (0.8964 – 0.9205)	0.8842 (0.8695 – 0.8984)
	<i>P</i> -value	Ref	<.001	Ref	<.001
	4 sets of sensitivity/specificity (%)	90.00/81.24	90.00/70.20	90.00/74.77	90.00/67.51
		95.00/70.49	95.00/56.15	95.00/58.59	95.00/46.72
	77.44/90.00	72.00/90.00	73.28/90.00	69.28/90.00	
	65.92/95.00	57.76/95.00	60.32/95.00	55.68/95.00	
Prediction window ≤ 90 days	Cases/Controls, <i>n</i>	971/10108			
	AUC (95% CI)	0.9286 (0.9210 – 0.9361)	0.8963 (0.8862 – 0.9062)	0.9031 (0.8928 – 0.9127)	0.8763 (0.8634 – 0.8879)
	<i>P</i> -value	Ref	<.001	Ref	<.001
	4 sets of sensitivity/specificity (%)	90.00/80.05	90.00/67.82	90.00/73.16	90.00/65.58
		95.00/69.87	95.00/55.53	95.00/59.81	95.00/45.31
	76.73/90.00	69.62/90.00	71.37/90.00	66.53/90.00	
	64.06/95.00	54.99/95.00	57.16/95.00	51.18/95.00	
Prediction window ≤ 730 days	Cases/Controls, <i>n</i>	1495/11993			
	AUC (95% CI)	0.9190 (0.9118 – 0.9255)	0.8830 (0.8744 – 0.8918)	0.8926 (0.8844 – 0.9010)	0.8622 (0.8522 – 0.8719)
	<i>P</i> -value	Ref	<.001	Ref	<.001
	4 sets of sensitivity/specificity (%)	90.00/76.88	90.00/65.71	90.00/70.15	90.00/62.11
		95.00/65.85	95.00/52.31	95.00/57.45	95.00/43.38
	73.78/90.00	64.82/90.00	67.56/90.00	60.27/90.00	
	60.13/95.00	49.23/95.00	51.97/95.00	44.95/95.00	

AUC: area under the curve; CI: confidence interval.

^aFull-feature (including structured and unstructured features) models and structured-feature-only models were compared. Two full-feature models were included: Ensemble of eXtreme Gradient Boosting (EXGB) model and the Least Absolute Shrinkage and Selection Operator (LASSO). Two structured-feature-only models were included: the Structured-EXGB (S-EXGB) model and the structured-LASSO (S-LASSO) model. All the models were trained in a training dataset and tested in a test (blind or hold-out) dataset. Full-feature models performed significantly better than structured-feature-only models. We chose four common sets of metrics (i.e., sensitivity and specificity) based on two sensitivities at 90% and 95% and two specificities at 90% and 95%; given a pre-selected sensitivity or specificity, the corresponding metrics were measured from the test dataset.

^bAll 95% confidence intervals were measured through 2000 stratified bootstrap replicates.

history. In the 30-day prediction window, AUCs ranged from 0.875 to 0.949 across 18 subgroups. In the 730-day prediction window, AUC ranged from 0.82 to 0.942. [Figure 6](#) shows individual AUCs across the 18 subgroups and the two prediction windows.

Predictive features

EXGB identified 1726 features (from the initial 2126 features) from the five data sources: clinical notes ($n=1299$), demographics ($n=4$), diagnoses ($n=227$), healthcare utilization ($n=3$), and medication ($n=193$). cTAKES generated features from clinical notes exclusively. [Tables 3](#) and [4](#) list demographics and the 10 most significant risk and protective features (among diagnoses, healthcare utilization, medication, and NLP of clinical notes) from the best XGB model with ORs and aORs, respectively. Three critical risk factors with both ORs and aORs > 2 from NLP were suicide attempt (which was picked up in clinical notes but not recorded as a

previous diagnosis), depressive disorder, and drug abuse. While all of the cases were first-time attempters according to structural diagnosis data, 12.8% (654) of cases and 1.5% (616) of controls had a previous suicide attempt concept according to NLP assessment of unstructured narrative notes. With respect to service use within 2 years prior to an index visit, a previous emergency department visit was a risk factor, whereas any outpatient healthcare visit was protective ([Table 4](#)). [Supplementary Table S1](#) lists top 10 features from the LASSO, S-LASSO, and S-EXGB models.

Social determinants of health

Among 46 SDOH within the 1299 NLP features, we identified increased risk and protective factors with prevalence $\geq 1\%$ and statistically significant odds ratios (i.e., 95% CI that excluded 1.0): one risk factor for suicide attempt (*divorced marital status*) and three

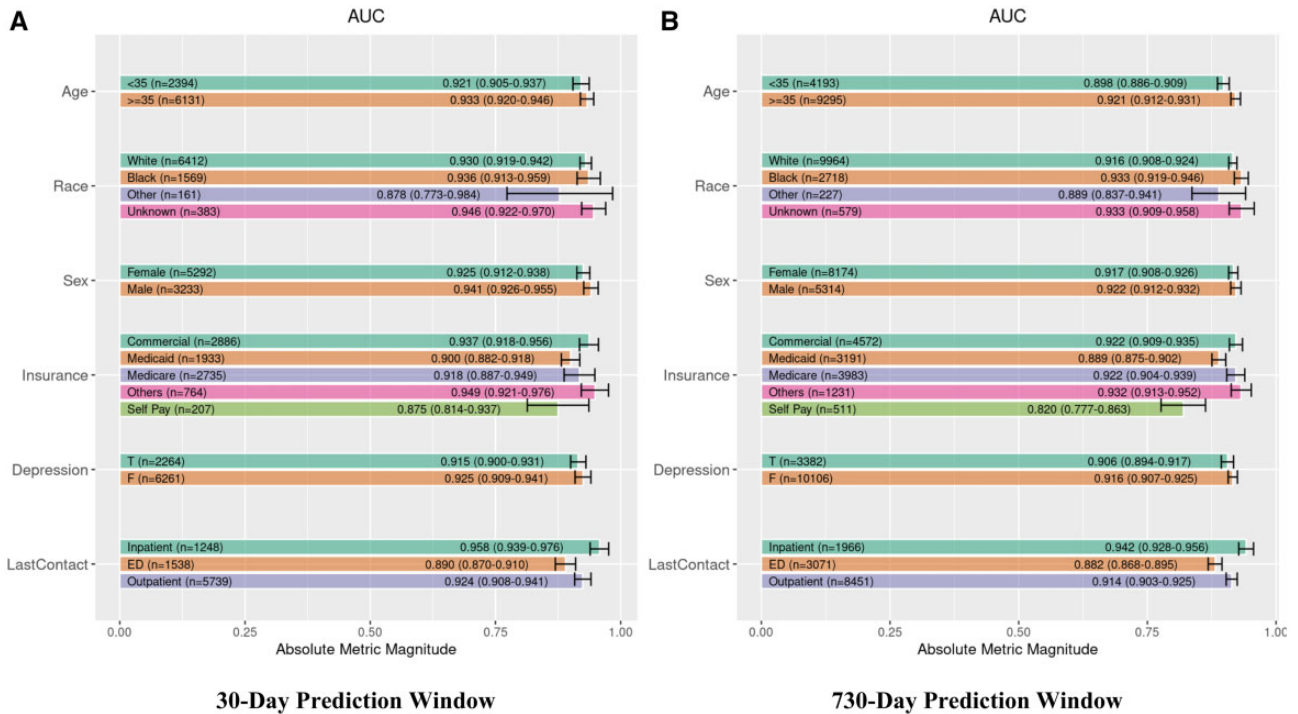


Figure 6. Robustness analysis of Ensemble eXtreme Gradient Boosting (EXGB) model across 18 subgroups based on demographics (age, race, gender, insurance), depression diagnosis, and point of historical most recent clinical contact. Plot (A) shows the EXGB performance in 30-day prediction window. Plot (B) shows the EXGB performance in 730-day prediction window. Age was measured in years. Abbreviations: T, present; F, absent; LastContact, point of historical most recent clinical contact.

Table 3. Unadjusted and adjusted odds ratios^a with 95% CI for demographic features

Feature	Cases, No. (%) (n = 5099)	Controls, No. (%) (n = 40 139)	Unadjusted Odds Ratio (95% CI)	Adjusted Odds Ratio(95% CI)
Demographic: sex				
Male	2133 (41.83)	15 839 (39.46)	1.1 (1.04–1.17)	1.36 (1.27–1.44)
Female	2966 (58.17)	24 300 (60.54)	1 (ref)	ref
Demographic: age				
10–14	253 (4.96)	1481 (3.69)	7.16 (5.76–8.9)	7.62 (6.13–9.48)
15–24	1 368 (26.83)	4697 (11.70)	12.19 (10.15–14.63)	13.93 (11.59–16.74)
25–34	1 103 (21.63)	5281 (13.16)	8.74 (7.27–10.51)	9.6 (7.98–11.54)
35–44	901 (17.67)	5485 (13.67)	6.87 (5.71–8.28)	7.4 (6.14–8.92)
45–54	903 (17.71)	8217 (20.47)	4.6 (3.82–5.54)	4.81 (4–5.8)
55–64	439 (8.61)	9436 (23.51)	1.95 (1.6–2.37)	1.98 (1.62–2.41)
65+	132 (2.59)	5542 (13.81)	1 (ref)	ref
Demographic: race				
Black	885 (17.36)	8075 (20.12)	0.84 (0.78–0.91)	0.68 (0.63–0.73)
Not specified	155 (3.04)	1300 (3.24)	1.33 (1.11–1.58)	1.12 (0.93–1.34)
Other	152 (2.98)	878 (2.19)	0.91 (0.77–1.08)	0.78 (0.66–0.93)
White	3907 (76.62)	29 886 (74.46)	1 (ref)	ref
Demographic: insurance				
Medicaid	2180 (42.75)	8398 (20.92)	2.85 (2.64–3.07)	2.59 (2.4–2.8)
Medicare	789 (15.47)	12 663 (31.55)	0.68 (0.62–0.75)	1.35 (1.22–1.49)
Others	530 (10.39)	3617 (9.01)	1.61 (1.44–1.79)	1.62 (1.45–1.81)
Self-pay	314 (6.16)	1356 (3.38)	2.54 (2.22–2.91)	2.1 (1.83–2.42)
Commercial	1286 (25.22)	14 105 (35.14)	1 (ref)	ref

EXGB: Ensemble XGB; XGB: extreme gradient boosting.

^aThe adjusted odds ratios (aORs) were estimated while controlling for sex, age, race, and insurance. The boldfaced numbers represent the highest OR/aOR in the increased risk categories or the lowest OR/aOR in the decreased risk categories.

Table 4. Unadjusted and adjusted odds ratios in top 10 increased-risk and decreased-risk features with status present (true)

Feature	Cases, No. (%) (n = 5099)	Controls, No. (%) (n = 40 139)	Unadjusted Odds Ratio (95% CI ^a)	Adjusted Odds Ratio ^b (95% CI ^a)
Top 10 increased-risk features ^c from best model in EXGB with all odds ratios and 95% CIs > 1 ranked by feature importance				
One or more emergency department visits in 2 years	4196 (82.29)	22 405 (55.82)	3.68 (3.41–3.96)	3.06 (2.83–3.3)
Other psychological or physical stress not elsewhere classified (ICD-9 62.8^a)	748 (14.67)	481 (1.20)	14.17 (12.58–15.95)	13.17 (11.62–14.93)
Episodic mood disorders (ICD-9 296 ^a)	1901 (37.28)	4051 (10.09)	5.30 (4.96–5.65)	5.36 (5–5.74)
Suicide attempt (UMLS C0038663 ^d)	654 (12.83)	616 (1.53)	2.27 (1.8–2.87)	2.03 (1.58–2.6)
Depressive disorder (UMLS C0011581)	2556 (50.13)	12 510 (31.17)	2.5 (2.11–2.94)	2.59 (2.19–3.07)
Anxiety, dissociative and somatoform disorders (ICD-9 300 ^a)	2252 (44.17)	9460 (23.57)	2.57 (2.42–2.72)	2.89 (2.71–3.08)
Drug abuse (UMLS C0013146)	2127 (41.71)	9635 (24.00)	2.28 (1.95–2.67)	2.11 (1.79–2.49)
Depressive disorder, not elsewhere classified (ICD-9 311 ^a)	2182 (42.79)	8617 (21.47)	2.74 (2.58–2.91)	3.17 (2.97–3.38)
Suicidal (UMLS C0438696)	1524 (29.89)	3111 (7.75)	1.41 (1.23–1.63)	1.32 (1.14–1.54)
Depressed mood (UMLS C0344315)	1427 (27.99)	4413 (10.99)	1.58 (1.22–2.05)	1.54 (1.18–2.02)
Top 10 Decreased-risk Features from best model in EXGB with all odds ratios and 95% CIs <1 ranked by feature importance				
One or more outpatient visits in 2 years	3619 (70.97)	35 127 (87.51)	0.35 (0.33–0.37)	0.44 (0.41–0.48)
Hypertensive disease (UMLS C0020538)	1707 (33.48)	20 328 (50.64)	0.6 (0.49–0.74)	0.76 (0.61–0.95)
Anger (UMLS C0002957)	656 (12.87)	1229 (3.06)	0.55 (0.33–0.92)	0.51 (0.29–0.88)
Hypersensitivity (UMLS C0020517)	3661 (71.80)	31 540 (78.58)	0.61 (0.57–0.66)	0.71 (0.66–0.77)
Neoplasms (UMLS C0027651)	655 (12.85)	11 508 (28.67)	0.63 (0.54–0.75)	0.72 (0.61–0.85)
Mecarazole (UMLS C0065839)	12 (0.24)	856 (2.13)	0.11 (0.06–0.2)	0.11 (0.06–0.19)
Effusion (UMLS C0013687)	382 (7.49)	5843 (14.56)	0.69 (0.59–0.8)	0.75 (0.64–0.87)
Diabetes mellitus (ICD-9 250 ^a)	547 (10.73)	9066 (22.59)	0.41 (0.38–0.45)	0.66 (0.6–0.73)
Encounter for antenatal screening of mother (ICD-9 V28 ^a)	109 (2.14)	2052 (5.11)	0.41 (0.34–0.49)	0.21 (0.17–0.26)
Anesthetics (NDF-RT CN200 ^e)	1473 (28.89)	19 963 (49.73)	0.41 (0.39–0.44)	0.53 (0.5–0.57)
SDOH from EXGB ranked by the unadjusted odds ratio				
Divorced state (UMLS C0086170)	404 (7.92)	1809 (4.51)	1.82 (1.63–2.04)	2.29 (2.03–2.57)
Marriage, life event (UMLS C0024841)	280 (5.49)	1857 (4.63)	0.17 (0.06–0.48)	0.19 (0.07–0.56)
Family support (UMLS C0150232)	157 (3.08)	578 (1.44)	0.34 (0.21–0.56)	0.29 (0.17–0.48)
Rehabilitation therapy (UMLS C0034991)	1071 (21.00)	6719 (16.74)	0.43 (0.26–0.71)	0.45 (0.26–0.77)

Note: Selected features were limited to a minimum of 1% prevalence in case or control group. The 95% confidence intervals for the selected features were limited to their range either all >1 or <1. The ranking was based on feature importance from the best performing extreme gradient boosting (XGB) model among the ensemble XGB (EXGB). The adjusted odds ratios were estimated while controlling for sex, age, race, and insurance (see Table 3). The SDOH were identified from EXGB ranked by the unadjusted odds ratio. The boldfaced number represents the highest odds ratio (OR) or adjusted OR (aOR) in the increased risk categories or the lowest OR/aOR in the decreased risk categories.

ICD-9: International Classification of Diseases, Ninth Revision; UMLS: Unified Medical Language System; NDF-RT: National Drug File—Reference Terminology.

^aAll 95% confidence intervals were measured through 2000 stratified bootstrap replicates.

^bWe used the Firth logistic regression method⁵² to calculate adjusted ORs.

^cExcluding demographic features that were listed in the top portion of the table.

^dUMLS Concept Unique Identifier (CUI).

^eNDF-RT code.

protective factors (*marriage, family support, and rehabilitation therapy*) (Table 4).

Sensitivity analysis results

We conducted a sensitivity analysis to address a potential limitation of ICD-9/10 coded suicide attempt, that is, cases and controls may have been mislabeled.⁴⁹ After applying NLP to the search of cases and controls with CUIs having a suicide keyword, we removed those cases and controls and re-evaluated EXGB performance in 30- and 730-day prediction windows within the test dataset. In the 30-day window, 218 (34.9%) cases and 568 controls (7.2%) were removed; in the 730-day window, 381 (25.5%) cases and 755 (6.3%) controls were removed. The sensitivity analysis showed no significant prediction performance difference compared to the original (primary)

results in both 30-day and 730-day windows (30-day AUCs, 0.920 vs. 0.932, $P = .17$; 730-day AUCs, 0.910 vs. 0.919, $P = .1$).

DISCUSSION

In this single-center (18 hospitals) retrospective case-control study, we leveraged NLP and ML technologies to identify patients at risk of first-time suicide attempts using historical EHR data. A large data-driven approach was employed to systematically collect and assess 54 600+ features from both structured and unstructured EHR data. Our EXGB model comprising 2126 features as a result of feature engineering demonstrated high accuracy of suicide attempt prediction across four prediction time windows—7, 30, 90, and 730 days—between an index visit, and the last clinical encounter prior to

the index visit (last point of clinical contact). Prediction of suicidal behavior within short windows of time is particularly salient to clinicians, given the paucity of existing tools to predict imminent suicidal risk.²⁰ We demonstrated that NLP of unstructured clinical narrative notes added significantly to the predictive power of ML applied solely to structural data. In addition, our EXGB model was robust across multiple demographic strata, last point of clinical contact, and depression diagnosis history. This suggests that our EXGB model could be deployed across various populations and clinical settings in the region, and may be potentially applied to similar populations in other healthcare systems. Additionally, the large number of features used by the EXGB model could potentially facilitate personalized intervention.

The strengths of this study include: (1) a focus on the prediction of first-time attempts, which account for two-thirds of suicide decedents;^{4–6} (2) demonstration of the added value of NLP into a prediction of suicidal behavior through an integration of NLP and ML; (3) development of a large data-driven approach with four ML models assessing 54 600+ features from EHR data; (4) evaluation of the robustness (bias) of a model to various demographic and clinical stratifications.

This study chose index visits among emergency department and inpatient facilities where practitioners may not be skilled at an assessment of suicidal risk, and where patients may be at increased suicidal risk.^{10,11,28} Although our models did not include index visits from outpatient facilities, the prior 2 years of EHR data included outpatient, emergency department, and inpatient visits. Most importantly, we showed consistent performance across the three types of last point of clinical contact prior to suicide attempt (i.e., emergency department, inpatient, and outpatient), suggesting that our model may have utility in different clinical settings.

This study has limitations. First, patients were included from a single, large regional healthcare system in the Northeastern United States and thus may not generalize to other healthcare settings. Moreover, patients may have sought healthcare in other systems, thereby underestimating the suicide attempts in study patients. Second, this is a retrospective case-control study with cases and controls selected from emergency department and inpatient index visits. This may have underestimated the contribution of some risk factors, since patients seen in these settings are at increased risk for a suicide attempt. Third, the study did not analyze EHR utilization shifts over time. To the extent that less complete data could occur in earlier years, the results of this effect would be to underestimate the performance of our model in more recent years when more complete EHR data became available. Last, we did not compare our models against conventional suicide prediction models. Besides technical limitations to estimating other predictive scores retrospectively, our main focus was to highlight the importance of unlocking the information wealth of available unstructured data to improve predictions and inform clinicians about SDOH. Moreover, our findings of AUC's > 0.90, along with relative high sensitivity and specificity, contrast favorably with the performance of some commonly used scales in the literature. For example, in one large multi-site study, four commonly used scales were tested in emergency department settings to predict a suicide attempt within six months. The AUCs ranged from 0.49 to 0.71, and scales either had high sensitivity and low specificity, or vice versa.⁵¹

It is well-known that documentation of suicide risk in specific settings, and of medical charting in general is neither complete nor highly accurate,⁴⁹ and consistent with previous reports, using NLP found that coded diagnoses underestimated the rate of suicide

attempts.⁵⁰ Any biases or disparities in the receipt of healthcare could be reflected in our algorithm. The algorithm is static and does not consider session-to-session changes in clinical status. While EXGB was the most accurate of the four ML models, its results are not transparent, and a slightly less accurate but more readily understood algorithm, such as one based on LASSO regression, may be more clinically useful.⁸ The left censoring of data prior to 2005 may have led to mislabeling of cases and controls from historical ICD codes. Thus, our data might include some cases or controls with previous suicide attempts. We leveraged the use of narrative notes and NLP to identify such occurrences and found no significant difference in prediction performance in our sensitivity analyses, which bolsters confidence that our findings are robust with or without undetected cases with a previous suicide attempt.

This study demonstrates that NLP of unstructured data added predictive power to the ML models (S-EXGB and S-LASSO) based solely on structured data. With the NLP of unstructured data, we identified several SDOH that affected suicide risk prediction including marital conflict (a risk factor) and family support (a protective factor). These SDOH could potentially furnish the clinician with relevant intervention targets. NLP also identified suicidal behavior that was not documented in the structural diagnosis codes. Thus, while the ML models without NLP of unstructured data performed well (e.g., AUC = 0.9 in a 30-day window), NLP may provide a better utility on further identifying clinically actionable findings that might be missed by the ML models solely using structured data.

Future work based on our study will require the following:

1. replication in other healthcare systems with diversity of populations and organizational structure using a prospective design;
2. extending this work to develop algorithms that are sensitive to changes in suicidal risk;
3. developing a continuous learning system that can periodically update models based on the most recent data;
4. exploring how to present risk data and potential treatment options suggested by the individual's risk level to the clinician and how the clinician should present these results to patients;
5. working with patients, healthcare providers, insurance providers, and regulatory agencies to find the effective, equitable, and ethical ways to apply these algorithms in healthcare settings.

CONCLUSIONS

Our ML and NLP framework using both structured and unstructured EHR data demonstrated accurate and robust first-time suicide attempt prediction. Using recently developed NLP analyses of unstructured textual data in EHRs provided a significant boost to the overall accuracy of these ML models. Thus, our large data-driven approach may be of value for healthcare systems to better identify patients with first-time suicide attempt risk and be used to provide timely interventions, especially in nonmental healthcare facilities, which is the most common point of contact for the majority of patients at high suicidal risk. Moreover, this approach may enable personalized intervention given the enormous amount of information identified from individual patients, particularly from unstructured clinical records.

FUNDING

This research was supported by the Beckwith Institute and the National Institute of Mental Health (P50-MH115838).

AUTHOR CONTRIBUTIONS

Study concept and design: FRT, NDR, CB, and DAB. Analysis and interpretation of data: FRT, LS, VR, NDR, SI, and DAB. Collection or assembly of data: FRT, LS, and NDR. Drafting of the manuscript: FRT, NDR, CB, SI, and DAB. Critical revision of the manuscript for important intellectual content: FRT, LS, VR, NDR, CB, SI, CGW, and DAB. Funding: DAB. Study Supervision and Coordination: FRT, NDR, and DAB.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

ACKNOWLEDGMENTS

We would like to gratefully acknowledge Nadine Melhem from the Department of Psychiatry, and Hoah-Der (Howard) Su from the Department of Biomedical Informatics at the University of Pittsburgh for their contributions in this article. We would also like to thank Lezhou (Joe) Wu and Allan Simpaio at the Children's Hospital of Philadelphia for their feedback.

DATA AVAILABILITY

The data underlying this article were provided by the University of Pittsburgh Medical Center via the Institutional Review Board at the University of Pittsburgh (askirb@pitt.edu). Data access may be granted on a case-by-case basis to researchers who meet necessary criteria.

CONFLICT OF INTEREST STATEMENT

DB receives royalties from UpToDate, Guilford Press, and eRT; research support from NIMH, AFSP, and Once Upon A Time Foundation; consultation fees from Healthwise; on scientific boards of AFSP and the Klingenstein Third Generation Foundation. NR has NIMH funding and received an honorarium from Axsome Therapeutics for membership in a scientific advisory committee. FRT has NIMH funding. None declared for others.

REFERENCES

- Mack KA, Clapperton AJ, Macpherson A, *et al.* Trends in the leading causes of injury mortality, Australia, Canada and the United States, 2000–2014. *Can J Public Health* 2017; 108 (2): e185–e191.
- Case A, Deaton A. Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proc Natl Acad Sci USA* 2015; 112 (49): 15078–83.
- Appleby L, Hunt IM, Kapur N. New policy and evidence on suicide prevention. *Lancet Psychiatry* 2017; 4 (9): 658–60.
- Ross V, Kölves K, De Leo D. Beyond psychopathology: a case-control psychological autopsy study of young adult males. *Int J Soc Psychiatry* 2017; 63 (2): 151–60.
- Kodaka M, Matsumoto T, Yamauchi T, Takai M, Shirakawa N, Takeshima T. Female suicides: Psychosocial and psychiatric characteristics identified by a psychological autopsy study in Japan. *Psychiatry Clin Neurosci* 2017; 71 (4): 271–9.
- Nock MK, Dempsey CL, Aliaga PA, *et al.* Psychological autopsy study comparing suicide decedents, suicide ideators, and propensity score matched controls: results from the study to assess risk and resilience in service members (Army STARRS). *Psychol Med* 2017; 47 (15): 2663–74.
- Belsher BE, Smolenski DJ, Pruitt LD, *et al.* Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry* 2019; 76 (6): 642–51.
- Simon GE, Johnson E, Lawrence JM, *et al.* Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Am J Psychiatry* 2018; 175 (10): 951–60.
- Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *J Child Psychol Psychiatry* 2018; 59 (12): 1261–70.
- Ahmedani BK, Simon GE, Stewart C, *et al.* Health care contacts in the year before suicide death. *J Gen Intern Med* 2014; 29 (6): 870–7.
- Schaffer A, Sinyor M, Kurdyak P, *et al.* Population-based analysis of health care contacts among suicide decedents: identifying opportunities for more targeted suicide prevention strategies. *World Psychiatry* 2016; 15 (2): 135–45.
- Barak-Corren Y, Castro VM, Javitt S, *et al.* Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry* 2017; 174 (2): 154–62.
- Adkins DE. Machine learning and electronic health records: a paradigm shift. *Am J Psychiatry* 2017; 174 (2): 93–4.
- 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 Psychiatry Res* 2018; 26 (3): 1–14.
- López Pineda A, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Rich Tsui F. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *J Biomed Inform* 2015; 58: 60–9.
- Walsh CG, Ribeiro JD, Franklin JC. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clin Psychol Sci* 2017; 5 (3): 457–69.
- 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.
- Kho AN, Rasmussen LV, Connolly JJ, *et al.* Practical challenges in integrating genomic data into the electronic health record. *BMJ* 2013; 15 (10): 772–8.
- McCoy TH, Castro VM, Roberson AM, Snapper LA, Perlis RH. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry* 2016; 73 (10): 1064–71.
- Glenn CR, Nock MK. Improving the short-term prediction of suicidal behavior. *Am J Prev Med* 2014; 47 (3): S176–S180.
- Posada JD, Barda AJ, Shi L, *et al.* Predictive modeling for classification of positive valence system symptom severity from initial psychiatric evaluation records. *J Biomed Inform* 2017; 75: S94–104.
- World Health Organization. Social determinants of health. Accessed January 8, 2021. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1
- Health.gov. Social determinants of health. Accessed January 8, 2021. <https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health>
- Halfon N, Larson K, Russ S. Why social determinants? *HCQ* 2010; 14 (sp1): 8–20.
- Hedegaard H, Schoenbaum M, Claassen C, Crosby A, Holland K, Proescholdbell S. Issues in developing a surveillance case definition for nonfatal suicide attempt and intentional self-harm using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD–10–CM) Coded Data. *Natl Health Stat Report* 2018; 108: 1–19.
- Bachmann S. Epidemiology of suicide and the psychiatric perspective. *Int J Environ Res Public Health* 2018; 15 (7): 1425–3.
- Ahmedani BK, Westphal J, Autio K, *et al.* Variation in patterns of health care before suicide: a population case-control study. *Prev Med (Baltim)* 2019; 127: 105796.
- Knesper DJ. *Continuity of Care for Suicide Prevention and Research: Suicide Attempts and Suicide Deaths Subsequent to Discharge from the Emergency Department or Psychiatry Inpatient Unit.* Newton, MA: Education Development Center, Inc; 2010.
- Lopez PA, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Rich T. F. Comparison of machine learning classifiers for influenza detection from

- emergency department free-text reports. *J Biomed Inform* 2015;58: 60–69.
30. Apache Software Foundation. cTAKES. <https://ctakes.apache.org>
 31. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
 32. Posada JD, Barda AJ, Shi L, *et al*.. Predictive modeling for classification of positive valence system symptom severity from initial psychiatric evaluation records. *J Biomed Inform* 2017; 75: S94–S104.
 33. De Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011; 18 (5): 557–62.
 34. Yan Z, Lacson R, Ip I, *et al*. Evaluating Terminologies to Enable Imaging-Related Decision Rule Sharing. *AMIA. Annu Symp Proceedings AMIA Symp*. Published online 2016.
 - 35..nlm. UMLS Terminology Services. <https://www.nlm.nih.gov/research/umls/index.html>
 36. Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Med Inform Decis Mak* 2018; 18 (74): 13–9. 10.1186/s12911-018-0654-2
 37. Chen T, Guestrin C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. ACM Press; 2016: 785–94. doi.org/10.1145/2939672.2939785
 38. Tsui F, Ye Y, Ruiz V, Cooper GF, Wagner MM. Automated influenza case detection for public health surveillance and clinical diagnosis using dynamic influenza prevalence method. *J Public Health (Oxf)*. 2017;40 (4): 1–8.
 39. Kent JT. Information gain and a general measure of correlation. *Biometrika* 1983; 70 (1): 163–73.
 40. Speybroeck N. Classification and regression trees. *Int J Public Health* 2012; 57 (1): 243–6.
 41. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning*. 2nd Ed. New York: Springer; 2009.
 42. DeLong ER, DeLong DM, Clark-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristics curves: a non-parametric approach. *Biometrics* 1988; 44 (3): 837–45.
 43. Breiman L, Friedman J, Stone C, Olshen R. *Classification and Regression Trees (Wadsworth Statistics/Probability)*. New York, NY: CRC Press; 1984.
 44. Breiman L. *Mach Learn* 2001; 45 (1): 5–32.
 45. Pearce N. What does the odds ratio estimate in a case-control study? *Int J Epidemiol* 1993; 22 (6): 1189–92.
 46. Tchetgen Tchetgen EJ. On a closed-form doubly robust estimator of the adjusted odds ratio for a binary exposure. *Am J Epidemiol* 2013; 177 (11): 1314–16.
 47. Angwin J, Larson J, Mattu S, Kirchner L. *Machine Bias*. ProPublica. 2016. <http://dx.doi.org/10.1108/17506200710779521>
 48. Dastin J. *Amazon Scraps Secret Ai Recruiting Tool That Showed Bias Against Women*. Reuters. 2018. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
 49. Stanley B, Currier GW, Chesin M, *et al*. suicidal behavior and non-suicidal self-injury in emergency departments underestimated by administrative claims data. *Crisis* 2018; 39 (5): 318–25.
 50. Zhong QY, Mittal LP, Nathan MD, *et al*. Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. *Eur J Epidemiol* 2019; 34 (2): 153–62.
 51. Steeg S, Quinlivan L, Nowland R, *et al*. Accuracy of risk scales for predicting repeat self-harm and suicide: a multicentre, population-level cohort study using routine clinical data. *BMC Psychiatry* 2018; 18 (1): 113.
 52. Wang X. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Front Genet*. 2014; 5: 187.