npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01266-7

Deep learning for identifying personal and family history of suicidal thoughts and behaviors from EHRs

Prakash Adekkanattu[1], Al'ona Furmanchuk[2], Yonghui Wu[3], Aman Pathak[3], Braja Gopal Patra[1], Sarah Bost[3], Destinee Morrow[4], Grace Hsin-Min Wang[3], Yuyang Yang[2], Noah James Forrest[2], Yuan Luo[2], Theresa L. Walunas[2], Weihsuan Lo-Ciganic[3,5], Walid Gelad[5], Jiang Bian[3], Yuhua Bao[1], Mark Weiner[1], David Oslin[6] & Jyotishman Pathak[1]

Personal and family history of suicidal thoughts and behaviors (PSH and FSH, respectively) are significant risk factors associated with suicides. Research is limited in automatic identification of such data from clinical notes in Electronic Health Records. This study developed deep learning (DL) tools utilizing transformer models (Bio_ClinicalBERT and GatorTron) to detect PSH and FSH in clinical notes derived from three academic medical centers, and compared their performance with a rule-based natural language processing tool. For detecting PSH, the rule-based approach obtained an F1-score of 0.75 ± 0.07, while the Bio_ClinicalBERT and GatorTron DL tools scored 0.83 ± 0.09 and 0.84 ± 0.07, respectively. For detecting FSH, the rule-based approach achieved an F1-score of 0.69 ± 0.11, compared to 0.89 ± 0.10 for Bio_ClinicalBERT and 0.92 ± 0.07 for GatorTron. Across sites, the DL tools identified more than 80% of patients at elevated risk for suicide who remain undiagnosed and untreated.

Suicide represents a critical public health challenge globally, ranking as the second leading cause of death among people between 15 and 29 years old[1]. In the United States, it is the tenth leading cause of death overall and ranked second for those aged 10 to 34, as reported by the Centers for Disease Control and Prevention[2]. In 2020, suicide rate among U.S. veterans—a high-risk population—was 1.57 to 1.66 times greater than in the non-veteran adult population, when adjusted for age and sex differences[3]. However, patients at risk of suicide often remain underdiagnosed for a variety of reasons such as stigma associated with mental health issues and the limitations of diagnostic codes in capturing the complexity of mental health conditions[4]. Recognizing warning signs and addressing risk factors with effective early interventions are crucial in suicide prevention and mental health promotion[5–8]. Personal and family history of suicidal thoughts and behaviors (PSH and FSH, respectively) significantly increase future suicide risk, emphasizing the importance of identifying these factors[9,10].

Based on the existing literature, suicidal thoughts and behaviors (STBs) broadly encompasses thoughts, behaviors, actions, and emotions linked to suicide and self-harm[11]. STBs manifests in various forms, including suicidal ideation (SI), suicidal behavior (SB), and suicide attempts (SAs), each carrying distinct implications for individuals. SI entails a range of thoughts and preoccupations with death and suicide, varying in intensity from fleeting thoughts to detailed plans[12]. Although its definition may vary and often evolves, SB generally refers to any self-directed actions with potential lethality, ranging from preparatory acts to completed suicides[11]. Thus, it is critical to develop effective methods to identify individuals with STBs that often are underdiagnosed to improve the management of suicide prevention and treatment.

Understanding, predicting, and preventing STBs is a complex and challenging task. Over the years significant efforts have been made in developing and testing STB risk assessment tools in a wide variety of population[13]. Existing tools have inadequate reliability and low positive predictive value (PPV) in distinguishing between low and high-risk patients[14]. Moreover, majority of these tools offer STB risks assessment over longer time windows (e.g., 6- or 12 months) which is of limited clinical utility for suicide prevention interventions[13]. This requires newer approaches, such as electronic health record–derived algorithms, Internet-based screening methods, and ketamine's potential benefit for preventing attempts, according to Mann et al.[15]. More recently, with the availability of patient Electronic Health Records (EHRs) and advances in computational methods such as natural language processing (NLP), machine learning

[1]Weill Cornell Medicine, New York, NY, USA. [2]Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [3]University of Florida College of Medicine, Gainesville, FL, USA. [4]Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [5]University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [6]Corporal Michael J Crescenz Veterans Affairs Medical Center, Philadelphia, PA, USA. ✉e-mail: pra2008@med.cornell.edu

(ML), and deep learning (DL), there are opportunities to develop automated tools to assess and intervene STB risks in real-time. Identifying risk factors in real time should be a critical component of any automated risk assessment tool. The most prominent example of such a suicide risk prediction and prevention effort is the REACH VET (Recovery Engagement and Coordination for Health—Veterans Enhanced Treatment) program from the Veterans Health Administration[16,17]. REACH VET program, already in practice since 2017, uses a machine learning model to identify patients at high risk of suicide. Inputs to the model are 61 risk variables derived from the veterans' health records. Each patient received a risk score, and those who fell in the top 0.1% risk stratum were identified to receive intervention. The risk model, however, used only structured data available in EHRs and reported a C statistics that varied from 0.836 (all health systems) to 0.634 (psychiatric inpatient settings). Clinical notes in EHR are one other critical information available to clinicians during encounter with patients and can significantly improve the predictive value of any risk assessment tools[18].

EHRs have been widely used to study phenotyping and risk prediction models[19]. Notably, the majority of suicide decedents had a physician visit in the year prior to their death and 45% had a visit in the last month of life, highlighting the potential of EHR data in suicide risk prediction and prevention[20]. While International Classification of Disease (ICD-9/10) diagnosis codes in EHRs are used to document STBs, their sole use has proven insufficient[21]. This inadequacy is partly because details like personal and family history of STBs are often only recorded in clinical notes instead of ICD-9/10 diagnoses codes. The ICD-9 coding system had no defined diagnosis codes to document the historical aspect of STBs. Personal history of self-harm was introduced only in 2016 when ICD-10 coding system was formally adopted across the US. For family suicide history, ICD-10 introduced "z81.8" as the code for self-harm in the family. However, FSH is often considered as a risk factor and many physicians may be reluctant to use this code for diagnosis purpose. NLP and ML/DL techniques show promise in effectively mining clinical notes for information that are not readily available in structured data. Most recently, transformer-based architectures have facilitated building high-performance models and fine-tuning has made it possible to effectively utilize these models for a wide variety of tasks. Since these language models (e.g., BERT, ChatGPT, LLaMA, FLAN) have been trained using very large datasets, they possess contextual knowledge, and fine-tuning them with problem-specific data can achieve substantial improvement in performance[22]. In particular, the BERT (Bidirectional Encoder Representations from Transformers) models effectively capture the interaction between a key clinical concept and their surrounding context, and have shown superior performance in various NLP tasks including text classification[23].

Advancements in automated detection of STBs from clinical notes have shown varied success[24–33]. These approaches, ranging from rule-based algorithms to ML techniques, primarily focus on identifying suicidal tendencies without distinguishing between current and historical events. Fernandes et al. developed an NLP method to detect SI and SA using a psychiatric clinical research database, but no distinction is made between current or historic events[29]. Similarly, Carson et al. developed an ML method to identify SB among psychiatrically hospitalized adolescents, but the classifier was trained to detect SB without any distinction on whether these events are current or historic[25]. Previously, we also reported a weak supervision NLP method for the detection of current SI[28]. To date, there have been no studies reported specifically on personal and family histories of STBs, despite the critical but different insight that such information may offer. Our study addresses this gap by developing and evaluating novel NLP and DL-based tools to detect both PSH and FSH from clinical narratives, tested across diverse patient cohorts in three academic medical centers in the U.S.

Table 1 | Demographics characteristics and ICD diagnosis of patients for which the gold standard corpora were developed and used for evaluating the performance of the NLP and DL methods at the three sites

	WCM	NM	UF
Total patients (n)	134	313	341
Total notes (N)	301	400	400
Avg age (SD) on date of note documented	48.43(18.2)	44(14.5)	52.18(15.1)
Age categories			
<18	2(1.5%)	1(0.3%)	4(1.2%)
18–39	47(35.1%)	127(40.6%)	66(19.4%)
40–59	41(30.6%)	129(41.2%)	157(46%)
≥60	44(32.8%)	56(17.9%)	114(33.4%)
Sex			
Female	81(60.4%)	198(63.3%)	231(67.7%)
Male	53(39.6%)	115(36.7%)	110(32.3%)
Race			
Native/Alaskan	0(0.0%)	2(0.6%)	0(0%)
Asian	4(3.0%)	5(1.6%)	2(0.6%)
Black/AA	10(7.5%)	44(14.1%)	124(36.4%)
Hawaiian/PI	0(0.0%)	1(0.3%)	0(0%)
White	86(64.2%)	229(73.2%)	203(59.5%)
Other/Declined /Unknown	34(25.4%)	32(10.2%)	12(3.5%)
Ethnicity			
Hispanic	12(9.0%)	44(14.1%)	10(2.9%)
Non-Hispanic	120(89.6%)	255(81.5%)	328(96.2%)
Other/Declined/Unknow	1(0.7%)	32(10.2%)	3(0.9%)
Marital Status			
Single	83(61.9%)	167(53.4%)	NA
Married	30(22.4%)	99(31.6%)	NA
Divorced	13(9.7%)	NA	NA
Widowed	7(5.2%)	NA	NA
Other	1(0.7%)	47(15%)	NA
ICD diagnosis of SI on analyzed encounter			
Yes	1(0.7%)	9(2.9%)	5(1.5%)
No	133(99.3%)	304(97.1%)	336(98.5%)
ICD diagnosis of SB on analyzed encounter			
Yes	0(0.0%)	6(1.9%)	2(0.6%)
No	134(100.0%)	307(98.1%)	339(99.4%)
ICD diagnosis of SI before analyzed encounter			
Yes	1(0.7%)	31(9.9%)	27(7.9%)
No	133(99.3%)	282(90.1%)	314(92.1%)
ICD diagnosis of SB before analyzed encounter			
Yes	1(0.7%)	23(7.3%)	17(5.0%)
No	133(99.3%)	290(92.7%)	324(95.0%)
Provider specialty			
Psychiatric	31(23.1%)	669(21.1%)	2(0.6%)
Non-Psychiatric	103(76.9%)	247 (78.9%)	339(99.4%)

Fields where data is not available are indicated by "NA".

Results

The patient cohorts are predominantly women across all three study sites: 60.4% at WCM, 63.3% at NM, and 67.7% at UF. The study involved examining 301 notes from 134 patients at WCM, 400 notes from 313 patients at NM, and 400 notes from 341 patients at UF. Notably, demographic variations (Table 1) were evident across the sites, particularly in age, race, and ethnicity. Both WCM (32.8%) and UF (33.4%) had a higher representation of older patients (aged 60 or older) compared to NM (17.9%). The UF cohort (36.4%) had a higher representation of Black patients when compared to WCM (7.5%) and NM (14.1%) cohorts. Additionally, a larger proportion of non-Hispanic patients were observed at UF when compared to WCM and NM sites (96.2% vs 89.6% and 81.5%). Across all sites, more than 90% of the patients had no recorded diagnosis of STBs (either pre-existing or concurrent with the note date), according to the ICD-9/10 codes in their medical records.

The rule-based NLP tool demonstrated varying performance across different sites (Table 2). Specifically, the macro-average F1-scores ranged from 0.81 to 0.63 for PSH and 0.80 to 0.58 for FSH when evaluated against the corresponding gold standards. Notably, the highest performance for both outcomes was recorded at WCM. The lowest performance for PSH was observed at UF, primarily as a result of a reduced recall of 0.60. Similarly, for FSH, the lowest performance occurred at NM, mainly because of a low recall of 0.44.

The performance of the DL-based tools is shown in Table 3. For PSH, utilizing Bio_ClinicalBERT-based model yielded F1-score of 0.88, 0.73, and 0.88 at WCM, NM and UF, respectively. In contrast, the GatorTron-based model demonstrated superior performance for PSH with F1-score of 0.92, 0.78, and 0.83 at WCM, NM and UF, respectively.

For FSH, both methods showed comparable effectiveness: Bio_ClinicalBERT-based DL achieved F1-score of 0.88, 0.81, and 1.00 at WCM, NM and UF, respectively, while GatorTron-based DL reported slightly higher scores of 0.88, 0.90, 1.00 at WCM, NM and UF, respectively.

In the WCM cohort of 134 patients, while manual annotation identified 45 (33.6%) patients with a PSH and 14 (10.4%) patients with a FSH, only 1 (2.2%) PSH patient and 1 (7.1%) FSH patient had relevant ICD codes on or before the notes report date. ICD-9 code 950.3 (Suicide and self-inflicted poisoning by tranquilizers and other psychotropic agents) was observed for one patient who was positively identified as having a family history of suicide by the NLP classifier. Conversely, the DL-based GatorTron tool accurately detected 42 (93.3%) of the PSH cases and 12 (85.7%) of the FSH cases from patient notes. No ICD code exists for FSH, precluding direct comparison of underdiagnosis in EHRs. In the NM cohort of 313 patients, while manual annotation identified 97 (31.0%) patients with PSH and 57 (18.2%) with FSH, only 9 (9.3%) PSH and 6 (10.5%) FSH patients had ICD codes for STB on or before the notes report dates. GatorTron, however, successfully identified 78 (80.4%) of the PSH cases and 51 (89.5%) of the FSH cases from notes. In the UF cohort of 341 patients, manual annotation identified 64 patients with PSH and 12 with FSH. Of these, only 5 (7.8%) PSH and 0 FSH patients had any ICD code for SI or SB on or before the notes report date. GatorTron, successfully identified 57 (89.0%) of the PSH cases and 12 (100%) of the FSH cases.

Table 2 | Performance of the rule-based NLP tools at the three sites

	WCM (*n* = 301)	NM (*n* = 400)	UF (*n* = 400)
PSH			
Accuracy	0.82	0.79	0.84
Precision (macro average)	0.81	0.67	0.77
Recall (macro average)	0.83	0.69	0.60
F1-score (macro average)	0.81	0.68	0.63
FSH			
Accuracy	0.94	0.91	0.98
Precision (macro average)	0.79	0.85	0.99
Recall (macro average)	0.80	0.44	0.62
F1-score (macro average)	0.80	0.58	0.69

Discussion

Individuals with a PSH or a FSH have an increased risk for future suicide events. However, none of the existing methods reported for the detection of STB specifically looked at the historical nature of these events. We developed rule-based NLP and DL tools to detect PSH and FSH from clinical notes and compared results based on clinical diagnoses. The algorithms were validated by comparison to manually annotated clinical notes of patients with different characteristics from three different academic medical centers in the U.S. For both PSH and FSH, the DL methods showed higher performance than the rule-based NLP method across all three sites. Given that prior efforts to detect STBs from clinical notes did not focus specifically on the historical aspect of suicidality, a direct comparison of the performance of our current tools with existing literature is not feasible. Fernandes et al. developed an NLP method to detect SI and SA using a psychiatric clinical research database and reports a sensitivity of 0.88 and a precision of 0.92 for SI and sensitivity of 0.98 and a precision of 0.83 for SA[29]. The reported study implemented both a rule-based NLP and ML methods to identify SI and SA. Also, using a weak supervision NLP method, efforts from our group recently reported an F1-score of 0.82 for the current SI[28]. The rule-based NLP implemented in this study excluded all historical mentions of SI. Similarly, Carson et al. developed an ML method using terms extracted from clinical notes to identify SB among psychiatrically hospitalized adolescents and reported a sensitivity of 0.83, specificity of 0.22, and AUC of 0.68[25]. The study analyzed clinical notes for suicide attempt from 73 adolescent patients who were selected from a self-reported survey of having suicide attempt in the past year prior to an index hospital admission. Relevant terms were extracted from clinical notes through NLP, mapped to corresponding UMLS CUI's, and used in training and testing multiple ML-based classifiers for suicidal attempt. While the current rule-based tool demonstrates good performance at the development site (WCM), its performance at external sites (NM and UF) was lower, suggesting room for further improvement.

Table 3 | Performance of the Transformer based DL tools at the three sites

	WCM (*n* = 301)		NM (*n* = 400)		UF (*n* = 400)	
	Bio_ClinicalBERT	GatorTron	Bio_ClinicalBERT	GatorTron	Bio_ClinicalBERT	GatorTron
PSH						
Accuracy	0.89	0.93	0.74	0.79	0.91	0.87
Precision	0.90	0.93	0.75	0.81	0.86	0.80
Recall	0.89	0.92	0.74	0.79	0.91	0.87
F1-score	0.88	0.92	0.73	0.78	0.88	0.83
FSH						
Accuracy	0.90	0.91	0.84	0.91	1.00	1.00
Precision	0.88	0.85	0.79	0.89	1.00	1.00
Recall	0.90	0.91	0.84	0.91	1.00	1.00
F1-score	0.88	0.88	0.81	0.90	1.00	1.00

The Transformer-based BERT model, on the other hand, has comparable performances at both the development sites and external sites making it a better choice for detecting history of STB. Moreover, the DL-based tools were fine-tuned on relatively small number of notes (gold standard corpus) from each site which reduces the overall development effort required. In contrast, the rule-based NLP tools relied on an iterative process of manually defining lexicon and implementing specific set of rules using a large development corpus.

Error analysis at the three sites further suggests that documentation practices vary widely when reporting historical aspects of STBs. For instance, at NM, notes with suicidal events documented using a timestamp accounted for the majority of false negative cases and contributed towards the low recall. This includes examples such as *"on xx/xx was hospitalized x wk for suicide attempt by overdose"*, *"Suicidal ideation xx/xx/xxxx"*, and *"suicide attempt xxxx"*. At both NM and UF, some false negative cases of PSH arise due to the fact that annotators interpreted history of STB from certain statements that do not conform to the concept-modifier rule we established within the NLP algorithm. Examples include mentions of a specific date in the past such as *"who presents to NNNNN in xx/xx for medication management for depression after visiting the ED for passive suicidal ideations."*, *"No longer suicidal"*, and *"Suicide attempts: His only suicide attempt was as a teen"*. False positives for PSH are mainly due to the inadequacies of the heuristics rule we applied when there are multiple instances of suicide history mentions in a document. Examples include *"Prior suicide attempts: none Prior self-mutilation: none Prior suicidal ideation: current Prior violence: none"* and *"History of suicidal/homicidal ideation/behaviors: Pt denies"*. In the case of FSH, at both NM and UF, the (family) concept-modifier rule that we established in the NLP algorithm was found to be inadequate to capture some of the documented cases of FSH. Examples of false negatives include *"Family Psychiatric History: Mother - depression Great grandfather - suicide"* and *"she endorsed some psychotic symptoms that appeared to surface when the patient is under significant stress (suicide attempt by father, getting married, and presently)"*. Examples of false positives include *"Family history unknown: Yes Columbia Suicide Severity Rating Scale Columbia Suicide Severity Rating Scale (C-SSRS)"*, and *"Relatives or friends with suicidal ideation? yes, foster father committed suicide when she was xx"*.

The DL tools we developed demonstrated robustness to the varying documentation practices and patient populations presented across the three sites. We analyzed clinical notes from diverse healthcare settings at WCM, NM, and UF to understand documentation practices in different medical environments and clinical settings. WCM's cohort primarily came from general outpatient settings, focusing on mental health diseases and health maintenance, with detailed notes on personal and family health histories, including mental health and suicidality indicators. At NM, the data encompassed a wider range of settings, including ED, inpatient, and outpatient services. Notes from the ED and inpatient settings often capture more immediate and acute health concerns, potentially including emergent personal mental health crises like STBs. ED notes, while comprehensive, varied in family mental health details. Outpatient notes at NM, similar to WCM, would likely include comprehensive health histories of patients but with varying degrees of details regarding mental health, depending on the clinician's specialty and primary reason for the outpatient visit. At UF, the clinical notes were specifically collected from patients with at least 1 outpatient visit and prescribed opioids in an outpatient setting. In such a setting, provider notes are expected to be centered around oncological care, treatment plans, and follow-ups. However, these notes can also be rich sources of information regarding a patient's mental health, as dealing with a cancer diagnosis can significantly impact psychological well-being. The likelihood of encountering explicit mentions of personal or family history of STB might be higher in this context, given the profound emotional and psychological implications of cancer on patients and their families. The comparable performance of the DL methods at the three sites, despite the fact that notes were originated at different clinical settings, suggest the general suitability of this method in detecting history of STB from clinical notes.

Prior research has shown that major psychiatric illness, previous suicide attempts, ideations, and behaviors, as well as suicide among family members are primary risks for future suicide events[34,35]. Gathering individual, family, and community level suicide risk relevant data is therefore essential in developing effective suicide risk assessment tools. To be practically useful, such tools will be brief enough to be conducted in a primary care or psychiatric setting and will identify a threshold beyond which preventive action should be taken. Although physicians have access to clinical notes for each individual patient, manually going through tens of hundreds of notes is technically challenging, and information such as history of STBs may be overlooked. In recent years, there have been some efforts developing STB risk prediction tools using information extracted from clinical notes. In a case-control study on investigating the potential impact of including NLP extracted data in suicide risk prediction using a ML model, Levis et al. found that variables derived from clinical notes indeed improved the prediction performance[18]. Similarly, Adamou et al. implemented ML algorithms utilizing text-mining techniques to predict within a specified period which people are most at risk for suicide at the time of referral to a mental health service and found that including clinical variables extracted from notes significantly improved the risk prediction performance[36]. Bittar et al. found that including text features along with structured data from EHR in suicide risk prediction models significantly improves the prediction performance[37]. Ben-Ari et al. implemented a Random Forests method to predict STB over a 10 year period using data extracted from clinical records of 250,000 veterans and reported a ROC of 0.86[38]. Similarly, McCoy et al. implemented regression models for predicting suicide or accidental death after discharge from hospitals and found that incorporating variables extracted from narrative discharge notes improves stratification of risk for death by suicide[39]. None of these reported studies however extracted PSH or FSH as one of the clinical variables from notes, or included them as part of any risk prediction model. In this regard, the present study on extracting PSH and FSH from clinical notes has significant utility in developing and implementing more effective suicide risk prediction models.

Previous research indicates that mental health conditions were inadequately recorded as structured ICD or SNOMED codes in EHRs, but were more likely to be documented in patient notes[40]. Our prior study found that among patients identified through NLP-based approach as having PHQ-9 scores—a clinical instrument measuring depression severity—31% showed scores suggesting major depressive disorder, yet lacked a corresponding structured ICD or SNOMED diagnosis code[41]. The study used interface terminology items within Epic corresponding to ICD-9 codes starting with 311.*, 300.4*, 292.2*, or 292.3* and ICD-10 codes F32.* and F34.1* for defining depression in general. For major depressive disorder in particular, the study used interface terminology items corresponding to ICD-9 codes 292.2 or 292.3 and ICD-10 codes F32.*. Sources of diagnosis codes in the EHR included the patient's problem list, medical history, encounter diagnoses, and billing diagnoses. The current study reveals a similar pattern of underdiagnosis when relying solely on ICD codes. The observed low ICD codes for STB could be interpreted as due to a multitude of factors. At both WCM and UF, our cohorts partly consist of patients who received care during the period when ICD-9 coding system were still being used for documentation of diagnosis, for which there was no specific codes for PSH or FSH. We should also point out that traditionally ICD codes were used in EHRs mainly for the purpose of billing. Clinicians not always report ICD codes for a given health condition, thus further underscores the importance of extracting information relevant to PSH and FSH from clinical notes. In this study we investigated under-diagnosis from the perspective of clinical documentation where we looked at the availability of ICD diagnosis code for PSH and FSH in the EHR dataset. We did not consider the wider question of when and under what conditions clinicians should diagnose STBs, which is beyond the scope of our current study.

Although a comprehensive discussion is beyond the scope of this study, under-diagnosis of STBs is also a consequence of stigma and discrimination associated with mental health illnesses and suicidality. Stigma and perceived discrimination among individuals with mental illnesses and suicidality are well documented[4,42]. When documenting in EHRs and communicating diagnostic information to patients and their families, physicians are often faced with conflicting bioethical considerations of what level of information is appropriate in a given situation. Physicians may be reluctant to give a diagnosis code due to concerns about its impact on individuals due to the perceived stigma and discrimination. In particular, self-stigma (occurs when negative stereotypes are internalized) leading to low self-esteem, shame and hopelessness might impede individual seeking help from mental health professionals.

The current study, while advancing the field of suicidality detection using NLP and DL tools through a multi-site approach, faces several limitations. Primarily, the lexicon for STB and history modifiers was restricted, leading to moderate performance at external sites due to missed keywords specific to those locations. This issue is particularly pronounced in self-harm behavior identification, where diverse operational definitions exist and behaviors like cutting and burning are not uniformly documented. Note that we excluded terms such as "cutting" and "burning" from our lexicon during development as these terms can occur in various other contexts outside of STBs. However, such terms can also occur in the context of life threatening self-harm behavior of patients and may have adversely contributed to the performance of the algorithm. Local customization of the lexicon could enhance algorithm accuracy. Secondly, the rule-based NLP and DL tools were tested across three academic medical institutions, which may not reflect the broader healthcare system. While the data from NM encompassed diverse clinical settings and specialties, broader testing across various contexts is essential to confirm the efficacy and adaptability of our tools. Thirdly, modern EHR systems use pre-defined templating component in organizing and documenting clinical notes. The extent to which these templates are included in notes varies widely across the three sites. While notes from WCM and UF are mostly in free-text format with little or no templating components, those from NM exhibit a mix of semi-structured and unstructured formats varying with the type of note. Our rule-based NLP algorithm did not account for possible templating structure and may have impacted negatively on the performance.

In conclusion, our study highlights the significant potential of rule-based NLP and DL tools in identifying personal and family histories of STB, which are often only recorded in free-text notes within various EHR systems. This approach marks an important step towards developing an effective risk prediction model in suicide prevention efforts. Our findings indicate that traditional methods of structured information collection might miss up to 90% of patients exhibiting suicidal thoughts or behaviors, as these are often only mentioned in unstructured clinical notes. For about 80% of these cases, our developed models prove effective in extracting this critical information, underscoring the value of these tools in enhancing mental health assessments. The DL-based approach achieves a higher and more robust performance across diverse health care systems compared to rule-based NLP. Future works include development and testing of practical STBs risk prediction models by including PSH, FSH and information such as social and behavioral determinants of heath from clinical notes as part of potential predictors. To this end our group has already implemented several ML models to predict the risk of STBs among patients prescribed opioids using structured data in EHR (unpublished work).

Methods

Study setting and data sources

The rule-based NLP and DL tools were developed and validated at three academic medical centers: Weill Cornell Medicine (WCM), Northwestern Medicine (NM), and the University of Florida Health (UF) to enhance the generalizability and portability of our tool. This multi-site study was conducted with the approval of The Weill Cornell Medicine Institutional Review Board (Protocol No. 22-05024878), The Northwestern University Institutional Review Board (Protocol No. STU00218389), and The University of Florida Institutional Review Board (Protocol No. IRB202001100), ensuring adherence to ethical standards and patient privacy regulations. Informed consent was waived as it is not practicable to obtain consent from large numbers of patients for a retrospective cohort study. In order to enhance the generalizability of the NLP and DL tools, we assembled clinical notes from diverse patient cohorts seeking care from outpatient ambulatory services, emergency department (ED), inpatient care from multiple specialties. For PSH, we implemented a rule-based NLP method and two DL methods using the Bio_ClinicalBERT and GatorTron Transformer models. Similarly, for FSH we implemented a rule-based NLP method and two DL methods using the Bio_ClinicalBERT and GatorTron Transformer models. In the gold standard corpora from each site, we gathered demographic data and recorded diagnoses of SI and SB for the respective patients.

The training data for this study was sourced from WCM, an academic medical center in New York City affiliated with NewYork-Presbyterian Hospital. The dataset comprised of more than 13.8 million clinical encounter notes derived from patients ($N = 177{,}993$) who were either prescribed antidepressants or diagnosed with mental health conditions between 2000-2020. Clinical notes consisted of progress notes (49.4%), telephone encounters (32.3%), patient instructions (2.1%), letters (2.0%), nursing notes (0.4%), and unknown types (13.8%). The notes, authored by clinicians from various specialties including internal medicine, psychiatry, anesthesiology, and pain medicine, offer a rich, unstructured collection of information, reflecting the diversity of clinical environments and the variability levels of detail provided. When queried the 13.8 million notes with a filter of having a character string "suicide", it resulted 194,204 notes, from which we randomly selected 1,301 notes. Of the 1,301 notes, 1000 notes were used for the development of the rule-based NLP method. The remaining 301 notes were used for the development of gold standard for evaluating the NLP method. The same gold standard was used for the training and testing of the DL tools.

NM is a comprehensive academic medical center located in Chicago, IL. The NM Enterprise Data Warehouse is an integrated data platform that provides secure, centralized access to clinical and ancillary data sources from all inpatient and outpatient settings. It consolidates data from Northwestern Memorial HealthCare, the Feinberg School of Medicine at Northwestern University, and Northwestern Medicine Regional Medical Group. The 400 notes used for the validation study were randomly collected from the integrated system between January-December, 2018. The gold standard corpus consisted of 3 (1%) assessment & plan notes, 23 (7.3%) consult notes, 4 (1.3%) discharge/summary notes, 21 (6.7%) ED notes, 26 (8.3%) History & Physical notes, 12 (3.8%) plan of care notes, 167 (53.4%) progress notes, 63 (20.1%) psychiatric note, 3 (1%) telephone encounters, and 16 (5.1%) notes of other types. The clinical notes used for validation were written by 66 (21.1%) psychiatric specialists and 247 (78.9%) non-psychiatric specialists.

The UF Health Institutional Data Repository (IDR) is a clinical data warehouse that aggregates data from the university's various clinical and administrative information systems, including the Epic (Epic Systems Corporation) system. At UF, the corpus was developed using a random sample of 400 clinical notes derived from a cohort of individuals with at least one prescription of opioids between 2016 and 2019 recorded in the IDR. Patients with pain conditions or those prescribed opioids are at an increased risk of STBs. Adult patients aged ≥18 who had at least one outpatient visit and at least one eligible opioid prescribing order (excluding injectable and buprenorphine approved for opioid use disorder) were included in the patient sample. The gold standard corpus consisted of 13 (3.3%) consult notes, 6 (1.5%) discharge summary, 11 (2.8%) ED notes, 10 (2.5%) H&P notes, 319 (79.8%) progress notes, 8 (2.1%) psychiatric inpatient notes, and 33 (8.4%) other or unknown types.

Evaluation of the NLP and DL tools

The performance of the rule-based NLP and DL tools was evaluated using a gold standard corpus developed by manual annotation using an annotation tool. We used the manual in-file annotation (NM) and Brat annotation tool (WCM, UF) to identify relevant concepts in notes collected from the corresponding EHR systems[43]. At each site, we set up a secure instance of the annotation tool with the same annotation scheme. All annotators were given the same annotation guidelines. These guidelines (Supplementary Note 1)

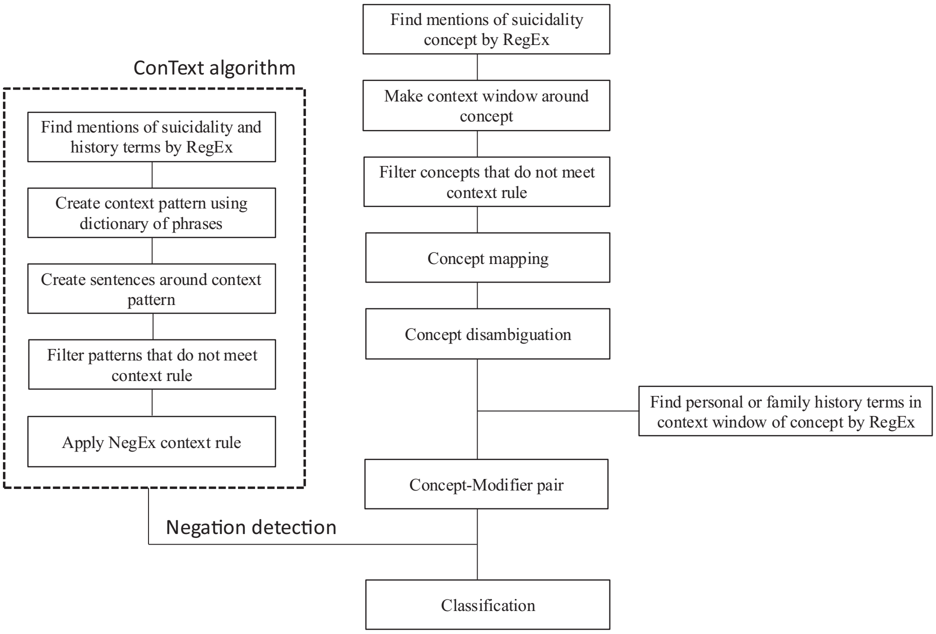
ConText algorithm
Find mentions of suicidality and history terms by RegEx
Create context pattern using dictionary of phrases
Create sentences around context pattern
Filter patterns that do not meet context rule
Apply NegEx context rule
Negation detection
Find mentions of suicidality concept by RegEx
Make context window around concept
Filter concepts that do not meet context rule
Concept mapping
Concept disambiguation
Find personal or family history terms in context window of concept by RegEx
Concept-Modifier pair
Classification

Fig. 1 | Rule-based NLP algorithm implemented for detecting PSH and FSH in clinical notes. Rule-based algorithm implemented using the Leo NLP framework for the detection of PSH and FSH in clinical notes. For classification, the pipeline implemented a series of rules and logic in concept detection, context window selection, concept mapping, disambiguation, filtering and negation detection.

include finding all mentions of personal history of STBs and family history of STBs within the note. All instances of PSH and FSH along with their negation status were extracted. The label was assigned based on majority votes across negation statuses. In case of a tie, the negation status of the last instance is used. Using the above heuristics, for each note the annotator made a binary classification of positive or negative for PSH and positive or negative for FSH. After both annotators completed annotations on the documents, any disagreements between the two sets were resolved through joint sessions of annotations, giving us a final gold standard annotated document set.

At WCM, we developed the gold standard through a manual review of 301 encounter notes selected at random from a superset of notes containing the character string "suicide". These notes were not part of the 1000 notes previously selected for the development and testing. To establish the reference standard, two reviewers (PA, BGP) annotated all notes based on the previously defined guidelines. In instances of differing assessments, reviewers resolved discrepancies through joint sessions. The reviewers also confirmed all notes that did not have any mention of suicide history information. At NM, manual annotation of 400 notes was performed by three reviewers (YY, NF, AF). Four labels were created to identify the presence or absence of a personal and family history of STB at each note. Each note was annotated independently by two reviewers. If the two reviewers differed in their assessment (Supplementary Table 2) the discrepancy was resolved through a joint session of all three reviewers. The resolved labels in question were utilized for comparing with the labels generated by NLP and DL tools. At UF, a total of 400 notes were identified for manual annotations. Three annotators (AP, ML, SZ) classified each note for the presence or absence of STBs. Discrepancies among the annotators were resolved through discussions to reach a consensus.

During the validation of annotation, we observed generally a high inter-rater agreement between the manual annotators at all three sites. At WCM, the Cohen's Kappa measured was 0.89 for PSH and 0.89 for FSH. Before the final voting process among all three reviewers at NM, the Cohen's Kappa coefficient were 0.75 and 0.85 for PSH and FSH, respectively. At UF, Cohen's Kappa score of 0.89 was observed for PSH and FSH combined, ensuring sufficient agreement between reviewers.

Rule-based NLP tool development

To develop a novel rule-based tool, we employed the Leo NLP framework, an open-source tool provided by the U.S. Veterans Administration[44]. Specifically, we created a dedicated instance of Leo, termed *SuicideHistoryExtractor*, to detect historical instances of STB from clinical notes. Our approach involved two distinct pipelines to extract PSH and FSH, utilizing a dual lexicon strategy. This strategy comprised target STB and key historical modifiers to accurately identify and contextualize suicidality references (Supplementary Table 3).

We developed an extraction logic through a structured, iterative process. Initially, we focused on defining key terms associated with STB, including SI, SB, and SA while excluding non-life-threatening behaviors like cutting and burning. The rationale behind the exclusion of terms "cutting" and "burning" is based on our preliminary investigation of clinical notes where these terms were observed in other contexts and contributed to high false positive. Some examples from real-world data include: *"when he ran into a burning building to rescue people"*, *"he was found in his residence burning 20+ candles"*, *"Focused on burning sensation and pregnancy"*, *"Says at times he's "burned out"*, *"Will restart mirtazapine 7.5 mgs. Pt may begin by cutting this in half w/ pill-cutter."*. We should, however, acknowledge that by excluding these terms from our lexicon may introduce some false negative cases of STBs. Utilizing string matching, filters, and a series of regular expressions, we captured various expressions of these concepts (Supplementary Table 3). Context analyses were then performed by concept mapping, disambiguation, and filtering. We then identified historical indicators of STB using modifiers and paired these with core concepts using regular expressions and string matching. The ConText algorithm was employed to discern negated instances[45]. For FSH, we searched for family-related terms within a defined proximity of the concept-modifier pair (Supplementary Table 3). The final document classification combined majority polling and analysis of the last mentioned in the document. This comprehensive approach, illustrated in Fig. 1, ensured accurate and context-sensitive extraction of suicidality-related information.

For the rule-based algorithm development, we selected a sample of 1,000 notes, randomly chosen from a larger set previously flagged for containing the string "suicide", from WCM EHR. This selection strategy was aimed at enriching our dataset with more instances of suicide history, a relatively rare occurrence in general clinical documentation. The sample was divided into four batches of 100, 200, 300, and 400 notes for detailed analysis. Through this process, we continuously identified and addressed shortcomings in our algorithm, particularly in areas of lexicon usage, context analysis, and rule-based filtering and validation. This iterative refinement was conducted until the algorithm consistently and accurately extracted all relevant instances of history of STB.

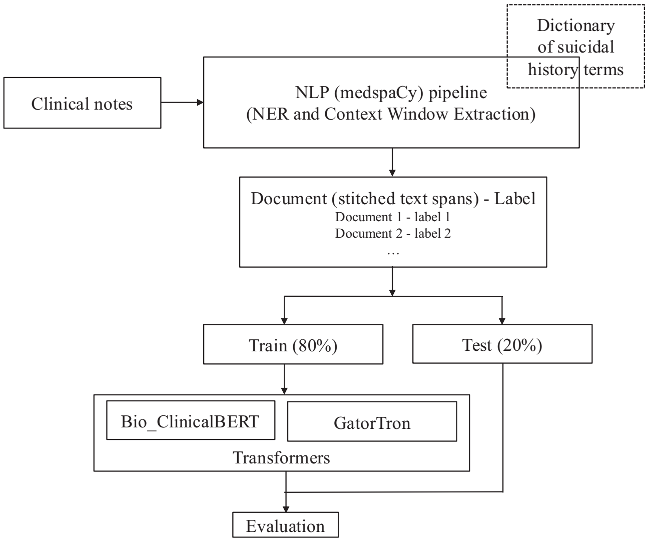
Dictionary of suicidal history terms
Clinical notes
NLP (medspaCy) pipeline (NER and Context Window Extraction)
Document (stitched text spans) - Label
Document 1 - label 1
Document 2 - label 2
…
Train (80%)
Test (20%)
Bio_ClinicalBERT
GatorTron
Transformers
Evaluation

Fig. 2 | Architecture of the transformer-based method implemented for detecting PSH and FSHin clinical notes. Transformer-based DL method implemented for the detection of PSH and FSH in clinical notes. Using spaCy and medspaCy libraries, key concepts and associated context texts were extracted and joined into a larger text representation per document and served as input to the transformer models. Bio_ClinicalBERT and the GatorTron transformer models were fine-tuned and tested using the annotated gold-standard corpus.

For each note, the NLP method attempted to extract all instances of history of STB. For notes with multiple instances, we applied a majority polling heuristic for document-level classification, consistent with the approach used in creating our gold standard. In cases where a single instance of "STB history" or "family STB history" was identified, classification was based on the presence or absence of negation. For notes with multiple instances, the majority and last-instance rule was applied. This methodology was uniformly applied across both PSH and FSH data extraction processes.

At each participating institute, the *SuicideHistoryExtractor* was deployed to analyze notes from the corpuses defined above in the "Study Setting and Data Sources" section, and the system-level performance was evaluated using the gold standard corpus detailed above in the "Evaluation of the NLP and DL tools" section. The NLP system produced two types of outputs: a raw output enumerating extracted entities, and a classification output providing a document-level classification for each note. The effectiveness of the rule-based tool was evaluated by comparing its classifications against the gold standard, using precision, recall, and F1-score metrics.

Transformer based DL tool development

We employed the BERT framework to construct a transformer-based classifier for analyzing history of STB in clinical texts, as illustrated in Fig. 2. The classifiers, pre-trained on extensive data in clinical domains, are adept at understanding context, and their effectiveness is further enhanced by fine-tuning to capture intricate interactions between clinical concepts and their contextual environment. We used a custom NLP pipeline to transform the raw text into a smaller string with key concepts and their surrounding words. The raw text instances are first tokenized using a clinical domain tokenizer implemented in the medspaCy library[46]. Key concept (anchor n-gram) for history of STB is identified using NER TargetRule in medspaCy. The target rule was developed using a dictionary of terms and phrases based on the concept-modifier identified in our rule-based NLP algorithm for PSH and FSH. Clinical notes in our corpus are structured in multiple sections and STB terms often appear in sections such as "chief complains", "history of present illness", "past medical history", review of systems, and "plans". A context window of n words to the left and n words to the right of the matched concept are extracted. A given document may have multiple instances of concept terms and we extracted text spans for all those instances. These text spans were then combined to form a text representation of that document and used for further modeling. We assigned labels to the combined text same as the label assigned by the annotators for the whole document. The optimum context window size was determined by n varying from 8, 16, 24, and 32 words to find the best performance. For our classification task, we used the Bio_ClinicalBERT[47] and the GatorTron[48] models; both of which had shown good performance when applied to clinical notes. We used Hugging Face's transformers library[49] to initialize both these models and fine-tune them using code written in PyTorch[50]. The Bio_ClinicalBERT was pretrained on MIMIC-III[51] and the GatorTron model was pretrained on EHR notes at UF[48]. We trained both models using a batch size of 16, a fixed learning rate of 1e-05, a dropout probability of 0.3, average cross-entropy loss, and AdamW[52] optimizer. Since the initial layers of the models only learn very general features, we kept them unchanged and only fine-tuned the last layers for our classification task. We tokenized and fed our input training data to fine-tune the models and then used the fine-tuned models for the test set classification. The model was trained for 5 epochs. Each model was evaluated via 10-fold cross-validation by randomly training using 80% of the data while keeping out 20% for testing to avoid overfitting. Model's performance was measured in terms precision, recall and F1-score. Since the sample cohort was unbalanced for PSH and FSH across all sites, we used the weighted average from scikit learn python library to compute these measures[53]. All experiments were run in HIPAA-compliant computing environments at the participating sites, equipped with adequate computational resources.

We further benchmarked the performance of our NLP and DL-based tools against a traditional method that relies exclusively on diagnostic codes. This comparison was conducted using the data outlined in Supplementary Table 1. This benchmarking process was crucial in assessing the relative effectiveness of the NLP and DL tools in capturing a more comprehensive and accurate picture of patients' mental health status compared to methods dependent solely on diagnostic codes. Clinical notes assembled for both WCM and UF gold standards are composed of notes written before and after the official implementation date of ICD-10, which is 1st October 2015. For WCM corpus, there were 253 notes with a report date before 1st October 2015 and 48 notes with a report date on or after 1st October 2015. Of the 400 notes annotated at UF, 34 notes were reported before 1st October 2015 and 366 notes were reported on or after 1st October 2015. For the NM cohort, all 400 notes were generated in year 2018, well after the integration of ICD-10 in EHR systems.

Data availability

The data of this study are not publicly available due to privacy and ethical restrictions. Data to support the findings of this study are available upon reasonable request.

Code availability

Code for data collection and deep learning models are available on request.

Received: 4 March 2024; Accepted: 14 September 2024;
Published online: 28 September 2024

References

1. Arensman, E., Scott, V., De Leo, D. & Pirkis, J. Suicide and suicide prevention from a global perspective. *Crisis* **41**, S3–S7 (2020).
2. Curtin, S. C. State suicide rates among adolescents and young adults aged 10–24: United States, 2000–2018. (2020).
3. Morral, A. R., Schell, T. L. & Smart, R. Comparison of suicide rates among US veteran and nonveteran populations. *JAMA Netw. Open* **6**, e2324191 (2023).
4. Carpiniello, B. & Pinna, F. The reciprocal relationship between suicidality and stigma. *Front. Psychiatry* **8**, 229216 (2017).
5. Beghi, M. et al. Suicidal behaviour in older age: a systematic review of risk factors associated to suicide attempts and completed suicides. *Neurosci. Biobehav. Rev.* **127**, 193–211 (2021).

6. Hill, K., Somerset, S., Armstrong, D., Schwarzer, R. & Chan, C. Saving lives: a systematic review on the efficacy of theory-informed suicide prevention programs. *Community Ment. Health J.* **58**, 454–473 (2022).
7. Park, E.-H., Hong, N., Jon, D.-I., Hong, H. J. & Jung, M. H. Past suicidal ideation as an independent risk factor for suicide behaviours in patients with depression. *Int J. Psychiatry Clin. Pr.* **21**, 24–28 (2017).
8. Probert-Lindström, S., Berge, J., Westrin, Å., Öjehagen, A. & Pavulans, K. S. Long-term risk factors for suicide in suicide attempters examined at a medical emergency in patient unit: results from a 32-year follow-up study. *BMJ Open* **10**, e038794 (2020).
9. Brent, D. A. & Melhem, N. Familial transmission of suicidal behavior. *Psychiatr. Clin. North Am.* **31**, 157–177 (2008).
10. Runeson, B. & Åsberg, M. Family history of suicide among suicide victims. *Am. J. Psychiatry* **160**, 1525–1526 (2003).
11. Klonsky, E. D., May, A. M. & Saffer, B. Y. Suicide, suicide attempts, and suicidal ideation. *Annu. Rev. Clin. Psychol.* **12**, 307–330 (2016).
12. Paffard, M. in *Acute Medicine: A Symptom-Based Approach* (eds. Whitehead, D., Haydock, S. & Fritz, Z.) 415–420 (Cambridge University Press, 2014).
13. Velupillai, S. et al. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. *Front. Psychiatry* **10**, 36 (2019).
14. Carter, G. et al. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *Br. J. Psychiatry* **210**, 387–395 (2017).
15. Mann, J. J., Michel, C. A. & Auerbach, R. P. Improving suicide prevention through evidence-based strategies: a systematic review. *Am. J. psychiatry* **178**, 611–624 (2021).
16. Kessler, R. C. et al. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. *Int J. Methods Psychiatr. Res.* **26**, e1575 (2017).
17. Walsh, C. G. et al. Prospective validation of an electronic health record–based, real-time suicide risk model. *JAMA Netw. Open* **4**, e211428 (2021).
18. Levis, M., Westgate, C. L., Gui, J., Watts, B. V. & Shiner, B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychol. Med.* **51**, 1382–1391 (2021).
19. Pathak, J., Kho, A. N. & Denny, J. C. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J. Am. Med. Inform. Assoc.* **20**, e206–e211 (2013).
20. Luoma, J. B., Martin, C. E. & Pearson, J. L. Contact with mental health and primary care providers before suicide: a review of the evidence. *Am. J. Psychiatry* **159**, 909–916 (2002).
21. Stanley, B. et al. Suicidal Behavior and Non-Suicidal Self-Injury in Emergency Departments Underestimated by Administrative Claims Data. *Crisis* **39**, 318–325 (2018).
22. Torbarina, L. et al. Challenges and opportunities of using transformer-based multi-task learning in NLP through ML lifecycle: A position paper. *Nat Lang Process J.* **7**, 100076 (2024).
23. Garrido-Merchan, E. C., Gozalo-Brizuela, R. & Gonzalez-Carvajal, S. Comparing BERT against traditional machinelearning models in text classification. *J Comput Cogn Eng.* **2**, 352–356 (2023).
24. Bejan, C. A. et al. Improving ascertainment of suicidal ideation and suicide attempt with natural language processing. *Sci. Rep.* **12**, 15146 (2022).
25. Carson, N. J. et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PLoS ONE* **14**, e0211116 (2019).
26. Cliffe, C. et al. Using natural language processing to extract self-harm and suicidality data from a clinical sample of patients with eating disorders: a retrospective cohort study. *BMJ Open* **11**, e053808 (2021).
27. Cook, B. L. 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 (2016).
28. Cusick, M. et al. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *J. Psychiatr. Res.* **136**, 95–102 (2021).
29. Fernandes, A. C. et al. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Sci. Rep.* **8**, 7426 (2018).
30. Meerwijk, E. L. et al. Suicide theory-guided natural language processing of clinical progress notes to improve prediction of veteran suicide risk: protocol for a mixed-method study. *BMJ Open* **12**, e065088 (2022).
31. Tsui, F. R. et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open* **4**, ooab011 (2021).
32. Xie, F., Ling Grant, D. S., Chang, J., Amundsen, B. I. & Hechter, R. C. Identifying suicidal ideation and attempt from clinical notes within a large integrated health care system. *Perm. J.* **26**, 85–93 (2022).
33. Zhong, Q.-Y. 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.* **34**, 153–162 (2019).
34. Rihmer, Z., Belsö, N. & Kiss, K. Strategies for suicide prevention. *Curr. Opin. Psychiatry* **15**, 83–87 (2002).
35. Ribeiro, J. D. et al. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychol. Med.* **46**, 225–236 (2016).
36. Adamou, M. et al. Mining free-text medical notes for suicide risk assessment. In: *Proceedings of the 10th hellenic conference on artificial intelligence* 1–8 (Association for Computing Machinery, 2018).
37. Bittar, A., Velupillai, S., Roberts, A. & Dutta, R. Text classification to inform suicide risk assessment in electronic health records. *Stud Health Technol Inform.* **264**, 40–44 (2019).
38. Ben-Ari, A. & Hammond, K. Text mining the EMR for modeling and predicting suicidal behavior among US veterans of the 1991 Persian Gulf War. In: *2015 48th Hawaii International Conference on System Sciences* (eds. Bui, T. & Sprague, Jr., R.) 3168–3175 (IEEE Computer Society, 2015).
39. McCoy, T. H., Castro, V. M., Roberson, A. M., Snapper, L. A. & Perlis, R. H. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry* **73**, 1064–1071 (2016).
40. Kariotis, T. C., Prictor, M., Chang, S. & Gray, K. Impact of electronic health records on information practices in mental health contexts: scoping review. *J. Med Internet Res.* **24**, e30405 (2022).
41. Adekkanattu, P. et al. Ascertaining depression severity by extracting patient health questionnaire-9 (PHQ-9) scores from clinical notes. *AMIA Annu. Symp. Proc.* **2018**, 147–156 (2018).
42. Oexle, N., Waldmann, T., Staiger, T., Xu, Z. & Rüsch, N. Mental illness stigma and suicidality: the role of public and individual stigma. *Epidemiol. Psychiatr. Sci.* **27**, 169–175 (2018).
43. NLPLAB. Brat Rapid Annotation tool. https://brat.nlplab.org.
44. VINCI-Leo. *Department-of-veterans-affairs/Leo* https://github.com/department-of-veterans-affairs/Leo.
45. Harkema, H., Dowling, J. N., Thornblade, T. & Chapman, W. W. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. Biomed. Inf.* **42**, 839–851 (2009).
46. Eyre, H. et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu. Symp. Proc.* **2021**, 438 (2021).
47. Alsentzer, E. et al. Publicly available clinical embeddings. In: *Proc. 2nd Clinical Natural Language Processing Workshop* (eds. Rumshisky, A., Roberts, K., Bethard, S. & Naumann, T.) 72–78 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019). https://doi.org/10.18653/v1/W19-1909.

48. Yang, X. et al. A large language model for electronic health records. *NPJ Digit Med.* **5**, 194 (2022).
49. Wolf, T. et al. Transformers: State-of-the-art natural language processing. In: *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds. Liu, Q. & Schlangen, D.) 38–45 (Association for Computational Linguistics, 2020).
50. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).
51. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
52. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://arxiv.org/abs/1711.05101 (2017).
53. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

This research is funded in part by NIMH grant 5R01MH121907-04, 5R01MH121922-05 and NIDA 1R01DA050676.

Author contributions

Study conception and design: P.A., J.P., A.F., Y.W., D.O. Data collection: P.A., J.P., A.F., Y.L., Y.W., S.B., T.L.W., W.L.J. Analysis: P.A., A.F., A.P., G.H.W., B.G.P., Y.Y., N.J.F. Interpretation of results: P.A., J.P., A.F., Y.W., S.B., A.P., D.M. Participated in discussion: P.A., J.P., A.F., Y.W., B.G.P., S.B., A.P., S.H., D.M., T.L.W., W.L.J., W.G., J.B., Y.B., M.W., D.O. Draft manuscript preparation: P.A., J.P., A.F., Y.W., A.P., Y.B., W.L.J., B.G.P., D.M., J.B. Managed the overall study: J.P. All authors reviewed the results and approved the final version of the manuscript

Competing interests

One of the authors, Y.L., is an Associate Editor of npj Digital Medicine. The author is not involved in the editorial review of this manuscript. Other authors declare no financial or non-financial competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01266-7.

Correspondence and requests for materials should be addressed to Prakash Adekkanattu.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.