European Journal of Epidemiology (2019) 34:153–162
https://doi.org/10.1007/s10654-018-0470-0

PSYCHIATRIC EPIDEMIOLOGY

Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem

Qiu-Yue Zhong[1] · Leena P. Mittal[2] · Margo D. Nathan[2] · Kara M. Brown[2] · Deborah Knudson González[3] · Tianrun Cai[4] · Sean Finan[5] · Bizu Gelaye[1] · Paul Avillach[1,5,6] · Jordan W. Smoller[1,7] · Elizabeth W. Karlson[4] · Tianxi Cai[6,8] · Michelle A. Williams[1]

Received: 20 May 2018 / Accepted: 1 December 2018 / Published online: 10 December 2018

Abstract

We developed algorithms to identify pregnant women with suicidal behavior using information extracted from clinical notes by natural language processing (NLP) in electronic medical records. Using both codified data and NLP applied to unstructured clinical notes, we first screened pregnant women in Partners HealthCare for suicidal behavior. Psychiatrists manually reviewed clinical charts to identify relevant features for suicidal behavior and to obtain gold-standard labels. Using the adaptive elastic net, we developed algorithms to classify suicidal behavior. We then validated algorithms in an independent validation dataset. From 275,843 women with codes related to pregnancy or delivery, 9331 women screened positive for suicidal behavior by either codified data (N = 196) or NLP (N = 9,145). Using expert-curated features, our algorithm achieved an area under the curve of 0.83. By setting a positive predictive value comparable to that of diagnostic codes related to suicidal behavior (0.71), we obtained a sensitivity of 0.34, specificity of 0.96, and negative predictive value of 0.83. The algorithm identified 1423 pregnant women with suicidal behavior among 9331 women screened positive. Mining unstructured clinical notes using NLP resulted in a 11-fold increase in the number of pregnant women identified with suicidal behavior, as compared to solely reliance on diagnostic codes.

Keywords Natural language processing · Suicidal behavior · Pregnant women · Electronic medical · Records · Classification algorithm

Introduction

Suicide is one of the leading causes of maternal deaths [1, 2]. For example, based on the 1997–1999 report of the Confidential Enquiries into Maternal Death [1], 12% of maternal deaths were due to psychiatric disorders and 10% to suicide

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10654-018-0470-0) contains supplementary material, which is available to authorized users.

✉ Qiu-Yue Zhong
qyzhong@mail.harvard.edu

[1] Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA, USA

[2] Division of Women's Mental Health, Department of Psychiatry, Brigham and Women's Hospital, Boston, MA, USA

[3] Department of Psychiatry and Behavioral Neurosciences, Morsani College of Medicine, University of South Florida, Tampa, FL, USA

[4] Department of Medicine, Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA, USA

[5] Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA, USA

[6] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[7] Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

[8] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[1]. Up to 20% of postpartum deaths are accounted for by suicide [3]. Suicide may be prevented if prompt action and immediate interventions are taken to mitigate risk as part of prenatal and postpartum care. Although studied for decades, the epidemiology of nonfatal suicidal ideation and behavior (hereinafter referred to as “suicidal behavior”) among pregnant women has been difficult to characterize due, in part, to low prevalence and incomplete ascertainment. In a recent study on suicidal behavior in U.S. pregnant women using the National Inpatient Sample, the prevalence of suicidal behavior was 142.2 per 100,000 pregnancy- and delivery-related hospitalizations in 2012 [4]. In prior electronic medical record (EMR)-based studies, the identification of suicidal behavior has typically relied on codified data such as the International Classification of Diseases (ICD) billing codes. However, suicidal behavior is often “under-coded” with ICD codes capturing only a small proportion of suicidal behavior cases [5–8]. Notably, in a review [9], Walkup et al. noted that the sensitivity (the proportion of true suicidal cases that were identified as positive by the ICD codes) associated with the use of ICD diagnostic codes to classify patients with suicidal behavior ranged from 0.13 to 0.65 when compared to the more labor-intensive, time consuming manual chart review of clinical notes or review of death registry [9]. Such low sensitivity of diagnostic codes suggests that a sizable portion of suicidal cases may be missed when case-finding relies on ICD codes alone [10].

Recognizing the limitations associated with reliance on codified data, efforts have been made to identify patients with suicidal behavior from clinical notes using advanced informatics approaches such as natural language processing (NLP). NLP is a process whereby information residing in unstructured clinical notes are extracted and then converted into a more structured analyzable layout [11]. For example, using Medical Language Extraction and Encoding System (MedLEE), a clinical NLP engine, Haerian et al. [12]. showed that their NLP algorithm identified nearly nine-fold more cases of suicidal behavior by drug overdose as compared to the ICD codes (4087 for NLP verse 469 for ICD). In another study, Anderson et al. [7]. developed a rule-based NLP algorithm searching for positive mentions of suicidal behavior in clinical notes to identify suicidal behavior among 15,761 patients with at least one diagnostic code of depression. The application of the NLP algorithm resulted in a 34-fold increase in the number of patients identified with suicidal ideation (1025 for NLP versus 30 for ICD) and a five-fold increase in patients identified as having attempted suicide (86 for NLP versus 16 for ICD). However, neither of these two NLP algorithms went beyond searching for concepts directly related to suicidal behavior, and no further machine learning algorithm was used. Taken together, these studies illustrate the clinical relevance of applying NLP approaches to identifying patients with suicidal behavior. In our study, we sought to develop classification algorithms specific to pregnant women, a population that is understudied when it comes to understanding the determinants and sequelae of suicidal behavior [13].

Here, we used EMRs from a large healthcare system (Partners HealthCare) to develop a classification algorithm that would accurately identify pregnant women with suicidal behavior. We extracted diagnostic data from both structured codified data and unstructured clinical notes processed by NLP. We assessed the diagnostic validity of the algorithm against gold-standard labels obtained from manual chart reviews by psychiatrists and a trained researcher.

Methods

Data source and study population

EMR data were extracted from the Partners HealthCare System Research Patient Data Registry (RPDR), a clinical data warehouse that gathers medical records for nearly 4.6 million patients from Massachusetts General Hospital (MGH) and Brigham and Women’s Hospital (BWH), as well as community and specialty hospitals in the Boston area. The RPDR is updated every 1.5–2 months. A quality assurance process is performed every time the RPDR is updated to ensure the data is being uploaded correctly. The RPDR includes socio-demographic data, vital signs, laboratory and test results, problem list entries, prescribed medications, billing codes, and clinical notes for healthcare services provided within the system [14]. For clinical notes, the RDPR includes the ambulatory notes, discharge summaries, EPIC progress reports (such as emergency department (ED) observation progress notes, labor and delivery notes, lactation notes, progress notes, and significant event notes), operative notes, and pathology, cardiology, endoscopy, pulmonary, and radiology reports. The clinical notes were directly entered into the EPIC EMR system by doctors. There is no technical constraint as to the number of characters or words allowed in a note. The Institutional Review Board of Partners HealthCare (Protocol Number: 2016P000775/BWH) and Harvard T.H. Chan School of Public Health (Protocol Number: IRB16-0899) approved all aspects of this study.

We initially searched for women within a priori age range 10-64 years with at least one diagnostic code related to pregnancy or delivery (International Classification of Diseases-10 [ICD-10]: Z3A.*, O0.*- O9.*; ICD-9: 640.*-679.*, V22.*, V23.*, V24.*, V27.*, V28.*; Diagnosis-Related Group [DRG]: 370–384) in the EMRs from January 1, 1996 to March 31, 2016 (Fig. 1). A set of 275,843 women (hereinafter referred to as “datamart”) were identified.

We first screened for suicidal behavior using diagnostic codes (initial screen) including the ICD codes and the

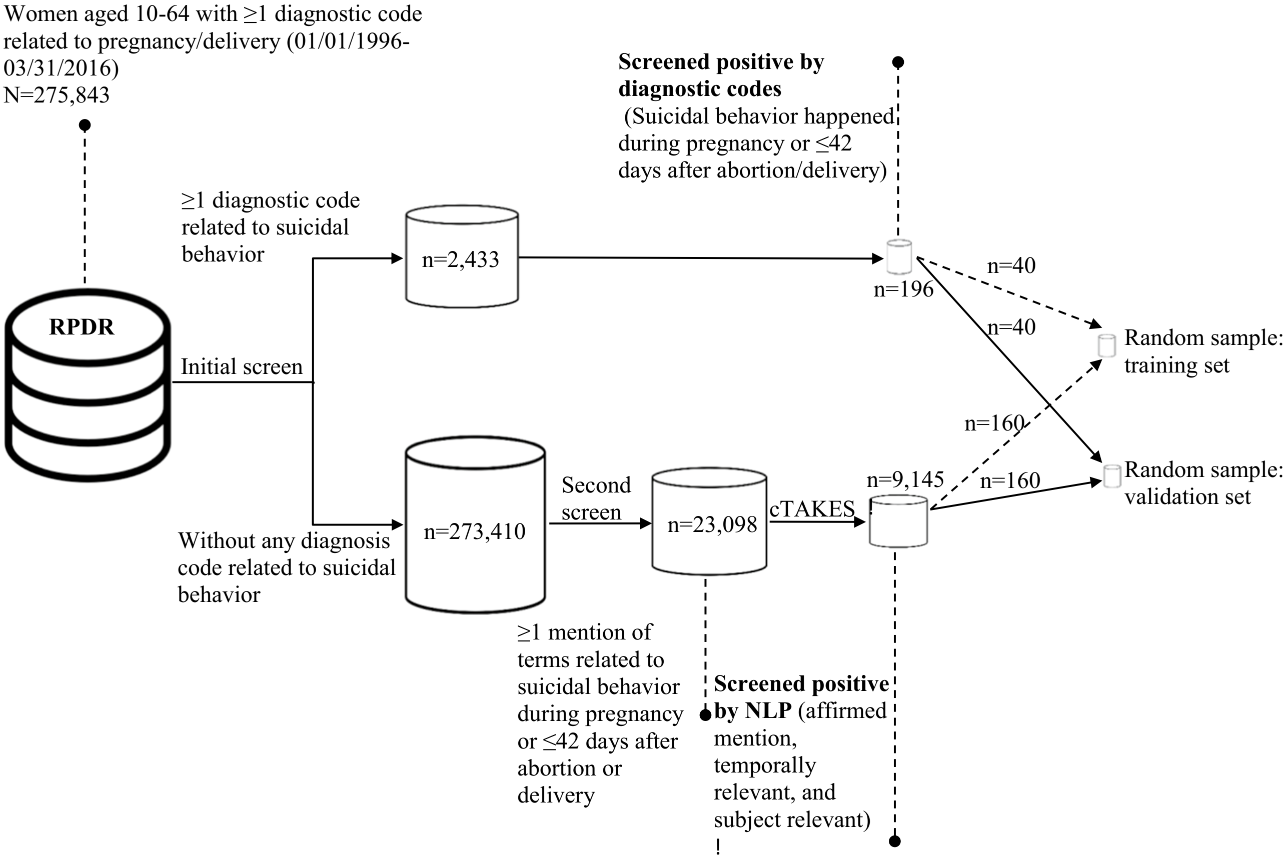
Women aged 10-64 with ≥1 diagnostic code related to pregnancy/delivery (01/01/1996-03/31/2016)
N=275,843
RPDR
Initial screen
≥1 diagnostic code related to suicidal behavior
n=2,433
Screened positive by diagnostic codes
(Suicidal behavior happened during pregnancy or ≤42 days after abortion/delivery)
n=196
n=40
n=40
Random sample: training set
Random sample: validation set
n=160
n=160
n=9,145
Without any diagnosis code related to suicidal behavior
n=273,410
Second screen
n=23,098
cTAKES
≥1 mention of terms related to suicidal behavior during pregnancy or ≤42 days after abortion or delivery
Screened positive by NLP (affirmed mention, temporally relevant, and subject relevant)
!

Fig. 1 Process for screening suicidal behavior using diagnostic codes vs. NLP among pregnant women. *NLP* Natural Language Processing; *RPDR* Research Patient Data Registry

Longitudinal Medical Record (LMR) codes. The LMR codes were assigned to problem list conditions in the ambulatory EMR system used across Partners HealthCare System (Supplemental Table 1). In addition to the explicit diagnostic codes for suicidal ideation (e.g., ICD-9 V62.84) and suicide attempt (e.g., ICD-9 E95*), we included additional sets of ICD code categories (poisoning by analgesics, antipyretics, and antirheumatics; poisoning by sedatives and hypnotics; and poisoning by psychotropic agents) with positive predictive value ≥ 0.8 for suicidal behavior, based on a previous study [15]. Among the 275,843 women with at least one diagnostic code related to pregnancy or delivery, 2433 women had at least one diagnostic code related to suicidal behavior, of whom 196 (hereinafter referred to as the "diagnostic code group") had a diagnostic code that occurred during pregnancy, or within 42 days after abortion or delivery (Fig. 1) [16]. This 42-day threshold was chosen a priori based on the World Health Organization (WHO) definition of maternal death [16].

In addition, among the 273,410 women without any diagnostic codes related to suicidal behavior, we searched clinical notes and identified 23,098 women with at least one mention of the terms related to suicidal behavior during pregnancy or within the 42 days after abortion or delivery (second screen) [5] (Supplemental Table 2).

Using the clinical Text Analysis and Knowledge Extraction System (cTAKES, http://ctakes.apache.org/) [17], we then processed the clinical notes of the 23,098 women, along with the notes of 196 women from the diagnostic code group. The process along with examples of clinical notes has been described in detail previously [10]. In summary, cTAKES is a comprehensive clinical NLP tool that processes clinical notes and identifies terms in English. cTAKES maps the terms to a subset of the Unified Medical Language System (UMLS) Metathesaurus [18] and assigns each term a UMLS concept unique identifier (CUI). Here we used the default fast dictionary lookup containing roughly 500,000 synonyms for 250,000 terms in the UMLS. cTAKES also extracts qualifying attributes (including negation, temporality, and subject status) associated with each CUI. To be considered as relevant, CUIs must be tagged as "affirmed" by the negation module, "overlap" or "before/overlap" by the *DocTimeRel* module (temporally relevant), and "patient" by the *Subject* module (subject relevant).

We calculated the proportions of affirmed, temporally relevant, and subject relevant CUIs related to suicidal behavior among all CUIs related to suicidal behavior (Supplemental Table 3) for each woman, and selected 9145 women (hereafter referred to as the "NLP group") with these proportions greater than or equal to 0.25. The 9145 women from the NLP group (screened positive by NLP), together with the 196 women from the diagnostic code group (screened positive by diagnostic codes), comprised the "screened positive group" (N = 9331). Of note, the prevalence of confirmed suicidal behavior was 1% among those who screened negative by cTAKES, and the prevalence of confirmed suicidal behavior was 30% among those who screened positive by cTAKES.

Training and validation data set

A total of 200 women, 40 from the diagnostic code group (n = 196) and 160 from the NLP group (n = 9145), were randomly selected as the gold-standard training set. Another 200 women, again 40 from the diagnostic code group and 160 from the NLP group, who were not part of the gold-standard training set, were randomly selected as the gold-standard validation set (Fig. 1).

Chart review to obtain gold-standard labels

Detailed medical record reviews of 400 charts (200 from the training set, and 200 from the validation set) were performed by one of the authors (Q.Y.Z., [400 charts]) and three experienced, board-certified psychiatrists (M.D.N. [200 charts], K.M.B. [100 charts], and D.K.G. [100 charts]) with expertise in women's mental health. Each chart was reviewed by two reviewers and interrater agreement between them was assessed using the Cohen's kappa coefficient [19, 20]. An independent, board-certified psychiatrist (L.P.M.) conducted a final review to achieve consensus and adjudicated disagreements among the first two reviewers. Review guidelines for assigning diagnostic status were adapted from the Columbia Classification Algorithm of Suicide Assessment (C-CASA) [21] (Supplemental Table 4). The reviewers assigned each woman a classification of (1) suicidal behavior, (2) non-suicidal behavior, (3) intermediate or potentially suicidal behavior, or (4) none of the above. For the purpose of classification algorithm development, we considered women in the latter three groups (including indeterminate or potentially suicidal events, non-suicidal behavior, or none) as "without suicidal behavior" (Supplemental Table 4).

Classification algorithms

We created a list of expert-curated features (Supplemental Table 5) derived from patients' clinical notes (Feeling hopeless, Feeling relief, Tired, Love, Feeling empty, Feeling content, Low self-esteem, Impulsive character, Isolation, Distractibility, Childhood adversity, Adult sexual abuse, Severe depression, Substance abuse problem, Personality Disorders, Psychotic Disorders, Seizures, Anxiety Disorders, Wound and injury, Abortion) or medications that are likely to be associated with suicidal behavior. Because the number of ICD codes and NLP mentions for the phenotypes of interest are typically the most predictive features [22–26], we also include the ICD and NLP counts of suicidal behavior (hereafter referred to as the "main ICD count" and the "main NLP count"). The main ICD count included counts of both ICD and LMR codes related to suicidal behavior (Supplemental Table 1), with multiple diagnostic codes that occurred on the same day being counted only once (i.e., if a woman had more than one diagnostic codes related to suicidal behavior on the same day, these diagnostic codes only contributed one to the main ICD count). The main NLP count corresponded to the counts of CUIs related to suicidal behavior (Supplemental Table 6). The main ICD and NLP counts, age at the index suicidal behavior, number of clinical notes, number of psychiatric hospitalizations along with counts for each concept in the expert-curated feature list derived from either clinical notes or medications (Supplemental Table 5) were included as predictors for the algorithm training. We trained the classification algorithms by fitting the adaptive elastic net penalized logistic regression, with gold-standard labels being the response variables and aforementioned features being the predictors while accounting for the sampling design (40 sampled from the diagnostic code group and 160 sampled from the NLP group). All count variables were transformed by $x \to \log(x+1)$. The adaptive elastic net can simultaneously perform feature selection and model estimation [27]. We selected the turning parameter controlling the amount of penalty for model complexity based on the Bayesian information criterion [28]. For this analysis, we trained algorithms to distinguish women with suicidal behavior vs. without suicidal behavior (Supplemental Table 4) given their feature information. A woman was classified as having suicidal behavior if the predicted probability of suicidal behavior exceeded a threshold value. We selected the threshold values (0.388 and 0.516) for classifying suicidal behavior by setting the specificity level at 0.90 and by setting the PPV level at 0.71, respectively. We chose this PPV level given that the PPV of at least one diagnostic code related to suicidal behavior was also 0.71.

We applied the model to the validation set to evaluate algorithm performance. We calculated the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, positive predictive value (PPV), and negative predictive value (NPV) corresponding to the specified specificity levels. We then applied the classification algorithm to the entire screened positive group to identify pregnant women with suicidal behavior. All analyses were completed using R [29].

Results

Manual chart review in training and validation data sets

In the training set of 200 women, for the 40 women sampled from the diagnostic code group, 29 (73%) women were confirmed to have suicidal behavior and 11 did not have suicidal behavior; for the 160 women sampled from the NLP group, 44 (28%) women had suicidal behavior and 116 had no indication of suicidal behavior. In the validation set of 200 women, for the 40 women from the diagnostic code group, 32 (80%) women had suicidal behavior and 8 had no suicidal behavior; for the 160 women sampled from the NLP group, 40 (25%) women had suicidal behavior and 120 had no suicidal behavior. We compared the distribution of women across the two datasets and found that the inter-rater agreement for suicidal behavior categorization was substantial (Cohen's kappa statistic = 0.75, 95% CI 0.69–0.81).

Summary of model performances

We validated 3 algorithms to classify suicidal behavior: (1) main ICD count only, (2) main ICD and NLP counts, and (3) main ICD and NLP counts along with expert-curated features. Summary statistics documenting the performance of each of the four algorithms in the validation set are presented in Tables 1 and 2, and Fig. 2. The AUC for the suicidal behavior algorithm based solely on ICD codes was 0.53. Inclusion of NLP count resulted in a substantial increase (AUC = 0.67) in AUC. The algorithm using the expert-curated features (Tables 1, 3) achieved an overall AUC of 0.83. For this algorithm, by setting the specificity level at 0.90, we obtained an estimated sensitivity of 0.58, a PPV of 0.63, and a NPV of 0.88 (Table 1). By setting a PPV comparable to the PPV of at least one diagnostic code related to suicidal behavior (0.71), we obtained an estimated sensitivity of 0.34, a specificity of 0.96, and a NPV of 0.83 (Table 2). Given that multiple screen steps were implemented prior to the classification algorithm, we calculated the NPV of our classification rule in our entire datamart. Based on chart review in one previous study [10], the NPV was 1.00 for the initial screen and the NPV was 0.99 for the second screen. Taken together, for the classification algorithm using the expert-curated features with a PPV of 0.71, we obtained a projected NPV of 0.994 in our entire datamart (Table 2).

Classification in the screened positive group

After running the classification algorithm using the expert-curated features by setting the PPV at 0.71 on the entire screened positive group, a set of 1423 pregnant women (125 from the diagnostic code group and 1298 from the NLP group) were identified as with suicidal behavior. The estimated prevalence of suicidal behavior among our study population was 515.87 per 100,000 pregnant women.

Discussion

In EMR-based studies, NLP combined with statistical classification algorithms has emerged as a useful tool [30] to improve the identification of phenotypes such as treatment resistant depression, cerebral aneurysms, Crohn's disease, ulcerative colitis, and rheumatoid arthritis [22, 31–34]. However, no NLP-based, supervised classification algorithm

Table 1 Model performances for algorithms to classify women with suicidal behavior

Algorithms	% Women with suicidal behavior classified by algorithms (95% CI)	Sensitivity (95% CI)	PPV (95% CI)	NPV (95% CI)	AUC (95% CI)
Specificity = 0.90					
Main ICD count	0.02 (0.01, 0.02)	0.14 (0.12, 0.16)	0.88 (0.80, 0.95)	0.80 (0.74, 0.85)	0.53 (0.52, 0.55)
Main ICD + NLP counts	0.15 (0.11, 0.18)	0.31 (0.14, 0.46)	0.48 (0.28, 0.59)	0.82 (0.75, 0.87)	0.67 (0.58, 0.75)
Main ICD + NLP counts + Expert-curated features[a]	0.21 (0.15, 0.25)	0.58 (0.34, 0.72)	0.63 (0.48, 0.71)	0.88 (0.81, 0.93)	0.83 (0.76, 0.89)

ICD international classification of diseases, *NLP* natural language processing, *AUC* area under the receiver-operating characteristic curve, *CI* confidence interval, *PPV* positive predictive value, *NPV* negative predictive value

[a] including main ICD count, main NLP count, age at the index suicidal behavior, number of clinical notes, number of psychiatric hospitalizations, Feeling hopeless, Feeling relief, Tired, Love, Feeling empty, Feeling content, Low self-esteem, Impulsive character, Isolation, Distractibility, Childhood adversity, Adult sexual abuse, Severe depression, Substance abuse problem, Personality Disorders, Psychotic Disorders, Seizures, Anxiety Disorders, Wound and injury, Abortion, Medication (Serotonin–norepinephrine reuptake inhibitor), Medication (Buprenorphine/Naloxone)

Table 2 Model performances for algorithms to classify women with suicidal behavior

Algorithms	%Women with suicidal behavior classified by algorithms (95% CI)	Sensitivity (95% CI)	Specificty (95% CI)	PPV (95% CI)	NPV[b] (95% CI)	Projected NPV[c] (95% CI)	AUC (95% CI)
PPV ≈ 0.71							
Main ICD count ≥1	0.02 (0.01, 0.02)	0.05 (0.03, 0.07)	0.99 (0.99, 1.00)	0.71 (0.58, 0.87)	0.78 (0.72, 0.84)	0.992 (0.990, 0.994)	0.53 (0.52, 0.55)
Main ICD + NLP counts	0.02 (0.01, 0.02)	0.06 (0.03, 0.09)	0.99 (0.99, 1.00)	0.69 (0.62, 1.00)	0.78 (0.73, 0.84)	0.992 (0.991, 0.994)	0.67 (0.58, 0.75)
Main ICD + NLP counts + Expert-curated features[a]	0.11 (0.07, 0.16)	0.34 (0.19, 0.54)	0.96 (0.95, 0.97)	0.71 (0.54, 0.81)	0.83 (0.78, 0.89)	0.994 (0.993, 0.996)	0.83 (0.76, 0.89)

ICD international classification of diseases, *NLP* natural language processing, *AUC* area under the receiver-operating characteristic curve, *CI* confidence interval, *PPV* positive predictive value, *NPV* negative predictive value

[a] including main ICD count, main NLP count, age at the index suicidal behavior, number of clinical notes, number of psychiatric hospitalizations, Feeling hopeless, Feeling relief, Tired, Love, Feeling empty, Feeling content, Low self-esteem, Impulsive character, Isolation, Distractibility, Childhood adversity, Adult sexual abuse, Severe depression, Substance abuse problem, Personality Disorders, Psychotic Disorders, Seizures, Anxiety Disorders, Wound and injury, Abortion, Medication (Serotonin–norepinephrine reuptake inhibitor), Medication (Buprenorphine/Naloxone)

[b] NPV for the algorthim

[c] NPV in the entire datamart. NPV was 1.00 for the initial screen and the NPV was 0.99 for the second screen

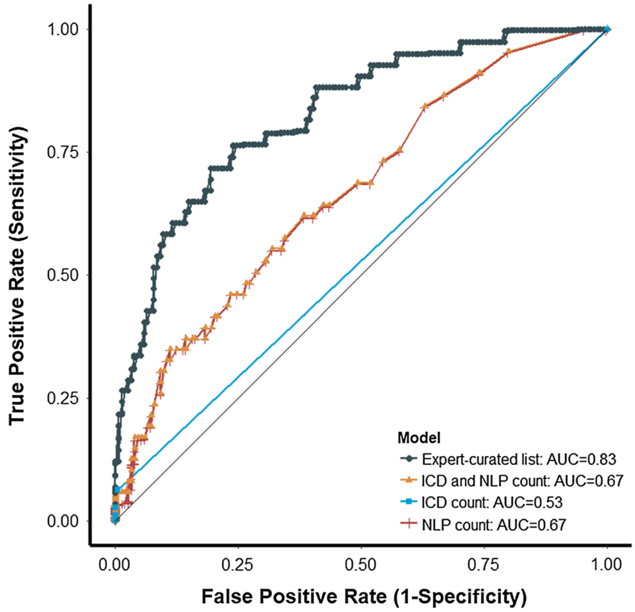

Fig. 2 Area Under the Curve (AUC) for algorithms to classify women with suicidal behavior

has been developed for suicidal behavior. In our large EMR system, the classification algorithm we developed to identify pregnant women with suicidal behavior supported the use of informatic-based phenotyping for suicide research. The algorithm, incorporating codified data, NLP mentions of suicidal behavior, and features curated by domain experts, yielded an AUC of 0.83. The addition of NLP resulted in a 11-fold increase in the number of pregnant women identified with suicidal behavior (from 125 to 1423). Based on our algorithm, over the study period of 20 years, the estimated prevalence of suicidal behavior was 515.87 per 100,000 women with diagnostic codes related to pregnancy or delivery.

The advantages of mining medical records using NLP to identify suicidal behavior cases were two-fold. First, NLP was particularly helpful in improving the sensitivity of case identification for suicidal behavior in EMRs. Prior research has similarly demonstrated the utility of NLP in improving case detection for a range of diseases [24, 33]. In our study, the use of NLP has successfully increased the number of suicidal behavior cases from 125 to 1423 while maintaining a similar PPV to diagnostic codes of suicidal behavior. Second, besides extracting codifiable concepts (e.g., "Suicidal behavior") from clinical notes that were not coded at the time of the encounter, we showed that NLP could also extract metrics that were not candidates for coding but could be used as candidate features (e.g., "Feeling hopeless," "Feeling relief," "Feeling content," and "Love") for classification algorithms. Such metrics that reflect individual symptoms and emotions, personalities, adverse psychosocial factors [35], as well as clinician perception not reflected elsewhere, might be helpful in identifying suicidal behavior, for which no established biological marker for diagnosis exists [36]. Moreover, although many risk factors have been identified for suicidal behavior, these factors only explain a small amount of the variance in suicidal behavior [15, 36]. Augmenting feature information extracted from clinical notes that are not otherwise reflected in codified data or not collected conventionally, such as positive and negative valence [37], may open new possibilities for identifying additional risk factors for suicidal behavior.

Our findings strongly suggest that NLP substantially improves the identification of patients with suicidal behavior using information in EMRs. In our study population, codified data missed more than 90% of suicidal behavior cases. This suggests that suicidal behavior is frequently "under-coded", perhaps due to concern about stigmatization [8] or financial disincentives (since suicidal behavior codes do not directly impact reimbursement) [8]. The estimated prevalence of suicidal behavior in our study was considerably higher than that reported in prior studies of pregnant women based on codified data. For example, using the National (Nationwide) Inpatient Sample, the largest all-payer inpatient care database in the U.S., the prevalence of suicidal behavior was 142.2 per 100,000 pregnancy- and delivery-related hospitalizations in 2012 [4]. Using a linked vital Statistics-Patient Discharge database of the State of California [38] from 1991 to 1999, the prevalence of suicide attempt was 40 per 100,000 pregnancies. Our study supports Walkup et al.'s conclusion that caution should be exercised when interpreting studies that rely solely on codified billing data as measures of suicidal behavior [9].

Using suicidal behavior definitions based on the C-CASA [21], we were able to achieve substantial but not excellent interrater agreement for the gold-standard labels. The percentage of disagreement between reviewers was 9.4% in our dataset. We found that the discrepancies between reviewers came primarily from disagreement in suicidal intent when retrospectively reviewing clinical notes. By definition, suicidal intent is used to distinguish between suicidal behavior and non-suicidal self-injurious behavior, which refers to direct, deliberate destruction of one's own body tissue in the absence of intent to die [39, 40]. However, less than one-third of patients with suicidal behavior express their suicidal intent to healthcare providers [41], and the documentation of suicidal behavior in clinical notes was variable and limited, particularly regarding suicidal intent. As some clinical notes lacked clearly stated suicidal intent, reviewers had to make inferences based upon the details of behavior and related clinical data. Furthermore, suicidal intent per se can be ambiguous [39], which further increases the difficulty in classifying suicidal behavior.

Table 3 Variables in the algorithms using expert-curated features

Variables	% With non-zero counts	Coefficient	95% CI
Main ICD count	18.14	1.417	0.000, 5.200
Main NLP count	96.20	1.128	0.576, 2.553
Age at the index suicidal behavior	100.00	− 1.641	− 4.107, 0.790
Number of clinical notes	100.00	− 0.473	− 1.707, 0.244
Number of psychiatric hospitalizations	19.41	0.172	− 1.337, 2.128
NLP: feeling hopeless	23.63	0.956	− 0.004, 2.726
NLP: feeling relief	16.03	0.679	− 0.152, 2.056
NLP: tired	28.69	0.449	− 0.271, 1.497
NLP: love	11.81	− 0.159	− 2.214, 1.586
NLP: feeling empty	0.00	—	—
NLP: feeling content	49.79	− 0.690	− 2.039, 0.212
NLP: low self-esteem	5.49	− 2.386	− 7.131, 0.007
NLP: Impulsive character	10.55	0.258	− 1.422, 1.759
NLP: isolation	7.17	0.217	− 1.953, 1.661
NLP: distractibility	0.00	—	—
NLP: childhood adversity	7.17	0.233	− 1.252, 5.335
NLP: adult sexual abuse	24.47	− 0.103	− 0.877, 0.793
NLP: severe depression	5.06	0.889	− 0.470, 3.919
NLP: substance abuse problem	48.10	− 0.381	− 1.550, 0.250
NLP: personality disorders	23.63	0.395	− 0.371, 1.247
NLP: psychotic disorders	29.54	− 0.579	− 1.809, 0.002
NLP: seizures	27.85	0.537	− 0.419, 1.844
NLP: anxiety disorders	68.35	− 0.164	− 0.748, 0.316
NLP: wound and injury	34.60	− 0.128	− 1.166, 0.573
NLP: abortion	32.91	− 0.098	− 0.883, 0.362
Medication: SNRI	2.95	− 1.210	− 9.745, 0.948
Medication: buprenorphine/naloxone	4.22	0.999	− 2.834, 11.735

CI confidence interval, *ICD* international classification of diseases, *NLP* natural language processing, *SNRI* serotonin–norepinephrine reuptake inhibitor

We note several caveats in interpreting our work. First, using the expert-curated list of features, we obtained an acceptable but not excellent AUC (0.83). Indeed, suicidal behavior is a complex classification problem [42]. Our estimated AUC might be due to some degree of diagnostic imprecision from gold-standard labels (as noted above), which was likely unavoidable considering the complexity of our phenotype. If inaccurately labeled data were used, the accuracy of classification might be compromised [43]. Another possible explanation for our AUC estimate could be that features manually curated by domain experts in our classification algorithm were not sufficiently informative. We included many known risk indicators and "warning signs" for suicidal behavior such as the presence of psychiatric disorders, younger age, hopelessness, social stressors such as history of childhood abuse or sexual abuse [35, 44, 45]. However, there is a limited understanding of suicidal behavior in the literature, and no single or panel of factors stood out as particularly strong in predicting suicidal behavior [15, 44]. To improve AUC, studies may benefit from investigating the features generated by NLP. In addition, automated feature selection methods that can choose the most highly informative features and machine learning algorithms that take the complex relationships among large numbers of potential features into consideration are needed [23, 42, 44, 46]. Second, we undoubtedly missed some cases of suicidal behavior. Although greatly improved as compared to the codified data, the sensitivity of our best performing algorithm was still relatively low. Further, suicidal behavior may not be well captured in clinical notes [47]. Recently the American College of Obstetrics and Gynecologists and the U.S. Preventive Services Task Force recommended that clinicians screen patients at least once during the perinatal period for depressive symptoms using validated tools [48]. These tools often include an item to screen for suicidal ideation. The incorporation of such screening tools in EMR systems may improve the capacity of future suicidal behavior identification. Third, given the small number of women classified as having suicidal behavior, we did not further classify the subtypes of suicidal behavior such as suicidal ideation and suicide attempt. Fourth, while our study population included more than 270,000 women observed for two decades, they still reflected the population in a single academic healthcare system. Further investigation to understand the extent to which our algorithms generalize to other healthcare systems is needed. Fifth, within our study period of more than 20 years, changes in coding [49] and diagnostic criteria [50] of suicidal behavior, and language used to describe suicidal behavior in clinical notes [51] might also affect the ascertainment of suicidal behavior.

Conclusion

To our knowledge, this is the first classification algorithm using NLP in addition to codified billing data to improve the case identification of suicidal behavior among pregnant women in EMRs. We showed that mining unstructured clinical notes using NLP substantially improves the detection of suicidal behavior. The addition of NLP resulted in a 11-fold increase in the number of pregnant women with suicidal behavior. Codified billing data alone are unlikely to be adequate for the identification of suicidal behavior cases in EMRs. We illustrated that augmenting feature information extracted from clinical notes by NLP that were not otherwise reflected in codified data or not collected conventionally might be helpful in identifying suicidal behavior. Our result is a step towards solving the complex classification problem of suicidal behavior. The framework we demonstrated here provides a high-throughput and cost-effective approach for studying suicidal behavior among pregnant women. We envision this algorithm will serve as a powerful foundation of EMR research for future epidemiologic, genetic, and clinical studies on suicidal behavior.

Acknowledgements This research was supported by awards from the National Institutes of Health (the National Institute on Minority Health and Health Disparities: T37-MD001449; and the National Center for Research Resources (NCRR), the National Center for Advancing Translational Sciences (NCATS): 8UL1TR 000170-09). The NIH had no further role in study design; in the collection, analysis and interpretation of data; in the writing of the manuscript; and in the decision to submit the paper for publication. The authors thank the Enterprise Research Infrastructure & Services at Partners HealthCare for the provision of computing resources. The authors also thank Laurie Bogosian and Stacey Duey of the Research Patient Data Repository at Partners HealthCare for the in-depth support. This research was done as partial fulfillment of the requirements of a Doctor of Science degree by one of the authors (QYZ) in the Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. The authors thank Dr. Michael G. Napolitano for valuable discussions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Oates M. Suicide: the leading cause of maternal death. Br J Psychiatry. 2003;183:279–81.
2. Oates M. Perinatal psychiatric disorders: a leading cause of maternal morbidity and mortality. Br Med Bull. 2003;67:219–29.
3. Lindahl V, Pearson JL, Colpe L. Prevalence of suicidality during pregnancy and the postpartum. Arch Womens Ment Health. 2005;8:77–87.
4. Zhong Q-Y, Gelaye B, Miller M, Fricchione GL, Cai T, Johnson PA, et al. Suicidal behavior-related hospitalizations among pregnant women in the USA, 2006–2012. Arch Womens Ment Health. 2016;19:463–72.
5. Thomas KH, Davies N, Metcalfe C, Windmeijer F, Martin RM, Gunnell D. Validation of suicide and self-harm records in the clinical practice research datalink. Br J Clin Pharmacol. 2013;76:145–57.
6. Lu CY, Stewart C, Ahmed AT, Ahmedani BK, Coleman K, Copeland LA, et al. How complete are E-codes in commercial plan claims databases? Pharmacoepidemiol Drug Saf. 2014;23:218–20.
7. Anderson HD, Pace WD, Brandt E, Nielsen RD, Allen RR, Libby AM, et al. Monitoring suicidal patients in primary care using electronic health records. J Am Board Fam Med. 2015;28:65–71.
8. Rhodes AE, Links PS, Streiner DL, Dawe I, Cass D, Janes S. Do hospital E-codes consistently capture suicidal behaviour? Chronic Dis Can. 2002;23:139–45.
9. Walkup JT, Townsend L, Crystal S, Olfson M. A systematic review of validated methods for identifying suicide or suicidal ideation using administrative or claims data. Pharmacoepidemiol Drug Saf. 2012;21(Suppl 1):174–82.
10. Zhong Q-Y, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, et al. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. BMC Med Inform Decis Mak. 2018;18:30.
11. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA. 2011;306:848–55.
12. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. In: AMIA Annual Symposium Proceeding 2012, pp. 1244–53 (2012).
13. Zhong Q-Y, Gelaye B, Smoller JW, Avillach P, Cai T, Williams MA. Adverse obstetric outcomes during delivery hospitalizations complicated by suicidal behavior among US pregnant women. PLoS ONE. 2018;13:e0192943.
14. Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc. 2017;24:339–44.
15. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. Am J Psychiatry. 2017;174:154–62.
16. World Health Organization. International statistical classification of diseases and related health problems. Geneva: World Health Organization; 2004.

17. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17:507–13.
18. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267–70.
19. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
20. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22:276–82.
21. Posner K, Oquendo MA, Gould M, Stanley B, Davies M. Columbia classification algorithm of suicide assessment (C-CASA): classification of suicidal events in the FDA's pediatric suicidal risk analysis of antidepressants. Am J Psychiatry. 2007;164:1035–43.
22. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res. 2010;62:1120–7.
23. Yu S, Chakrabortty A, Liao KP, Cai T, Ananthakrishnan AN, Gainer VS, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inform Assoc. 2017;24:e143–9.
24. Ananthakrishnan AN, Cai T, Savova G, Cheng S-C, Chen P, Perez RG, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm Bowel Dis. 2013;19:1411–20.
25. Xia Z, Secor E, Chibnik LB, Bove RM, Cheng S, Chitnis T, et al. Modeling disease severity in multiple sclerosis using electronic health records. PLoS ONE. 2013;8:e78927.
26. Castro V, Shen Y, Yu S, Finan S, Pau CT, Gainer V, et al. Identification of subjects with polycystic ovary syndrome using electronic health records. Reprod Biol Endocrinol. 2015;13:116.
27. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Ann Stat. 2009;37:1733–51.
28. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Berlin: Springer; 2013.
29. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing (2014).
30. Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. 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;2016:8708434.
31. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med. 2012;42:41–50.
32. Castro VM, Dligach D, Finan S, Yu S, Can A, Abd-El-Barr M, et al. Large-scale identification of patients with cerebral aneurysms using natural language processing. Neurology. 2017;88:164–8.
33. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350:h1885.
34. Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. PLoS ONE. 2015;10:e0136651.
35. O'Connor RC, Nock MK. The psychology of suicidal behaviour. Lancet Psychiatry. 2014;1:73–85.
36. Christensen H, Cuijpers P, Reynolds CF 3rd. Changing the direction of suicide prevention research: a necessity for true population impact. JAMA Psychiatry. 2016;73:435–6.
37. McCoy TH Jr, 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;73:1064–71.
38. Gandhi SG, Gilbert WM, McElvy SS, El Kady D, Danielson B, Xing G, et al. Maternal and neonatal outcomes after attempted suicide. Obstet Gynecol. 2006;107:984–90.
39. Andover MS, Morris BW, Wren A, Bruzzese ME. The co-occurrence of non-suicidal self-injury and attempted suicide among adolescents: distinguishing risk factors and psychosocial correlates. Child Adolesc Psychiatry Ment Health. 2012;6:11.
40. Nock MK, Joiner TE Jr, Gordon KH, Lloyd-Richardson E, Prinstein MJ. Non-suicidal self-injury among adolescents: diagnostic correlates and relation to suicide attempts. Psychiatry Res. 2006;144:65–72.
41. Turecki G, Brent DA. Suicide and suicidal behaviour. Lancet. 2016;387:1227–39.
42. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, et al. Letter to the editor: suicide as a complex classification problem: machine learning and related techniques can advance suicide prediction: a reply to Roaldset (2016). Psychol Med. 2016;46:2009–10.
43. Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R. Classification algorithms for phenotype prediction in genomics and proteomics. Front Biosci. 2008;13:691–708.
44. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull. 2017;143:187–232.
45. Nock MK. Suicide: global perspectives from the WHO World Mental Health Surveys. Cambridge: Cambridge University Press; 2012.
46. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clin Psychol Sci. 2017;5:457–69.
47. Kemball RS, Gasgarth R, Johnson B, Patil M, Houry D. Unrecognized suicidal ideation in ED patients: are we missing an opportunity? Am J Emerg Med. 2008;26:701–5.
48. Committee on Obstetric Practice. The American College of Obstetricians and Gynecologists Committee Opinion no. 630. Screening for perinatal depression. Obstet Gynecol. 2015;125:1268–71.
49. Stewart C, Crawford PM, Simon GE. Changes in coding of suicide attempts or self-harm with transition From ICD-9 to ICD-10. Psychiatr Serv. 2017;68:215.
50. Oquendo MA, Baca-Garcia E. Suicidal behavior disorder as a diagnostic entity in the DSM-5 classification system: advantages outweigh limitations. World Psychiatry. 2014;13:128–30.
51. Silverman MM. The language of suicidology. Suicide Life Threat Behav. 2006;36:519–32.