King's Research Portal

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Bittar, A. L., Velupillai, S. U., Roberts, A., & Dutta, R. (2019). Text Classification to Inform Suicide Risk Assessment in Electronic Health Records. In *Studies in Health Technology and Informatics* (Vol. 264, pp. 40-44)

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 27. Oct. 2024

Text Classification to Inform Suicide Risk Assessment in Electronic Health Records

André Bittar[a], Sumithra Velupillai[a,b], Angus Roberts[a], Rina Dutta[a,c]

[a] Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK,
[b] School of Electrical Engineering and Computer Science, KTH, Stockholm, Sweden
[c] South London and Maudsley NHS Foundation Trust, London, UK

Abstract

Assessing a patient's risk of an impending suicide attempt has been hampered by limited information about dynamic factors that change rapidly in the days leading up to an attempt. The storage of patient data in electronic health records (EHRs) has facilitated population-level risk assessment studies using machine learning techniques. Until recently, most such work has used only structured EHR data, to the exclusion of the unstructured text of clinical notes. In this article, we describe our experiments on suicide risk assessment, modelling the problem as a classification task. Given the wealth of text data in mental health EHRs, we aimed to assess the impact of using this data in distinguishing periods prior to a suicide attempt from those not preceding such an attempt. We compare three different feature sets, one structured and two text-based, and show that inclusion of text features significantly improves classification accuracy in suicide risk assessment.

Keywords: *Suicide, Risk Assessment; Natural Language Processing*

Introduction

Suicide is a serious public health problem, with almost one million people ending their lives worldwide each year [1]. In the United Kingdom, although suicide rates have dropped slightly since the early 1980s, in 2017 the Office for National Statistics nevertheless registered 5,821 suicides [2]. More than a quarter are in receipt of mental health services at the time of death [3], yet suicide risk remains immensely difficult for clinicians to assess, given the wide range of contributory factors, with the majority (88%) judged to be at 'low or no immediate risk' of suicide by clinicians at their final service contact. Current clinical methods for assessing when someone is at risk of a suicide attempt have been reported to be little better than chance [4]. New approaches to individualised risk assessment that integrate data from different sources are needed. With the availability of population-level patient data in the form of electronic health records (EHRs), novel methods for suicide risk assessment based on data mining and machine learning have been explored in recent years [5]. Indeed, machine learning models can capture complex associations between variables, making them particularly well-suited to the task of predictive analysis.

Different approaches to modelling the problem of suicide risk assessment have been explored. Several recent studies have focused on developing models that detect the suicide risk of individual patients in large historical population samples. For instance, Barak-Corren et al. (2017) trained a Naïve Bayes classifier on a cohort of more than 1.7 million patients (16,588 of whom had recorded suicidal behaviour) using a rich set of structured EHR features, including demographic, diagnostic, procedure and medication data [6]. Simon et al. (2018) combined structured EHR data and standard health questionnaire responses in a logistic regression model to predict suicide attempt and death in a cohort of nearly 3 million patients [7]. Similar work has also been carried out using state-of-the-art neural networks. For example, Bhat and Goldman-Mellor (2017) trained neural network classifiers on historical structured EHR data to detect the presence of a suicide attempt by individual patients in a given year [8].

A common factor in the aforementioned work and, more generally in the majority of research using machine learning for suicide risk assessment, is that it has relied exclusively on features derived from the structured fields of EHRs. However, much valuable data about patients is stored in EHRs as unstructured text [9]. To date, relatively little research on suicide risk has been done on mining this rich data source for features, although momentum is building. For example, Metzger et al. (2017) tested a series of machine learning classifiers to determine the prevalence of suicide-related emergency department admissions in a French hospital [10]. They used structured EHR data, as well as features extracted from clinical notes. Ben-Ari and Hammond (2015) performed a text search to identify Gulf War veterans who have made a suicide attempt [11]. They then used textual and structured features with a Random Forest classifier to predict first suicide attempts by patients in this cohort within a given year. McCoy et al. (2016) also used text-based features along with structured EHR data to gauge suicide risk after discharge [12]. They used an "off-the-shelf" Natural Language Processing (NLP) tool to extract the polarity of valence-conveying words (positive or negative) within the records and used this in regression models, finding that positive valence words were correlated with reduced suicide risk. Downs et al. (2017) used NLP to identify suicide-related mentions in a cohort of adolescents with Autism Spectrum Disorder, including a previously-developed negation detection module [13, 14]. Finally, Fernandes et al. (2018) used NLP to identify and classify mentions of suicidal ideation and suicide attempts in mental health records [15].

Methods

Classification for Suicide Risk Assessment

Most epidemiological case-control studies have used data spanning much wider time periods, typically years, and the risk factors have either been static (e.g. male gender, family history of psychiatric disorder) or lifetime ever variables (e.g. previous attempted suicide, any misuse of alcohol or drugs) [16]. In this study, we take a rather different approach, exploring whether it is possible to predict suicide attempts by building classification models based on structured and textual data from the 30-day period leading up to the event, when there is an opportunity to intervene as it can be a time of crisis. To the best of our knowledge, our approach is novel in using this critical time period. We tested a machine learning classifier to distinguish the 30-day periods prior to a hospital admission linked to a suicide attempt (hereafter referred to as 'suicidal window') and similar periods not preceding a suicide attempt from age- and sex-matched control patient records ('non-suicidal windows') extracted from a large database of EHRs. We used a linear-kernel Support Vector Machine (SVM) [17] classifier given this algorithm's well established performance in dealing with the high-dimensionality of text data [18]. We modelled the task using supervised binary classification. A further novel aspect of our work lies in our assessment of the impact of three different sets of features, namely, structured fields from the EHRs, a rich set of binary features derived from the clinical notes, and a bag-of-words representation of the full text of these same documents.

CRIS Clinical Cohort

We studied the de-identified EHRs of over 250,000 patients from the South London and Maudsley (SLaM) NHS Foundation Trust using the Clinical Record Interactive Search (CRIS) computer system comprising both structured data and over 3.5 million text documents [19]. Data from CRIS has been linked with the UK Hospital Episode Statistics (HES) data for Admitted Patient Care within a secure 'safe haven', and it is through this linkage that admission information was extracted. The documents in CRIS have been substantially enhanced, in particular, through the application of NLP (e.g. to identify symptoms) [20].

Our dataset was derived from the EHRs of 17,640 patients. It consisted of 21,175 suicide-related (case) and non-suicide-related (control) admissions, sampled according to a 1:4 case-control ratio. Cases were defined as any admission (acute physical or specialist mental health) where there was a suicide attempt (indicated by the presence of any of the following ICD codes: X6*, X7*, X80-4*, Y1*, Y2*, Y30-4*, Y87*) with the admission lasting at least 24 hours (starting and ending on different dates). Only admissions with a start date after the 1st of April 2006 and an end date before or including 31st March 2017 were considered. Of these case admissions only those which had at least 1 document in the 30 days prior and including the date of the hospitalised suicide attempt were retained. We also removed admissions with empty documents[1]. This left a total of 4,235 suicide-related admissions in the final dataset. Each control was matched by sex, had to be alive at the admission start date of the case, and were grouped into the same age group as cases (5 year age bands < 16, 16-19, 20-24 to 80-84, 85+ years). Each control also had at least one document in the 30 days prior to and including the date of their matched case's hospitalised suicide attempt. The total number of controls was 16,940. The controls were chosen to be as representative as possible of the population from which the cases were drawn and the ratio was based on the epidemiological principle that little statistical power is gained by further increasing the number of controls beyond approximately 4 per case [21]. Key descriptive characteristics of the dataset are shown in Table 1.

Features for Classification

Features examined included standard sociodemographic and clinical descriptors selected by a clinical academic psychiatrist (RD) *a priori* (e.g. ethnic group, marital status, employment status), as well as those shown in prior work to be associated with suicide attempts (e.g. past and current substance abuse) [22]. We used 14 (categorical) features from structured fields of the EHRs, including total document counts for the 30 days prior to but excluding the day of admission, based on the hypothesis that the volume of documentation would increase prior to a risk event such as a suicide attempt due to greater service use. We also included features derived from NLP applications routinely run on CRIS. Most of these applications are built using GATE, an open source NLP toolkit [20, 23]. All of these applications combine both rule-based pattern matching algorithms and supervised machine learning models, and some detect contextual information, such as negation or family history. We derived 68 binary features from these, one per application, each feature indicating the presence or absence of at least one match by the application on at least one document in the window, excluding the admission date. Applications included the detection of positive mentions of suicide attempt by the patient, mentions of the patient having disturbed sleep, and feelings of hopelessness and paranoia, to cite a few.

Finally, we also included as features the concatenated text from all documents in each window, excluding the admission date, represented as a TFIDF (term frequency-inverse document frequency) vector or "bag-of-words". TFIDF is a statistical weighting that reflects how important a particular term is in a given document in a collection, adjusted for the fact that some words occur more frequently than others [24]. When calculating word vectors, we applied the L2-Norm [25] to scale the vectors to unit length. This compensates for discrepancies in document length, such as those in our dataset. The total number of unique words in our document set was 201,538, making it a high-dimensional feature space.

	Cases	**Controls**
Patients **Female** **Male**	2,913 (16.5%) 1,730 (59.4%) 1,183 (41.6%)	14,727 (83.5%) 8,971 (60.9%) 5,756 (39.1%)
Admissions **Female** **Male**	4,235 (20%) 2,598 (61.3%) 1,637 (38.7%)	16,940 (80%) 10,392 (61.3%) 6,548 (38.7%)
Mean age (SD) years	34.4 (15.3)	34.4 (15.4)
EHR features for 30-day pre-admission windows		
Mean tokens (SD)	3455.5 (5732.4)	1344.5 (3179.9)
Total tokens	14,634,223	22,775,227
Mean docs (SD)	16.9 (31.4)	7.5 (18.1)
Total docs	71,404	127,047

Table 1: Characteristics of the dataset for suicide attempt related cases and non-suicide-related controls. Note that each window for EHR features is a 30-day period prior to the case admission date (inclusive).

Henceforth, we refer to the set of structured data as STRUCT, the text-based features derived from NLP applications routinely run on CRIS as GATE, and the bag-of-words features as TFIDF. For STRUCT and GATE, we encoded all features either as integers (e.g. age, marital status, employment status) or binary (0 or 1) values (e.g. sex, presence of the keyword *depression*). We list details of all 82 features in an online annex[2].

---

[1] Text from scanned documents is not always available.

Results

Experimental Setup

We randomly split the data into a training set (80%) and a test set (20%), ensuring the distribution of each class was the same across both sets (i.e. 1 case to 4 controls). The test data was held out until the final run in order to reduce the risk of overfitting and to provide a realistic estimation of performance on unseen data. We scaled all features to have zero-mean and unit-variance to ensure a balanced contribution of all features. We implemented the classifier and prepared data using the Scikit-learn (version 0.20.0) machine learning library for Python [26]. Our first step was to estimate the optimal parameters for the classifier (model tuning). We did this using grid search and ten-fold cross-validation on the training data, with F1-score as the evaluation metric[3]. F1-score is the harmonic mean of precision (positive predictive value) and recall (sensitivity), a metric often used in information retrieval and NLP [27]. For each tuning instance, we varied the feature set so as to tune the models to each representation of the data. This resulted in 7 different feature combinations: STRUCT, GATE, TFIDF, STRUCT+GATE, STRUCT+TFIDF, GATE+TFIDF, STRUCT+GATE+TFIDF. A flowchart of the general architecture is provided in the online annex. To gauge whether the differences in pairwise comparisons of feature sets were statistically significant, we used McNemar's test [28] (α=0.05) for classification disagreements on the training dataset.

Finally, we examined the text features that were most significant in distinguishing the two classes. For the final run on the held-out test data, we calculated the F1-score for the suicidal and non-suicidal windows and the mean of both.

Intensity of Documentation

As shown in Table 1, the mean number of documents for cases during the 30-day window is 19.9 (SD=34.0) and a lower 8.3 (SD=19.4) for controls. This divergence in mean document counts for each class supports the decision to include the number of documents in a window as a feature for classification. See Feature Importance for further comments.

Classification

In this section, we present results obtained on the held-out test data for each of the different feature sets. We report performance in terms of precision (P), recall (R) and F1-score (F). Where we make a comparison of results obtained on two feature sets, we also report the *p*-value of McNemar's pairwise test between them. Although we calculated figures for each of the two classes (suicidal windows and non-suicidal windows) separately, we are only interested in assessing the identification of suicidal windows. The correct classification of non-suicidal windows was relatively much simpler given the prevalence of this class in the dataset (mean F1-score for this class was 0.86, SD=0.08). Therefore, the figures we report are for suicidal windows only.

The classifier's performance using only structured features (STRUCT feature set) was relatively low (P=0.26, R=0.59, F=0.36). This increased with the addition of features extracted by GATE (STRUCT+GATE feature set) (P=0.49, R=0.58, F=0.53, p<0.001). However, best performance was obtained with the addition of the bag-of-words features (STRUCT+GATE+TFIDF feature set) (P=0.61, R=0.63, F=0.62, p<0.001), showing a good balance between precision and recall. Interestingly, bag-of-word features alone (TFIDF) provided the next best results (P=0.59, R=0.61, F=0.60, p=0.0042), only slightly lower than the combination of all feature sets, suggesting the importance of the bag-of-words representation. The use of GATE features on their own performed less well (P=0.50, R=0.57, F=0.54, p<0.001). This indicates that the TFIDF features captured a signal in the data that the targeted NLP applications did not. Using the combination of GATE features and bag-of-words (GATE+TFIDF), recall was heavily penalised, resulting in a significant reduction to the model's F1-score, despite increased precision (P=0.63, R=0.38, F=0.48, p<0.001). The combination of structured and bag-of-words features (STRUCT+TFIDF) provided a significant improvement over structured features alone (P=0.62, R=0.37, F=0.46, p<0.001), but also with a sharp increase in precision to the detriment of recall. Figure 1 provides a visual comparison of these results while Figure 2 shows a summary of all pairwise McNemar's tests across all feature sets.

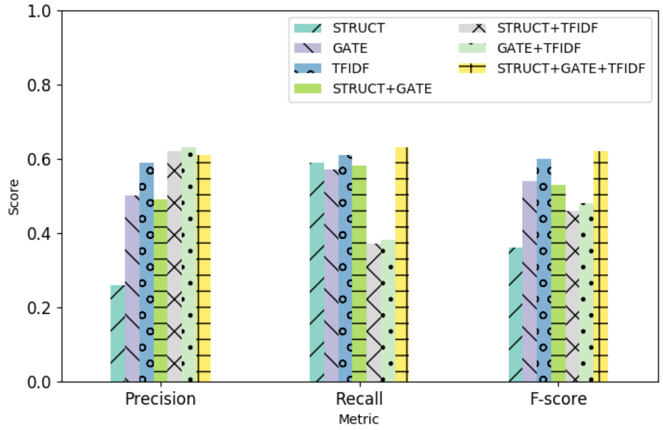

Figure 1 – Performance on the positive class (suicide attempt) in terms of precision (P), recall (R) and F1-score (F) on the test set for all feature sets.

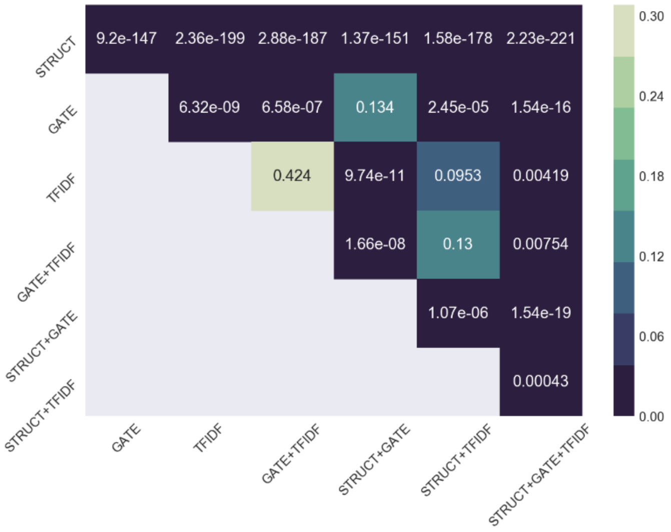

Figure 2 – Full results of pairwise McNemar's tests across feature sets on the training data. Cells show p-values (α=0.05), dark cells are more statistically significant, while lighter cells indicate the converse.

---

[2] https://github.com/KCL-Health-NLP/medinfo2019-sa-risk/

[3] All tuning parameter ranges and the final tuned configurations are provided in the online annex.

Feature Importance

During training, the (linear) SVM calculates a maximum-margin hyperplane to separate (as much as possible) the two classes in the data. The feature weights representing the coordinates of the vector orthogonal to the hyperplane are stored and their direction indicates the predicted class. We compared the values of these weights for each of the feature sets. The number of documents within the window is among the top-ranking features in the STRUCT feature set, indicating the importance of this feature in discriminating between classes. The most discriminating feature in the GATE feature set was the presence of a positive mention of a suicide attempt by the patient. These were also the top features in the STRUCT+GATE feature set. The top two features for the TFIDF feature set were the words *overdose* and *self-harm*, while other words, such as the adjective *suicidal*, and the medication names *paracetamol* and *zopiclone*, ranked among the top 15 discriminating features. More diverse keywords, as well as other features, ranked highly when TFIDF was combined with either STRUCT or GATE, or both. Given these results, we quantified and visualised the relative frequency of these terms within the texts of the suicidal and non-suicidal windows. To offset the imbalance in the dataset (differing mean number of tokens between the two groups) we calculated per-token frequencies for each day and divided term counts for control by 4 (to adjust for the 1:4 case-control ratio). Figure 3 shows a comparison for the term *overdose*. Mentions of this term are approximately four to six times more frequent for cases than for controls across the 30-day window. Plots for the other aforementioned terms showed the same trend. These preliminary findings on the TFIDF feature set suggest that significant mentions of suicide-related behaviour (overdose, self-harm, etc.) have been recorded in the suicidal windows, but not in the non-suicidal windows. Furthermore, despite the relatively low per-token frequency of the significant terms, the TFIDF weighting allowed for these to be picked up as important features by the classifier. A more in-depth and systematic examination of the occurrences of these terms would establish whether they represent independent suicide attempts that are not recorded using ICD codes in the HES data linked to the EHRs. Nevertheless, these preliminary results support the use of such word features in classification to inform suicide attempt risk assessment.

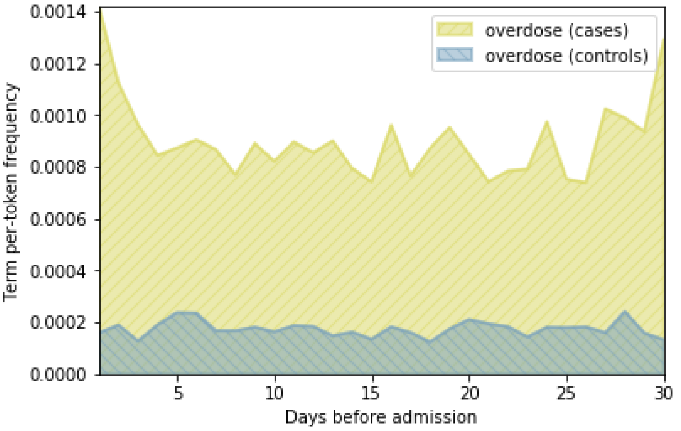

Figure 3 – Normalised relative per-token frequency of the term “overdose” for suicidal (case) and non-suicidal (control) windows.

Discussion

Using a case-control study design, we carried out an exploratory experiment on binary text classification to assess how this might help to inform suicide risk assessment. We evaluated the relative contributions of three different types of features, namely structured data, binary features derived from NLP applications routinely run on CRIS, and a TFIDF bag-of-words representation. Our results show that the use of text features significantly improves classification results, and the combination of structured and text-based features provided the best performance. An examination of the top textual features used in classification revealed the importance of certain terms for discriminating suicidal and non-suicidal windows.

The variety of ways in which the suicide risk assessment task has been modelled previously, including differences in data sets, algorithms and features, makes meaningful comparison of results between studies difficult and such was not the aim of this work.

Despite interesting results, clearly this work does have limitations. Firstly, the 14 structured features were selected to represent only a small sample of all available structured data. This means that although these features are clinically relevant for assessing suicide risk, certain features with less complete data were not tested. A broader selection of structured data may have led to better results with the STRUCT feature set. Another drawback lies in the two representations we used for the textual features. The binary (GATE) feature set is unable to account for the relative frequencies of identified terms. Thus, a single match by an application has the same "weight" as multiple matches in the document set. Although the TFIDF bag-of-words representation addresses this weakness, it does not capture the order and combination of words (e.g. multi-word expressions such as *suicidal ideation*), or phenomena such as negation (e.g. *no suicidal ideation*). Furthermore, our bag-of-words approach did not enable us to distinguish terms relating to the patient from those concerning other people (e.g. *family history of suicide, father took an overdose*). Whilst this was accounted for to some extent in the GATE features, more nuanced and targeted NLP could improve performance.

Conclusions

We have shown that the inclusion of text features in classification to inform suicide risk assessment using EHR data provides a statistically significant increase in performance over a dataset containing only structured data. Including the text allows access to word features that appear to be potential markers of impending suicide risk (*overdose, self-harm, suicidal*) that are also clinically plausible. Strikingly, the intensity of documentation within the 30-day period prior to an event may also be a significant factor in determining times of increased risk.

Acknowledgements

RD is funded by a Clinician Scientist Fellowship (research project e-HOST-IT) from the Health Foundation in partnership with the Academy of Medical Sciences which also funds AB. SV is supported by the Swedish Research Council (2015-00359), Marie Skłodowska Curie Actions, Cofund, Project INCA 600398. AR is funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London.

References

[1] World Health Organization, Mental health: suicide prevention, (2018).

http://www.who.int/mental_health/suicide-prevention/en/ (accessed November 2, 2018).

[2] Office for National Statistics, Suicides in the UK: 2017 registrations, (2018). https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/suicidesintheunitedkingdom/2017registrations (accessed September 25, 2018).

[3] The assessment of clinical risk in mental health services. National Confidential Inquiry into Suicide and Safety in Mental Health (NCISH), The University of Manchester, 2018.

[4] J.C. Franklin, J.D. Ribeiro, K.R. Fox, K.H. Bentley, E.M. Kleiman, X. Huang, K.M. Musacchio, A.C. Jaroszewski, B.P. Chang, and M.K. Nock, Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research., *Psychol. Bull.* **143** (2017) 187–232.

[5] E. Baca-García, I. Basurte-Villamor, J.M. Leiva-Murillo, M. de Prado-Cumplido, R. Santiago-Mozos, A. Artés-Rodríguez, and J. de Leon, Using Data Mining to Explore Complex Clinical Decisions: A Study of Hospitalization After a Suicide Attempt, *J Clin Psychiatry*. (2006) 9.

[6] Y. Barak-Corren, V.M. Castro, S. Javitt, A.G. Hoffnagle, Y. Dai, R.H. Perlis, M.K. Nock, J.W. Smoller, and B.Y. Reis, Predicting Suicidal Behavior From Longitudinal Electronic Health Records, *Am. J. Psychiatry*. **174** (2017) 154–162.

[7] G.E. Simon, E. Johnson, J.M. Lawrence, R.C. Rossom, B. Ahmedani, F.L. Lynch, A. Beck, B. Waltzfelder, R. Ziebell, R.B. Penfold, and S.M. Shortreed, Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records, *Am. J. Psychiatry*. **175** (2018) 951–960.

[8] H.S. Bhat, and S.J. Goldman-Mellor, Predicting Adolescent Suicide Attempts with Neural Networks, *ArXiv171110057 Cs Stat*. (2017). http://arxiv.org/abs/1711.10057 (accessed October 25, 2018).

[9] R. Stewart, and K. Davis, ‘Big data’ in mental health research: current status and emerging possibilities, *Soc. Psychiatry Psychiatr. Epidemiol.* **51** (2016) 1055–1072.

[10] M.-H. Metzger, N. Tvardik, Q. Gicquel, C. Bouvry, E. Poulet, and V. Potinet-Pagliaroli, Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study: text-mining and epidemiology of suicide attempts, *Int. J. Methods Psychiatr. Res.* **26** (2017).

[11] A. Ben-Ari, and K. Hammond, Text Mining the EMR for Modeling and Predicting Suicidal Behavior among US Veterans of the 1991 Persian Gulf War, in: 2015 48th Hawaii Int. Conf. Syst. Sci., IEEE, HI, USA, 2015: pp. 3168–3175.

[12] T.H. McCoy, V.M. Castro, A.M. Roberson, L.A. Snapper, and R.H. Perlis, Improving Prediction of Suicide and Accidental Death After Discharge From General Hospitals With Natural Language Processing, *JAMA Psychiatry*. **73** (2016) 1064.

[13] J. Downs, S. Velupillai, G. Gkotsis, R. Holden, M. Kikoler, H. Dean, A. Fernandes, and R. Dutta, Detection of Suicidality in Adolescents with Autism Spectrum Disorders: Developing a Natural Language Processing Approach for Use in Electronic Health Records, *Proc. AMIA Annu. Symp.* (2017) 641–649.

[14] G. Gkotsis, S. Velupillai, A. Oellrich, H. Dean, M. Liakata, and R. Dutta, Don’t Let Notes Be Misunderstood: A Negation Detection Method for Assessing Risk of Suicide in Mental Health Records, in: Proc. Third Workshop Comput. Lingusitics Clin. Psychol., Association for Computational Linguistics, San Diego, CA, USA, 2016: pp. 95–105.

[15] A.C. Fernandes, R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, and D. Chandran, Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing, *Sci. Rep.* **8** (2018).

[16] K. Hawton, C. Casañas i Comabella, C. Haw, and K. Saunders, Risk factors for suicide in individuals with depression: A systematic review, *J. Affect. Disord.* **147** (2013) 17–28.

[17] C. Cortes, and V. Vapnik, Support-vector networks, *Mach. Learn.* **20** (1995) 273–297.

[18] T. Joachims, Text categorization with Support Vector Machines: Learning with many relevant features, in: C. Nédellec, and C. Rouveirol (Eds.), Mach. Learn. ECML-98, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998: pp. 137–142.

[19] G. Perera, M. Broadbent, F. Callard, C.-K. Chang, J. Downs, R. Dutta, A. Fernandes, R.D. Hayes, M. Henderson, R. Jackson, A. Jewell, G. Kadra, R. Little, M. Pritchard, H. Shetty, A. Tulloch, and R. Stewart, Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource, *BMJ Open*. **6** (2016).

[20] R.G. Jackson, R. Patel, N. Jayatilleke, A. Kolliakou, M. Ball, G. Gorrell, A. Roberts, R.J. Dobson, and R. Stewart, Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project, *BMJ Open*. **7** (2017).

[21] D.A. Grimes, and K.F. Schulz, Compared to what? Finding controls for case-control studies, *The Lancet*. **365** (2005) 1429–1433.

[22] M.A. Ilgen, M.L. Burnette, K.R. Conner, E. Czyz, R. Murray, and S. Chermack, The association between violence and lifetime suicidal thoughts and behaviors in individuals treated for substance use disorders, *Addict. Behav.* **35** (2010) 111–115.

[23] R.G. Jackson, M. Ball, R.P. BMBCh, R.D. Hayes, R.J.B. Dobson, and R. Stewart, TextHunter – A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research, (2014) 10.

[24] A. Rajaraman, and J.D. Ullman, eds., Data Mining, in: Min. Massive Datasets, Cambridge University Press, Cambridge, 2011: pp. 1–17.

[25] Roger A. Horn, and Charles R. Johnson, Norms for Vectors and Matrices, in: Matrix Anal., Cambridge University Press, Cambridge, England, 1990.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, Scikit-learn: Machine Learning in Python, *Mach. Learn. PYTHON*. (n.d.) 6.

[27] N. Chinchor, MUC-4 Evaluation Metrics, in: Proc. 4th Conf. Message Underst. MUC4 92, Stroudsburg, PA, USA, 1992: p. 8.

[28] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*. **12** (1947) 153–157.

Address for correspondence

André Bittar, andre.bittar@kcl.ac.uk