Towards Automatic Suicide Risk Assessment

Abstract

Background: Suicide has been considered as an important public health issue for years, and is one of the main causes of death worldwide. Despite prevention strategies being applied, the rate of suicide has not changed substantially over the past decades. Suicide risk has proven extremely difficult to assess for medical specialists, and traditional methodologies deployed have been ineffective. Advances in machine learning make it possible to attempt to predict suicide with the analysis of relevant data aiming to inform clinical practice. Aims: (a) test our artificial intelligence based, referral-centric methodology in the context of the NHS, (b) determine whether statistically relevant results can be derived from data related to previous suicides, and (c) develop ideas for various exploitation strategies. Method: The analysis used data of patients who died by suicide in the period 2013-2016 including both structured data and free-text medical notes, necessitating the deployment of state of the art machine learning and text mining methods. Results and Conclusions: The results of this pilot study indicate that machine learning shows promise to predict within a specified period which people are most at risk of taking their own life at the time of referral to a mental health service. Keywords: automated machine learning, suicide prevention, risk assessment, clinical data, text mining

1. Introduction

Suicide has been considered as an important public health issue for a very long time (Oliven 1954), and presently, its scale in avoidable loss of life was described by the UK House of Commons as unacceptable (House of Commons Health Committee 2017). For such an important public health issue, suicide prevention strategies were launched worldwide (World Health Organisation 2014) but despite increasing efforts to reduce suicides through improved assessment and treatment, awareness campaigns and support services, the rate of suicide has not changed substantially over the past decades. During this period, medical knowledge and healthcare technologies developed rapidly and huge progress was made in combating other leading causes of death such as cancer, cardiovascular diseases and HIV.

Suicide risk has proven extremely difficult to assess for medical specialists as several variables are involved in its pathway. This led to the recommendation than in primary care, the rollout of suicide prevention initiatives is not recommended (O'Connor, Gaynes et al. 2013, LeFevre 2014, Milner, Witt et al. 2017) whilst clinical instruments already used to predict suicide were found not be clinically useful when classifying “high risk” individuals (Chan, Bhatti et al. 2016, Large, Kaneson et al. 2016, Carter, Milner et al. 2017), even when suicidal ideation and suicidal behaviours are used as predictors of suicide and other suicidal behaviours (Chapman, Mullin et al. 2015, Hubers, Moaddine et al. 2016, Ribeiro, Franklin et al. 2016).

These findings led to the realisation that current approaches to suicide risk assessment do not work in a satisfactory way at present, and that new paradigms are required (Hawgood and De Leo, 2016).
for suicide prediction and prevention. In addition, the need for clear structured approaches to risk assessment that are based on research evidence is now prominent (Department of Health 2012).

A particularly interesting approach in this context is the ability to automatically predict suicide risk by analyzing person-related data. This data could be clinical records, social care data, psychological assessments or social media entries, to name a few. As (Vahabzadeh, Sahin et al. 2016) explains, there is scope to assess the general impact of risk factors and their interaction and to develop a risk profile for an individual person.

This approach has been deployed in specific aspects of data-driven suicide prevention, for example in the US army (Kessler, Warner et al. 2015). This work developed a suicide risk assessment algorithm predicting the risk of suicide within 12 months from discharge, with good predictive power. Other works sought to predict the risk of suicide for US army veterans by analyzing clinical text notes using a learning algorithm on a genetic programming framework (Poulin, Shiner et al. 2014) whilst a thorough, data-driven, retrospective analysis of suicides in the Northwest of England (Saini, While et al. 2014) compared suicide risk assessment in primary and secondary care.

The pilot study reported in this paper was designed to (a) test an AI-based, referral-centric methodology and research design for automated risk assessment in the context of the NHS, (b) determine whether statistically relevant results can be derived from data related to previous suicides, and (c) develop ideas for various exploitation strategies.

2. Methodology

2.1. Data Sample

The sample was provided by a specialist NHS mental health provider (South West Yorkshire Partnership NHS Foundation Trust-SWYPFT) and consisted of the complete records of people who died by suicide in the period 2013–2016. Overall there were 130 such records. The data contained different type of information: demographics, referrals, appointments, progress notes, comprehensive assessments and inpatient stays. All data types except demographic data contained several events defining the clinical trajectory of the subjects. The information is represented either as free text or as semi-structured fields.

Demographics information includes date of birth, gender, marital status, ethnicity, religion, postcode, date and age of death. All this information is in (semi-)structured form. Table 1 summarizes the descriptive statistics of the demographics information.

2.2. Analysis Methods

2.2.1. Predictive Analysis Set-up

We devised a referral-centered analysis, with the goal of predicting whether a person will proceed to take their own life after their last referral to the Service. The objective was to realize a predictive model that assesses referrals according to the patient’s risk for dying by suicide in the next \( t \) months (\( t \) was set to 3 and 6 months; these were choices proposed to us by medical experts both to allow for reasonable time for interventions and medication effect to take place and to allow us to assess how the predictive performance changes as the time moves closer to a suicide event). Referrals taken at most \( t \) months before suicide formed the positive class, and the rest formed the negative class.

The referrals carry their own structured information, such as urgency, length of the episode, reason for discharge, etc. We paired this data with the demographic characteristics of the patient of the corresponding referral as well as information from the last free-text medical note taken immediately

---

1The Trust covers Barnsley, Calderdale, Kirklees and Wakefield in the North of England, and includes a number of towns as well as rural areas.
Table 1. Descriptive statistics of demographics information of data sample (suicide cases 2013-2016 used in this study). The “Age at death” is the only continuous attribute and is reported as average (standard deviation) in the whole population. Other variables are categorical and are reported as number (percentage) of subjects in each category.

<table>
<thead>
<tr>
<th>Age at death</th>
<th>average (std)</th>
<th>Gender</th>
<th># subjects (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole population</td>
<td>44.91 (15.93)</td>
<td>Male</td>
<td>86 (66%)</td>
</tr>
<tr>
<td>Men</td>
<td>44.61 (16.14)</td>
<td>Female</td>
<td>44 (34%)</td>
</tr>
<tr>
<td>Women</td>
<td>45.5 (15.66)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Religion</th>
<th># subjects (%)</th>
<th>Marital Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christian</td>
<td>28 (21.5%)</td>
<td>Divorced</td>
</tr>
<tr>
<td>Do Not Wish to Disclose</td>
<td>15 (11.5%)</td>
<td>Married</td>
</tr>
<tr>
<td>No Religion</td>
<td>18 (14%)</td>
<td>Not Disclosed</td>
</tr>
<tr>
<td>Not Known</td>
<td>56 (43%)</td>
<td>Not Known</td>
</tr>
<tr>
<td>Other</td>
<td>6 (5%)</td>
<td>Separated</td>
</tr>
<tr>
<td>Roman Catholic</td>
<td>7 (5%)</td>
<td>Single</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Widowed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Post Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Other Background</td>
<td>BD (Bradford)</td>
</tr>
<tr>
<td>Any Other White Background</td>
<td>HD (Huddersfield)</td>
</tr>
<tr>
<td>Not Known</td>
<td>HX (Halifax)</td>
</tr>
<tr>
<td>Patient Refused to Disclose</td>
<td>OL (Oldham)</td>
</tr>
<tr>
<td>White – British</td>
<td>S (Sheffield)</td>
</tr>
<tr>
<td></td>
<td>WF (Wakefield)</td>
</tr>
</tbody>
</table>

before the referral. Finally, we constructed new variables, e.g. for representing the number of clinical appointments that the patient had scheduled in the last $X$ months before a referral, where $X = 1, ..., 12$. In total, there were 828 referrals (a.k.a. samples) and 7,711 variables, 7,686 of which derived from preprocessing textual information (see Section 2.2.3 for further details).

On this data we defined a number of binary classification tasks, by varying several characteristics of the analysis. Parameters included the time threshold $t = 3$ or $6$, inclusion of textual features or not, and interpretable models or unrestricted. The latter choice is whether to run all available learning algorithms trying to optimize predictive performance or to restrict the analysis only to those learning algorithms that result in humanly interpretable models. Producing human-understandable models is of paramount importance for gaining insight into the mechanisms leading to an increase in the risk of dying by suicide; however, restricting the analysis to only a subset of possible learning algorithms may result in models with decreased predictive power.

### 2.2.2 Predictive Analysis Pipeline

JAD Bio (Borboudakis, Stergiannakos et al. 2017, Orfanoudaki, Markaki et al. 2017) is an automated multivariate statistical analysis pipeline comprising of a complete set/sequence of learning steps that lead to the production of the final predictive model (see Figure 1 for a schematic overview). In particular, JAD Bio performs (a) preprocessing of the data, (b) feature selection, (c) training of predictive models, (d) automated selection of the best configuration (i.e. a combination of feature...
selection and learning algorithms as well as specific values for their hyper-parameters) with which to construct the final model otherwise known as tuning, (e) construction of the final predictive model using the best configuration and all available data, and finally, (f) accurate performance estimation of the final model as well as its 95% confidence intervals.

Preprocessing methods change the values of the input variables, where necessary. In this particular analysis, the preprocessing methods employed included imputation of missing values, binarization of categorical variables, and standardization of continuous variables.
Feature selection is the process of identifying the most salient features for learning. The feature selection algorithm used for this analysis is the Statistical Equivalent Signatures (SES) (Lagani, Athineou et al. 2016). SES tries to identify as many as possible minimal sets of features that provide optimal classification accuracy, i.e., it reports multiple solutions to the feature selection problem.

JAD employs state-of-the-art supervised machine learning algorithms and trains a variety of multivariate advanced and basic predictive models. For binary classification problems, it uses the following learning algorithms: Support Vector Machines (SVMs) (Boser, Guyon et al. 1992) with linear, polynomial, and Gaussian kernels, Random Forests (RFs) (Breiman 2001), Decision Trees (DT) (Breiman, Friedman et al. 1984), and Ridge Logistic Regression (RLR) (Hoerl and Kennard 1970).

The tool automatically determines the set of configurations (on the basis of the statistical properties of the dataset, such as the number of training samples and the number of variables) to try. The best configuration, from which the final model will be constructed, is then identified using stratified, repeated 10-fold cross-validation (i.e. stratified cross-validation is repeated multiple times with different partitions of the data to folds). However, the cross-validated estimate of performance is optimistically biased (Varma and Simon 2006). JAD estimates the bias using a bootstrap-based method called BBC-CV (Tsamardinos, Greasidou et al. 2018), and removes it to return the final performance estimate.

The final predictive model that is returned is trained with the best-found configuration on all available data.

### 2.2.3 Text Mining

The clinical notes included unstructured free-text information that first had to be analyzed and converted into structured, measurable data, suitable for a machine learning analysis. To do so, all irrelevant information, such as html tags and punctuation, was removed from the text. Human names and stop-words (i.e. most common words in a language) were also identified and removed with the use of the Natural Language Toolkit (NLTK) (Bird, Klein et al. 2009). Next, all text was converted to lower case and the Porter 2 stemming algorithm from NLTK was applied to the resulting words. Stemming is the process of reducing inflectional forms and sometimes derivationally related forms of a word to a base form. Its application makes the data denser (reducing the variable space), and thus machine learning algorithms are expected to operate more efficiently. However, working only with the stem of a word comes at the risk of losing some information. Finally, the processed text was converted into a dataset of single words (bag-of-words) using the scikit-learn software (Pedregosa, Varoquaux et al. 2011). With bag-of-words modelling, text is represented as the multiset of its words, disregarding linguistic structure and structural markup, and only keeping word frequency. We did not take into consideration typographical errors and abbreviations of clinics and hospitals.

### 3 Results

For each of the predictive analyses we obtained a set of predictive variables, a predictive model, and a conservative estimate of the model’s generalization performance as well as 95% confidence interval for it. We report performances as Area Under the Receiver Operating Characteristic curve (AUC) (Huang and Ling 2005).

Table 2 shows the results for the predictive analyses for the time threshold $t$ being equal to three. Each row corresponds to a separate analysis whose characteristics (type of variables included in the analysis and interpretability of the returned model) are specified in the first two columns. The table also reports the number of samples $N$ and the number of variables $M$ involved in each of the analysis, as well as the best-performing configuration $C$ from which the final predictive model was produced and the number of variables it includes (selected by SES). Finally, the last column shows the estimated performance of the final predictive model and its 95% confidence intervals.
Table 2. Results for the predictive analyses for $t = 3$. $N$ and $M$ denote the number of samples and variables in the dataset, respectively. $C$ is the best-performing configuration from which the final model is constructed. The best prediction performance is highlighted in bold.

<table>
<thead>
<tr>
<th>analysis characteristics</th>
<th>$N$</th>
<th>$M$</th>
<th>$C$</th>
<th>#selected variables</th>
<th>AUC (CI 95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>structured variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>complete</td>
<td>828</td>
<td>25</td>
<td>SES, RLR</td>
<td>7</td>
<td>0.652 (0.589, 0.709)</td>
</tr>
<tr>
<td>human-interpretable</td>
<td>828</td>
<td>25</td>
<td>SES, RLR</td>
<td>7</td>
<td>0.662 (0.607, 0.719)</td>
</tr>
<tr>
<td><strong>structured and textual variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>complete</td>
<td>828</td>
<td>7711</td>
<td>SES, RFs</td>
<td>262</td>
<td><strong>0.705 (0.646, 0.760)</strong></td>
</tr>
<tr>
<td>human-interpretable</td>
<td>828</td>
<td>7711</td>
<td>SES, RLR</td>
<td>25</td>
<td>0.605 (0.545, 0.662)</td>
</tr>
</tbody>
</table>

The best overall performing model is obtained where (a) both structured and textual variable are included, (b) there is no restriction on the tested configurations, and (c) the time-point $t$ is equal to three months. The AUC achieved in this case is 0.705. The 95% CI of the AUC does not include the value 0.5 (CI = [0.646, 0.760]), thus the results are deemed statistically significant at the standard significance level of 5%.

For the complete (unrestricted) analysis we notice that the predictive models derived from the datasets that include both the structured and the textual variables perform better than those derived from datasets that include only the structured variables by an average of 0.05 AUC points. The difference in AUC was found to be significant, indicating that the free-text medical notes indeed contain predictive information. The statistical significance test we used to compare the AUC curves is described in (DeLong, DeLong et al. 1988) and we applied it with a significance level of 0.05.

4 Discussion

In our research design we wanted to create a process which would be the most applicable to the data available and most relevant in the clinical setting within the NHS, once robust predictive models on the basis of more data are derived (see further discussion below). The approach taken was to study which referrals are close to a suicide event as this information would aid the initial clinical assessment of patients; this assessment is crucial for deciding which referrals to prioritize and where resources should focus.

The best prediction model in our analysis had an AUC value of just over 0.7. While this result would be considered fair for engineering and some biomedical applications, in the context of mental health diagnoses the picture is very different. Many of the best-performing behaviour checklists and interventions in psychology and psychiatry currently available deliver AUC estimates in the 0.7–0.8 range under clinically realistic conditions (Youngstrom 2013). In fact, (Youngstrom, Meyers et al. 2006) argues that AUCs greater than 0.90 are more likely to indicate design flaws rather than exceptional discriminative validity. Suicide risk assessment is even more difficult as it seeks to predict future human behaviour. In fact, existing clinical instruments for predicting suicide were found to not be clinically useful when classifying “high risk” individuals (Carter, Milner et al. 2017), suggesting an AUC value of around 0.5. Thus, the results reported here would be a major step towards a more accurate assessment of suicide risk, if backed up by further studies.
In considering possible application in clinical practice, our results are tailored to the particular group of patients. Given that existing generic scales fail to provide accurate risk assessment, a strategy seeking to adapt to individuals or groups of individuals appears most promising (Carter, Milner et al. 2017). This focus on risk assessment models tailored to particular settings distinguishes this research from other works, e.g. (Walsh, Ribeiro et al. 2017), that aim at developing generic suicide prediction algorithms. We argue that locally adapted algorithms and models are better suited to take into account local characteristics, leading to potentially better automatic prediction models and deeper insights made available to clinicians.

In parallel to, and independently of our project, research sharing some of our objectives was carried out in the USA (Walsh, Ribeiro et al. 2017). Among others, that work excludes actual suicides and focuses on suicide attempts, while our work focuses specifically on actual suicides; moreover, such a distinction between attempts and suicides is difficult to implement in a clinical setting. In our work we concentrate on the clinically most significant problem: assess at referral point the risk of being close to suicide.

In summary, this work is novel because (a) it is the first of its kind in the context of the UK National Health Service (NHS), (b) it relates to people with mental health problems in secondary care, (c) it takes a referral-centric approach, and (d) the methodology takes into account both general and local parameters (such as patient address or mental health unit).

This work has a number of limitations. Firstly, only data of patients who died by suicide was available to us. Thus, negative samples (referrals taken prior to t months before suicide) are taken only from a population of referrals of patients who eventually died by suicide. The major assumption for our model to be clinically valid and its performance estimation accurate is that the distribution of negative referrals is the same between all patients and patients who eventually died by suicide. This assumption cannot be verified without further research, so the results of our pilot should be read with caution. Still they do provide experimental evidence validating our research design: medical data on past suicide cases contains statistically significant information regarding suicide risk; a referral-based setting is appropriate; both general and local parameters should be considered; both structured and textual information should be considered; a variety of analyses can support various exploitation strategies. Clearly, it is important to incorporate data of living patients in future studies, both because more machine learning techniques can be employed and because the data will contain more information. The combination of the two makes us expect better predictive power once control group data is included. In particular, it would allow to duly take into account the issue of low prevalence of suicide events in the general population, by using machine learning techniques specifically devised for case-control studies where one class is far more common than the other - a problem known as “imbalanced data” in the machine learning community (He and Garcia 2009).

In addition, this study was based on a fairly small data sample. Doing a similar analysis with more data, e.g. from other NHS Trusts, would provide a firmer foundation for the findings. Finally, more advanced text mining algorithms, such as tf/idf features, bag-of-phrases, removal of spelling mistakes and LDA (Blei, Ng et al. 2003), could have been employed.

We carried out a variety of analyses both to get as much insight as possible from the data, but also to support different potential exploitation / deployment strategies once the evidence base is sufficiently robust to allow for deployment in clinical practice. One approach would be fully automatic: data would be passed automatically from the patient information system to the prediction component, which would then calculate a risk factor. This deployment approach allows for the implementation of the best predictive algorithm (based on random forests) and would be fully integrated in the operations of a Trust. However, such a deployment would require changes to the Trust’s existing computing system, be potentially difficult to deploy in other Trusts/settings (that may be using different IT systems) and offer limited explanation of the actual score.
An alternative deployment strategy would use a lightweight standalone system and disregard textual variables. Or the existing guidelines of assessment could be revised to also include findings from this and subsequent studies. Such approaches are easy to implement within the Trust, easily transferable to other settings and deliver results understandable to experts. However, these solutions cannot implement the best predicting model (random forests).

In summary, the findings are encouraging, but this research should be seen as a first step towards improving suicide assessment in clinical settings through the use of machine learning. Further studies are needed before deployment in a clinical setting can be considered.

References

Orfanoudaki, G., M. Markaki, K. Chatzi, I. Tsamardinos and A. Economou (2017). "MatureP: prediction of secreted proteins with exclusive information from their mature regions." Scientific Reports 7. 10.1038/s41598-017-03557-4
Youngstrom, E. A. (2013). "A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC." Journal of pediatric psychology 39(2): 204-221. 10.1093/jpepsy/jst062