Using neural networks with routine health records to identify suicide risk

Abstract

**Background:** Every year approximately 800,000 people die by suicide worldwide, accounting for 1–2 in every 100 deaths. It is always a tragic event with a huge impact on family, friends, community and professionals. Unfortunately, suicide prevention and the development of risk assessment tools have been hindered by the complexity of the underlying mechanisms and the dynamic nature of people’s motivations and intent. Many of those who die by suicide have had contact with health services in the preceding year but identifying those most at risk is a challenge.

**Objective:** To explore the feasibility of using artificial neural networks (ANNs) based on routinely collected electronic health records (EHRs) to support the identification of those at high risk of suicide when in contact with health services.

**Methods:** Using the Secure Anonymised Information Linkage Databank UK, we extracted those who died by suicide between 2001 and 2015 and paired controls. Looking at primary (general practice: GP) and secondary (hospital admissions) EHRs, we built a binary feature vector coding the presence of risk factors at different times before the contact leading to death. Risk factors included: GP contacts and hospital admissions; diagnosis of mental health, injury and poisoning, substance misuse, maltreatment or sleep disorder; and prescription of opiates or psychotropics. We trained simple ANNs to differentiate between cases and controls and interpreted the output score as the estimated suicide risk score. We assessed system performance with 10x10 K-Folds repeated cross-validation and studied system performance and behaviour by representing the distribution of estimated risk across cases and controls and the distribution of factors across different estimated risks.

**Results:** We extracted a total of 2,604 suicide cases and 20 paired controls per case. Our system obtained an error rate of 26.78% ± 1.46 (64.57% of sensitivity and 81.86% of specificity). While the distribution of controls was concentrated around estimated risks < 0.5, cases where almost uniformly distributed between 0 and 1. Prescription of psychotropics, depression and anxiety and self-harm increased the estimated risk by ~0.4. At least 95% of those presenting these factors were identified cases.

**Conclusions:** Despite the simplicity of the implemented system, the proposed methodology obtained an accuracy similar to other published methods based on specialized questionnaire generated data. Most of the errors came from the heterogeneity of patterns shown by cases, some of which were identical to those of
controls. Prescription of psychotropics, depression and anxiety and self-harm were strongly linked with higher estimated risk scores, followed by hospital admission and long-term drugs and alcohol misuse. Other risk factors such as sleep disorder and maltreatment had more complex effects.

**Keywords:** suicide prevention; risk assessment; electronic health records; routine data; machine learning; artificial neural networks

**Introduction**

**Background**

The World Health Organization (WHO) recognizes suicide as a public health priority. WHO Member States have committed themselves to working towards the global target of reducing the suicide rate in countries by 10% by 2020 [1]. In Wales alone around 300 people die each year by suicide, accounting for around 1% of all deaths, three times as many people than in road traffic accidents [2]. The suicide rate has barely altered over the last decade, with any change that has occurred being generally upwards [3]. A death by suicide is always a tragic event with a huge impact on family, friends, the community and professionals. Each death by suicide in the UK has been estimated to cost in excess of £1,370,000 (direct and indirect costs) [4]. In light of the above, suicide prevention using a public health approach “has to be a national priority” [5].

Unfortunately, suicide risk prediction has proven to be a challenging problem for epidemiological studies and their application to practice. The pathway to suicide is mediated by highly complex processes, integrating a large number of risk factor variables which are extensively dependent on one another [6-9]. This results in difficulties around the positive identification of the relatively small number of individuals who will take their own lives from among the much larger group of individuals in whom some or all of the various risk factors have been identified. Assessment of immediate suicide risk requires a clinical evaluation. However, the majority of those who take their own lives present to services other than mental health in the last year of their life. The identification of those who are at risk so that appropriate questions can be asked in relation to suicidality would support suicide prevention efforts.

Short-term suicide risk prediction (i.e. days, weeks or months) would be particularly useful for targeted interventions. However, we know even less about the processes underlying short-term suicidality than when know about their long-term counterparts [10]. Distal or identified long-term risk factors have been found to have complex effects on short-term risk and therefore separate, specific research is needed.

At the same time, we now have data banks curating a wealth of electronic health records (EHRs) and administrative information which, when linked, could provide a representative picture of the biological, societal and health status of an individual at
any point in time. Use of this data at scale is expected to make a pivotal contribution to the study of many diseases [11], especially those with complex longitudinal histories such as suicide. However, the sheer volume of data and the complexity of the suicide factors-risk model have proven to be a challenge for traditional epidemiological and statistical modelling methods. As a result, existing screening tools are reportedly inefficient [12]. Advanced artificial intelligence (AI) techniques are currently better positioned to tackle the combined challenges of big data and suicide risk prediction.

**Prior Work**

Although the application of AI techniques in different areas of medicine is extensive [13, 14], the difficulties of processing routinely collected EHRs and big data in general have been reported elsewhere [15-18]. These include the volume, complexity, heterogeneity and changing nature of medical data as well as its poor mathematical characterization; the importance of physician's interpretations; and the legal, ethical and social implications. It is only recently that we have had the resources to record, maintain and analyse routinely collected EHRs with millions of records.

Most published work on machine learning (a branch of AI) and EHRs in diagnostic applications rely on fuzzy logic [19]. However recently, examples of artificial neural networks (ANNs) applications can also be found. Miotto and colleagues [20] created a deep neural network that received diagnosis codes from a hospital setting and created a ‘patient representation’ vector of 500 features. This vector was then fed to a random forest to predict 78 different diseases, including some mental disorders like schizophrenia. Suicide risk was not a part of this study. The model obtained an accuracy of more than 90% on more than 76,000 patients.

The application of AI in psychiatry is a field that has received relatively little attention but has great potential for innovation [11]. Some proposals found in the literature are optimisation of the delivery of momentary cognitive-behavioural interventions [21], early identification of post-traumatic stress disorder [22] and analysis of social-network information for mental health research [23]. AI studies focusing on suicide risk estimation are even more recent and scarce. Passos and colleagues administered questionnaires to 144 participants with major depressive disorder or bipolar disorder to extract risk-factor information [24]. Suicidality was estimated based on a previous history of suicide attempts. This data was then fed into various machine learning algorithms with the aim of identifying those at high risk of attempting suicide. A best performance of 72% accuracy was obtained with a relevance vector machine.

Kessler and colleagues used a population cohort of non-deployed U.S. Regular Army soldiers, including 147 deaths through suicide, who had a diagnosed mental disorder and at least one outpatient visit [25]. Between 10 and 14 factors were extracted after outpatient visits followed by suicide (cases) and visits not followed by suicide (controls), and used to build an elastic net classifier to predict suicidality.
in the 5 weeks subsequent to these visits. Their system obtained a sensitivity of 48% and a specificity of 84% when predicting suicidality in the following 5 weeks. The authors concluded that their system “outperformed mental health professionals to a large margin”.

**Goal of This Study**

We propose to explore the use of ANNs with routinely collected EHRs to estimate suicide risk within the general population. This takes Passos’ and Kessler’s works one step further by relying on routinely collected EHRs instead of on questionnaires recorded at a hospital setting. This means that our system would not depend on information collected only under specific circumstances (e.g. outpatient visits or hospital admissions) and can therefore be used to screen the whole population without increasing the workload of practitioners.

Our system aims at improving not only suicide risk assessment quality, but also its coverage. This is crucial when considering that only 35% of those who died in Wales by suicide between 2010 and 2015 were admitted to hospital in the year prior to death and around 40% had an emergency department admission. At the same time, 40% of those who died in Wales by suicide between 2001 and 2015 had never had a mental health record before their death, and this increases to 65% when considering only health contacts in the year prior to death. In contrast, around 83% of these cases had at least one contact with their GP during that period. The system exploits on these contacts to assess suicide risk and increase coverage.

In addition, our system has the potential to be able to perform risk assessment continuously over time automatically and on the background (i.e. seamlessly without human intervention) across healthcare settings. Rather than using this as an assessment of immediate ‘at risk’ or ‘not at risk’, it will be used to flag patients, even those attending for reasons other than mental health, so appropriate questions can be asked. The UK National Institute for Health and Care Excellence recommends that risk assessment tools and scales should not be used to predict future suicide or repetition of self-harm [26]. This is because of the dynamic nature of suicide risk, where an individual assessed as ‘not at risk’ on one occasion, may then become at risk subsequently and professionals may not be as responsive to these changes once someone is labelled in this way. This proposed system aims to flag at risk individuals at any contact with services so that relevant questions are asked and appropriately acted upon.

The goal of this study is to test the feasibility of this concept, validating the methodology from functionality (performance) and medical (validity of factors-risk model) points of view. Using an oversimplified system (shallow ANN) we obtain conservative results in terms of model complexity and performance. We combine data from primary and secondary care and use repeated cross-validation during evaluation. In addition, we explore the distribution of factors across different levels of estimated suicide risk to describe the behaviour of the system.
In the remainder of this document, we will first describe the data sources used and how we defined our cohorts of suicide cases and controls and the risk vectors used during experimentation. Next, we will give a brief introduction of ANNs, followed by a detailed description of the architecture evaluated here. We will then detail the analyses ran to assess the raw performance of the system and the factors-risk model it built. After presenting the results, we will discuss their interpretation as well as the potential of the proposed model, how it compares with the state of the art, its limitations and its implications for practice. We will finalise with the conclusions.

**Methods**

**Materials**

**Data sources**
We used data available within the Secure Anonymised Information Linkage (SAIL) Databank [27]. Ethical approval was granted from the HIRU Information Governance Review Panel (IGRP), an independent body consisting of a range of government, regulatory and professional agencies, which overseas study approvals in line with permissions already granted to the analysis of data in the SAIL databank [28; 29] under the SID-Cymru project [30] (approval number 0204).

For this study, we linked and analysed the National Statistics Annual District Deaths Extract (ADDE), the Welsh Demographic Service (WDS), the Welsh Primary Care GP dataset (WGP), the Patient Episode Database for Wales (PEDW) and the Emergency Department Data Set (EDDS). While all datasets were used to define the study case-control cohort, only WDS, WGP and PEDW were used to build the feature vector for experimentation.

Data availability varied across individuals and databases. While ADDE and PEDW datasets have a nationwide coverage, WPG contains data from 348 out of 474 (73%) GP practices in Wales. This variation was reduced by restrictions applied during the cohort definition (see below). At the same time, while the WGP and PEDW datasets were available over the full study period [2001, 2015], ADDE was only available from 2009. However, ADDE data was used only to determine a key date before death, not to train or test the ANN system, and therefore we do not expect this to bias our results significantly.

**Cohort definition**
We extracted our cohort from SID-Cymru, a population based electronic case-control study on completed suicide in Wales between 2001 and 2015 defined within SAIL [30]. There are approximately 32,000 deaths of Welsh residents registered each year of which around 350 are suicides or events of undetermined intent. It is conventional research practice to include the latter in the definition of suicide [31].

The case-control study cohort was built according to the following steps:
1. We identified those that died through suicide at age 10 or older between 2001 and 2015. Deaths of undetermined intent in those under 10 years of age may be related to abuse or neglect.

2. We followed their health history backwards from their death date to identify the full contact leading to death (CLD). This could include multiple entries within the WGP, PEDW and EDDS databases (e.g. attendance to A&E, admission to hospital, transfer to another hospital and finally GP letter received from hospital notifying death). A maximum CLD duration of 1 month was set to account for unrelated hospital stays. This CLD was subsequently removed from the analysis to avoid using information directly linked with the death of cases.

3. Only those residing in Wales at the time of death and with GP data available for at least 80% of the 5 years prior to CLD were included in the study. This ensured that similar data coverage was available for all cases and controls. The value of 5 years was chosen striking a valance between length of health history and number of cases retained.

4. For each case, 20 controls were randomly selected without replacement and excluding cases, after matching by gender and week of birth (±1 year). During control selection, those with similar time period of Welsh Residency and GP data coverage were prioritised to ensure similar coverage quality. Although this number if unnecessary large for traditional paired case-control studies, the proposed methodology benefits from increased data availability during training.

We identified a total of 2,604 suicide cases, 2,012 (77.3%) of which were males, and 58,080 controls. These had a perfect (deterministic) or very high (probabilistic) linkage score (between 0.95 and 1) within SAIL.

**Feature vector**

Only data from WDS, WGP and PEDW were used during experimentation. Not all events in WGP and PEDW represent a face-to-face contact with the patient, and that a single event may have multiple associated entries (e.g. multiple diagnoses).

We categorized each entry in WGP and PEDW into types of health event: depression and anxiety; other common mental disorders (CMD); other mental health; non-intentional injury and poisoning; self-harm; alcohol misuse; drugs misuse; possible maltreatment; physical sleep disorders; non-physical sleep disorders; and others. We also identified the prescription of opiates and psychotropics from within WGP (PEDW has no prescription information) and recorded whether there was any entry recorded in WGP or PEDW (representing a hospital admission). This makes a total of 15 factors (11 diagnoses, 2 prescriptions, a WGP entry and a hospital admission).
The above categories were defined in terms of ReadCodes for WGP and ICD10 for PEDW with the help of expert clinicians and based on previous publications when available (depression and anxiety [32], other common mental disorders (CMD) [33], non-intentional and intentional (self-harm) injury and poisoning [34; 35], alcohol misuse [36], drugs misuse [36, 37], possible maltreatment [38] and psychotropics [39]). Full code definitions can be seen in Tables A1 and A2, Multimedia Appendix 1.

We identified the presence of the above 15 health events during 4 non-overlapping time-frames:

1M: Between CLD and 1 month before CLD [CLD – 1 month, CLD].

6M: Between 1 and 6 months before CLD [CLD – 6 months, CLD – 1 month].

1Y: Between 6 and 12 months before CLD [CLD – 1 year, CLD – 6 months].

5Y: Between 1 and 5 years before CLD [CLD – 5 years, CLD – 1 year].

The final feature vector also included age at CLD and sex, resulting in length 62: 1 float age + 1 binary sex + 15 binary health events * 4 time-frames. This feature vector does not include data directly related to the CLD. Interactions between these factors are automatically designed by the ANN.

System design

Artificial neural networks

Artificial neural networks (ANNs) are biologically inspired computing systems capable of learning tasks through examples/experience, without the need of programming task-specific rules or any a priori knowledge of the solution [40].

ANNs are typically composed of an input layer, one or more optional hidden layers and an output layer (Figure 1). Each neuron in the input and output layers corresponds to one digit of the input and output vector respectively. The complexity of the input-output model is governed by the activation function of neurons, the number of hidden layers, the number of neurons in each layer and the connection between neurons/layers.

The term ‘black-box’ is sometimes used to describe ANNs. This has contributed to the widespread misconception of ANNs not being transparent, which in turn has gained them a bad reputation in fields such as medicine, where understanding how and why decisions are taken is important. However, ‘black-box’ alludes to the fact that the input-output model generated by the network is too complex to be expressed by a set of simple rules that are syntactically meaningful to us. Such a model can nevertheless be expressed as a mathematical equation. For example, a simple ANN composed of no hidden layers and a single output neuron with a logistic activation function is equivalent to the logistic regression model.
\[ y = S\left( b + \sum_{i} w_{ji} x_{i} \right), \]

where \( x_{i} \) are each of the input neurons (i.e. features), \( w_{ji} \) are the weights from the \( i \)-th input to the \( j \)-th neuron, \( b \) is a bias term, \( S(\cdot) \) is the sigmoid function and \( y \) is the output neuron (i.e. result). Typically, the input-output equation quickly grows in complexity, and therefore we opt not to represent it.

**Evaluated architecture**

We implemented a simple ANN with 7 different configurations: no hidden layers (nn0), 1 hidden layer of size 10, 50 or 100 (nn10, nn50, nn100) and 2 hidden layers with sizes 10, 50 or 100 (nn10-10 , nn50-50, nn100-100). All layers were fully connected (i.e. each neuron in layer \( i \) was connected with all neurons of the previous layer \( i-1 \)). The input layer was composed of the feature vector described above (i.e. 50 neurons). Hidden layers, when present, had a \( \text{tanh} \) activation function. The output layer had a single neuron with a sigmoid activation function, returning the score \( r \) of a sample belonging to a (suicide) case if \( r > 0.5 \) or a control \( r \leq 0.5 \). We interpreted this score as the estimated risk of suicide, differentiating between very low risk (VLR; \( r \leq 0.17 \)), low risk (LR; \( 0.17 < r \leq 0.33 \)), moderate-low risk (MLR; \( 0.33 < r \leq 0.5 \)), moderate-high risk (MHR; \( 0.5 < r \leq 0.67 \)), high risk (HR; \( 0.67 < r \leq 0.83 \)) and very high risk (VHR; \( r > 0.83 \)).

All ANNs were trained with the gradient descent algorithm and exponential learning rate decay. The mean square error was adjusted to account for data imbalance (20 controls per case) and the cost of both classes (case and control) was equal to 1. Training was performed sequentially with 3 different batch sizes: 25, 100 and all cases and their respective controls (i.e. total batch size 525, 2100 and full). Training within each batch size continued until a maximum number of epochs was reached, the change of cost function evaluated on the validation set was lower than a threshold or the change was in the negative direction (i.e. not improving).

Using the oversimplified system (i.e. reduced number of features and shallow ANNs) described above, we favoured obtaining conservative results in terms of model complexity and performance, which we hope would counteract some of the limitations of the study (described below). In addition, in a practical application the cost of misidentifying suicide cases and controls will probably not be the same. Whether the system should be tuned to have a high sensitivity at the cost of low specificity or vice versa depends on many factors and it’s out of the scope of this study. For simplicity, we equalized this cost to be the same for cases and controls.

**Statistical Analysis**

**System performance**

We followed a 10x10 k-fold cross-validation approach to evaluate the performance of the ANNs. On each iteration, 1 fold was used for testing, 1 for validation and 8 for
training. Cases were randomly distributed across folds, followed by their respective controls so that case-control pairs were maintained at all times during partitioning (this partitioning rule was also applied during batch partitioning in training).

On each iteration, as well as measuring the classification error obtained with the threshold resulting from training, we varied the threshold to compute the receiving operating characteristics (ROC) curve and the area under the ROC curve (AUROC). We compared performance between systems using a corrected resampled t-test [41] based on the average over sorted runs [42] for 10x10 k-folds, and P-values were further adjusted (Q-values) for multiple testing using the false discovery rate Benjamini and Hochberg (FDR-BH) method [43].

Finally, we repeated the above analysis shuffling the labels of each samples, i.e. we randomly assigned the label ‘case’ to one of the 20 paired controls of a case and rebranded the original case as ‘control’. This aims at evaluating whether our initial results are due to real relationships between labels and data, rather than to random idiosyncratic patterns in the data.

System behaviour
In addition to measuring system performance, we attempted to assess the factors-risk model obtained by the best performing ANN. Due to the dimensionality of the feature vector (i.e. number of input risk factors) and the freedom of the ANN to build complex models with numerous non-linear interactions, getting the full representation of the factors-risk model was not practical. However, the following results allowed us to gain some insights into how large a role each factor played in the computation of the risk score:

- The histogram of the number of cases and controls across estimated risk scores. This will provide information additional to the performance measurements about the classification capability for cases and controls.
- The histogram of the estimated risk difference when turning specific factors ‘on’ and ‘off’ across the whole dataset. This will show an estimate of role of each individual factor in the computation of the risk score, and how this role varies due to the interaction between factors within our dataset.
- The distribution of each factor (i.e. individuals presenting a factor) across estimated risk scores. This will work in conjunction with the previous point to draw an estimate of the role of each individual factor.
- The incidence of each factor within estimated risk scores. This will allow us to compare incidences across risk levels and between risk levels and cases/controls.

Results

System performance
The error rate of the described ANNs decreased slightly from 28.9% to 26.8% when increasing the number of hidden layers from 0 to 2 (Table 1). Overall, nn0 performed
worse than the rest. The performance difference between networks with 1 and 2 hidden layers, although small, is statistically significant ($q < 0.05$) (Table A3 of Multimedia Appendix 1).

Table 1: Mean and standard deviation of the error rate (Err.), sensitivity, specificity and AUROC obtained on the 10x10 k-folds experiments for each neural network.

<table>
<thead>
<tr>
<th></th>
<th>Err.</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>nn0</td>
<td>28.89% ± 1.47</td>
<td>57.28% ± 2.97</td>
<td>84.94% ± 0.54</td>
<td>0.78 ± 0.02</td>
</tr>
<tr>
<td>nn10</td>
<td>27.12% ± 1.42</td>
<td>64.19% ± 2.94</td>
<td>81.57% ± 0.57</td>
<td>0.79 ± 0.02</td>
</tr>
<tr>
<td>nn50</td>
<td>27.09% ± 1.42</td>
<td>64.25% ± 2.92</td>
<td>81.57% ± 0.58</td>
<td>0.79 ± 0.02</td>
</tr>
<tr>
<td>nn100</td>
<td>27.11% ± 1.42</td>
<td>64.18% ± 2.93</td>
<td>81.61% ± 0.61</td>
<td>0.79 ± 0.02</td>
</tr>
<tr>
<td>nn10-10</td>
<td>26.78% ± 1.46</td>
<td>64.57% ± 3.00</td>
<td>81.86% ± 0.58</td>
<td>0.80 ± 0.02</td>
</tr>
<tr>
<td>nn50-50</td>
<td>26.83% ± 1.43</td>
<td>64.52% ± 2.92</td>
<td>81.82% ± 0.59</td>
<td>0.80 ± 0.02</td>
</tr>
<tr>
<td>nn100-100</td>
<td>26.83% ± 1.47</td>
<td>64.54% ± 3.04</td>
<td>81.79% ± 0.61</td>
<td>0.80 ± 0.02</td>
</tr>
</tbody>
</table>

Figure 2 shows the ROC curve of the best performing network for each number of hidden layers (i.e. nn0, nn50 and nn10-10). ROC curves of nn10, nn50 and nn100 were virtually identical, as were curves of nn10-10, nn50-50 and nn100-100. In the false positive range (fpr=1-specificity) range between 0 and 15%, nn50 and nn10-10 perform better than nn0. Past this point, the ROC curves get closer together and for fpr>30% they become virtually identical. Despite the similarity between ROCs of nn50 and nn10-10, the difference in AUROCs between them is statistically significant ($q < 0.05$) (Table A4 of Multimedia Appendix 1). In general terms, nn10-10 and nn50 are capable of obtaining better sensitivity for more restrictive specificity values than nn0, but performs similarly well for higher specificity.

Finally, results after shuffling the labels were characteristic of a random process, i.e. 50% error rate and 0.5 AUROC.

**System behaviour**

The distribution of cases and controls across estimated risk scores reflects the results of Table 1 (Figure 3). Controls were mostly concentrated on scores below 0.5 (hence high specificity) resembling a Poisson distribution. Cases on the other hand were almost uniformly distributed (hence low sensitivity). Overall, few individuals received an estimated risk score $\leq 0.2$.

Prescription of psychotropics, depression and anxiety and self-harm seem to have the strongest effect on the estimated risk, increasing $r$ by $\sim 0.4$ when changing from ‘off’ to ‘on’ across all time-frames (Figure 4). Most of the increase in risk from prescription of psychotropics and depression and anxiety came on the first 6 months before CLD ($\Delta r \approx 0.3$), while self-harm had a more linear effect across time-
frames. The distribution of $\Delta r$ for prescription of psychotropics was the most concentrated around the peak. These 3 factors were followed in strength by hospital admissions and alcohol misuse, with $\Delta r \approx 0.25$. WGP entries, on the other hand, reduced the estimated risk by around 0.2.

Most samples were assigned a risk below the 0.5 threshold, with only 70 individuals resulting in a risk $r \leq 0.17$ (Table 2). In contrast, as many as 1,366 individuals obtained a very high estimated risk ($r > 0.83$). Age and gender distributions were virtually identical across risk levels, except for the very low risk range ($r \leq 0.17$) which was mainly composed of women (Table 2).

Table 2: Number of individuals, gender and mean age for controls, cases and estimated risk levels from very low to very high.

<table>
<thead>
<tr>
<th>Description</th>
<th># Individuals</th>
<th># Males (% [95% CI])</th>
<th>Mean age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls</td>
<td>52080</td>
<td>40240 (77.37% [76.9%, 77.6%])</td>
<td>48.04</td>
</tr>
<tr>
<td>Cases</td>
<td>2604</td>
<td>2012 (77.27% [75.9%, 78.6%])</td>
<td>48.04</td>
</tr>
<tr>
<td>VLR$^a$</td>
<td>70</td>
<td>4 (5.7% [2.6%, 12.1%])</td>
<td>54.32</td>
</tr>
<tr>
<td>LR$^b$</td>
<td>25744</td>
<td>17884 (69.5% [68.9%, 69.9%])</td>
<td>48.07</td>
</tr>
<tr>
<td>MLR$^c$</td>
<td>17818</td>
<td>15850 (88.9% [88.6, 89.3])</td>
<td>46.52</td>
</tr>
<tr>
<td>MHR$^d$</td>
<td>6011</td>
<td>4765 (79.3% [78.4, 80.1])</td>
<td>49.31</td>
</tr>
<tr>
<td>HR$^e$</td>
<td>3675</td>
<td>2703 (73.5 [72.3, 74.7])</td>
<td>53.03</td>
</tr>
<tr>
<td>VHR$^f$</td>
<td>1366</td>
<td>1046 (76.6 [74.6, 78.4])</td>
<td>47.75</td>
</tr>
</tbody>
</table>

$^a$Very low risk ($r \leq 0.17$); $^b$Low risk ($0.17 < r \leq 0.33$); $^c$Moderate-low risk ($0.33 < r \leq 0.5$); $^d$Moderate-high risk ($0.5 < r \leq 0.67$); $^e$High risk ($0.67 < r \leq 0.83$); $^f$Very high risk ($r > 0.83$).

Looking at how factors (individuals with factors ‘on’) were distributed across risk scores (Figure 5, and Tables A5 to A8 of Multimedia Appendix 1), in the month before CLD, 97% of those with a prescription of psychotropics, 96% of those with depression and anxiety and 95% of those with self-harm were classified as being at risk of suicide ($r > 0.5$) (Figure 5). More than 78% of those presenting with these factors or drugs or alcohol misuse at most of the considered time-frames (i.e. 1M, 6M, 1Y and 5Y) were classified as at risk. Moreover, more than half of the individuals with recorded self-harm in the 5 years before CLD, or depression and anxiety or alcohol/drugs misuse in the month before CLD, received a very high estimated suicide risk score ($r > 0.83$).
In terms of incidence (Figure 6, and Tables A9 to A12 of Multimedia Appendix 1), prescription of psychotropics across time-frames had an incidence between 77% and 90% on those with very high risk ($r>0.83$), and lower than 7% on those not at risk ($r\leq0.5$), except on the 5Y period, which had an incidence of 22% on those with moderate-low risk ($0.33<r\leq0.5$) (Figure 6). In comparison, between 35% and 48% of cases presented with this factor. More than 70% had a depression and anxiety event and a hospital event between 1 year and 5 years before CLD.

**Discussion**

**Principal Results**
The presented oversimplified system successfully differentiated between 2,604 suicide cases and 52,080 matched controls in 73.22% of tested instances during 10x10 k-folds cross-validation. It achieved this using only routinely collected EHRs from GP and hospital admissions in the 5 years before the case’s CLD.

The reduction in error rate as the number of hidden layers increased is representative of the complexity of the underlying suicide factors-risk model. In our case, results barely changed when the number of neurons in the hidden layers increased. In fact, performance differences between networks with the same number of layers came from a better tuning of the output scores resulting in an operational point closer to the optimal (i.e. equal error rate). Overall, we expect the advantages of having more layers and neurons to become obvious when larger feature vectors are fed in to the model.

The disparity that was observed between sensitivity and specificity and on the score distribution between cases and controls highlights the variation in the level of difficulty experiences when analysing both groups. Controls seem to follow more uniform patterns and are therefore easier to identify, hence the higher specificity and the clustering of controls below a 0.5 score. On the other hand, patterns of the cases are more heterogeneous, with some having feature vectors identical to controls, which explains the lower sensitivity and the almost uniform distribution of cases across risk scores.

The presented evaluations of the behaviour of nn10-10 do not unequivocally explain the factor-risk model built by the network. However, they do provide enough information to get a general idea of what is driving the output score upwards. The input factors prescription of psychotropics, depression and anxiety and self-harm, and to a lower degree drugs and alcohol misuse, were strongly linked with increasing estimated risk scores. This is in keeping with previous literature [6-8] and provides evidence for proof of concept and the feasibility of identifying high risk individuals based on routinely collected data using ANN. Similarly, gender and age were not related with risk estimation, also in line with findings of short-term risk studies [10].
On the other hand, some risk factors identified in the literature did not exhibit the same behaviour in our results. Physical sleep disorders seemed to decrease the estimated risk rather than increase it. Due to the relatively low incidence of this factor in our data, its effect may be attenuated by and highly dependent on more active factors. This would also explain the variation of its effects on estimated risk score (Figure 4). Furthermore, possible maltreatment also seemed to reduce the estimated risk. However, after a closer look, its effect seems to change sign as the maltreatment gets further away from the CLD, with possible maltreatment in the 5Y time-frame increasing the estimated risk. This may be related to long lasting effects of maltreatment and/or with help and support received in the first year after the maltreatment.

Potential of the proposal
Perfect estimation of suicide risk using EHRs will never be possible. Mainly because some individuals take their own life without ever seeking help or without presenting to health care services with signs of being at risk. In addition, of those that seek help or present with evidence, these signs may be missed by the professional or inaccurately/insufficiently recorded within the EHRs. Some of them may simply present insufficient evidence to distinguish them from controls (i.e. having exactly the same pattern as controls).

According to our data, around 90% of those that died through suicide in Wales had one or more contacts with health services in the year prior to their CLD, and approximately 30% of them had a contact related to their mental health. Therefore, the proposed methodology still has a good scope for application.

Comparison with Prior Work
To our knowledge, Passos’ [24] and Kessler’s [25] are the only two publications to date with proposals comparable to ours. They reported 72% and 66% of accuracy respectively, compared to 73% obtained by our best system. However, these results cannot be directly compared due to differences in the application setting, data used and evaluation process. Firstly, they applied and tested their systems on a hospital setting with only mental health patients. Secondly, their systems used smaller datasets and data extracted from questionnaires or outpatients visits with a specialist. Diagnoses in primary and secondary care (used here) rather than specialist mental health care services will be less specific. Similarly in primary care records there is little indication of severity.

Interestingly, while Kessler’s method also suffered from low sensitivity, Passos’ system obtained comparable sensitivity and specificity. This may be due to the latter using data from the questionnaire Structured Clinical Interview for DSM-IV axis-I Disorders, which records highly specific diagnoses. In addition, Passos’ system aimed at differentiating previous suicide attempters from non-attempters, rather than identifying future risk.
**Limitations**

Results presented here are limited by the purposely oversimplified system design used both in terms of the number of factors considered (only 15) and the design of the ANN (2 hidden layers maximum). Still, our system improved chance identification by almost 50%. As we move from feasibility to pilot study and increase the complexity of the system we expect to increase substantially the success rate.

The problem of suicide risk estimation suffers not only from a highly complex factors-risk model, but also from a lack of a quantitative measure of the real risk of suicide. Real risk of suicide is only known with certainty within a short time span before a recorded attempt (i.e. risk of 1). At any other time point, we do not know the real risk for any individual. Someone at risk may refrain from ever attempting suicide, whereas another person may become at risk and attempt suicide within a very short period of time. This will have implications for a more practical evaluation (compared to the feasibility analysis presented here), as we will need to find ways to assess performance fairly without knowing the real risk ourselves.

Without properly labelled data, we need to rely on clinicians to assess the factors-risk model constructed by the algorithm. In our case, most of the individuals with a self-harm event were classified as cases or as being at risk (i.e. $r > 0.5$). Some of them actually belonged to the control group, and we considered these as errors in our evaluation. However, should all these instances be considered errors? The answer to this question is not trivial, and have technical, clinical and ethical implications that we need to explore in more depth.

**Implications for practice**

Our proposal will be most practical in settings were professionals do not have specialist mental health training but are in contact with individuals. Nurses, emergency department staff, ambulance services and prison workers would be amongst the ones benefiting the most from a tool such as the one proposed here. These professionals face both the challenge of seeing large numbers of people where it is difficult to discern those at risk and of assessing the suicidality of individuals often without having received sufficient training and under staff shortages [44; 45]. As a result it can be a challenge to identify individuals for appropriate assessment and care [46]. Having an advanced assessment tool with complex factors-risk models that produces good estimations would be invaluable in these cases.

**Conclusions**

Prescription of psychotropics, depression and anxiety and self-harm were strongly linked with higher estimated risk scores, followed by hospital admissions and long-term drugs and alcohol misuse which is in keeping with the current literature. Other risk factors such as sleep disorders and maltreatment had more complex effects.

The system presented here is an oversimplified one, using a short feature vector and shallow ANNs to assess the practicality of using electronic health records in this way. As a feasibility study, we were more interested in (a) confirming the existence of...
discriminant information, and (b) validating the proposed methodology, than on obtaining high accuracy rates. Nevertheless, our system obtained an accuracy similar to other published methods based on specialized questionnaire data, hence supporting the hypothesis of the possibility of building a tool capable of estimating suicide risk in the general population using only routinely collected EHRs. We are a long way from employing such methods in clinical practice, but this is a first step to harness the potential of routinely collected electronic health records to support clinical practice in real time.

Prescription of psychotropics, depression and anxiety and self-harm were strongly linked with higher estimated risk scores, followed by hospital admissions and long-term drugs and alcohol misuse. Age and gender had no effect on risk. Interestingly, possible maltreatment had opposite effects in the short and long terms, decreasing risk when recent and increasing it when more than a year before CLD.

**Acknowledgements**

Please include all authors’ contributions, funding information, financial disclosure, role of sponsors, and other acknowledgements here. This description should include the involvement, if any, in review and approval of the manuscript for publication and the role of sponsors. Omit if not applicable.

**Conflicts of Interest**

Disclose any personal financial interests related to the subject matters discussed in the manuscript here. For example, authors who are owners or employees of Internet companies that market the services described in the manuscript will be disclosed here. If none, indicate with "none declared".

**Abbreviations**

ADDE: National Statistics Annual District Deaths Extract  
AI: artificial intelligence  
ANN: artificial neural networks  
AUROC: area under the ROC curve  
CMD: common mental disorders  
CLD: contact leading to death  
EDDS: Emergency Department Data Set  
EHR: electronic health records  
FDR-BH: false discovery rate Benjamini and Hochberg  
GP: general practice  
HR: high risk (0.67<r≤0.83)  
LR: low risk (0.17<r≤0.33)  
MHR: moderate-high risk (0.5<r≤0.67)  
MLR: moderate-low risk (0.33<r≤0.5)  
OR-CC: odds ratio between cases and controls  
OR-R: odds ratio between ‘at risk’ and ‘not at risk’ groups  
PEDW: Patient Episode Database for Wales  
ROC: receiving operating characteristics
SAIL: Secure Anonymised Information Linkage databank
VHR: very high risk (r>0.83)
VLR: very low risk (r≤0.17)
WDS: Welsh Demographic Service
WGP: Welsh Primary Care GP dataset
WHO: World Health Organization

References
26. NICE (2011), Self harm (longer term management) (CG133), paragraph 1.3.11, page 22.