An Integrated Framework for Suicide Risk Prediction

Truyen Tran[†‡], Dinh Phung[†], Wei Luo[†], Richard Harvey[♭],
Michael Berk[♮], Svetha Venkatesh[†]
[†]Pattern Recognition and Data Analytics, Deakin University, Australia
[‡]Department of Computing, Curtin University, Australia
[♭]Mental Health Drugs & Alcohol, Barwon Health, Victoria, Australia
[♮]School of Medicine, Deakin University, Australia
{truyen.tran,dinh.phung,wei.luo}@deakin.edu.au
{richardha,mikebe}@barwonhealth.org.au
svetha.venkatesh@deakin.edu.au

ABSTRACT

Suicide is a major concern in society. Despite of great attention paid by the community with very substantive medico-legal implications, there has been no satisfying method that can reliably predict the future attempted or completed suicide. We present an integrated machine learning framework to tackle this challenge. Our proposed framework consists of a novel feature extraction scheme, an embedded feature selection process, a set of risk classifiers and finally, a risk calibration procedure. For temporal feature extraction, we cast the patient's clinical history into a temporal image to which a bank of one-side filters are applied. The responses are then partly transformed into mid-level features and then selected in ℓ_1-norm framework under the extreme value theory. A set of probabilistic ordinal risk classifiers are then applied to compute the risk probabilities and further re-rank the features. Finally, the predicted risks are calibrated. Together with our Australian partner, we perform comprehensive study on data collected for the mental health cohort, and the experiments validate that our proposed framework outperforms risk assessment instruments by medical practitioners.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health; I.6.5 [**Simulation and Modelling**]: Model Development

Keywords

medical data analysis, suicide, risk modelling, risk prediction, one-side convolutional kernels, filter bank, machine learning

1. INTRODUCTION

Suicide is widely considered as a major problem in mental and public health, and is a main cause of death. WHO estimates that worldwide, suicide accounts for nearly 2% of deaths by 2000 [2]. Although there is a decreasing trend in the number of suicide-classified deaths in Australia, there is no decline in the suicidal ideation or attempts [11]. In a 2007 national survey, 2.9% of the population had suicidal ideation, and 0.4% had attempted suicide [8]. This poses grave challenges for mental health service providers, and the open question is how to improve early detection of suicide and prevention.

Mandatory practice in health services is to perform risk assessments, serving as one of the gate-keeper indicators in triage to determine nature of care. Such assessments have medico-legal consequences. However, the reliability and validation of suicide risk assessments is not well understood in terms of predictive power, and remains a controversial issue in risk management (e.g., see [18, 10]).

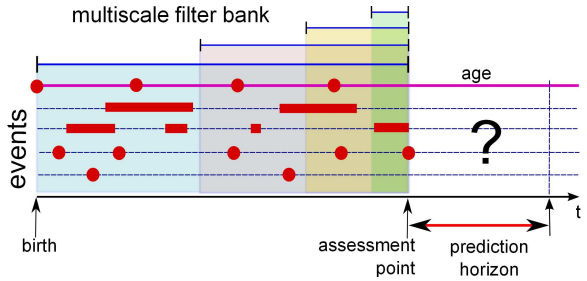

Figure 1: Clinical events represented as a temporal image, which is convoluted with one-sided filter bank.

We ask a bold question: Can we predict suicides automatically, given mental history, risk assessments and clinical intervention data? We aim to predict the probabilities within a given future period of sentinel events: low-risk, moderate-risk and high-risk events. Low-risk events implies no detected suicide risks, moderate-risk events are self-injuries that do not lead to fatal consequences, and high-risk events are those with fatal results. The convention is that if several events occur within the same period, the highest severity is considered. The cohort under study is from Barwon Mental Health, Drugs and Alcohol Services, the only provider in the region of 350,000 people in the central western region of Victoria in South-eastern Australia.

We depart from the standard medical practice of considering a small set of risk factors and limited risk levels based on expert knowledge (e.g., see [4]). We exploit large medical datasets, *generating thousands of potential signals* from multiple sources. We then employ machine learning to automatically select strong and reliable risk factors of future attempts or suicide. The goal is then to develop an automated tool that: (i) provides *objective measures* of risk factors quantifying uncertainties; (ii) detects *risk patterns* from the patient history; and (iii) computes *probabilities of outcomes*.

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

ACM 978-1-4503-2174-7/13/08 ...$15.00.

Key modelling considerations are *transparency* in modelling decisions and *interpretability* in results.

The problem is highly challenging (e.g., see [16]). Documented risk factors, such as those used in risk assessments may not correlate well with future outcomes. High risk events are rare and irregular. The data is aggregated from different sources, is incomplete (e.g., people reported dead without any noticeable history) and contains a significant noise (e.g., service providers under pressure might enter "junk" data to meet protocol requirements). The data is severely imbalanced. Time scales for event evolution can be very different, ranging from days to decades.

Our proposed framework implements three key concepts to tackle these challenges: automated scale-invariant feature generation and selection from big data, subsequent ordinal risk classification and finally, risk calibration to account for data imbalance. For feature extraction, we offer a novel conceptual view in which a patient's medical record is cast as a temporal image, features vs time, on which an *one-sided filter bank* can be applied (see Fig. 1 for an illustration and Sec. 3.1 for details). The extracted feature set is *temporally sensitive* and takes into account time sensitive nature of varied medical event evolutions. The feature pool is then pruned by a supervised procedure that penalises features that are weakly indicative of future attempts in a ℓ_1-norm framework, under the *extreme value theory* [6] (Sec. 3.2). A set of *ordinal classifiers* are designed to further re-rank and select surviving features, simultaneously computing probabilities of risk levels within a window of time (Sec. 3.3). Finally, the predicted risks are *calibrated* to meet operational requirements (Sec. 3.4).

To the best our knowledge, this is the first comprehensive data mining type of work for this important domain: Our database contains more than $10,000$ mental patients, with more than $25,000$ risk assessments. We consider up to $10,000$ extracted signals and a series of ordinal classifiers. Significantly, we show that using this data, our method's prediction outperforms the risk assessment instrument used by medical practitioners (Sec. 4).

The significance of our framework is that it is agnostic to disease type - given mixed type data comprising demography, clinical history (emergency attendances, admissions and diagnostic coding), and risk assessment instruments - questions with ordinal ratings, our framework automatically extracts the most relevant features and builds risk prediction classifiers. This allows us to easily apply the proposed framework across disease domains.

2. DESCRIPTIVE DATA ANALYSIS

2.1 Data Description

The Barwon Health hosts a data warehouse in which electronic medical records are pooled. We are mainly interested in data on emergency department presentations and admissions to the general hospital. For *emergency attendances* (ED), there are $42K+$ recorded mental cases for $8K+$ patients in the period of 2005–2012. For *hospital admissions* (HA), there are approximately $67K$ recorded mental cases in the period of 1995–2012. The number of recorded emergency attendances and admissions has been increasing over the year, e.g., from 7,068 admissions in 2009 to 8,143 in 2010 and 8,956 in 2012.

The most important piece of information is the diagnostic coding for any episodes. Each ED record contains only one main code, but an admission is typically associated with multiple codes, some of which reflects mental status. The diagnosis coding conforms with

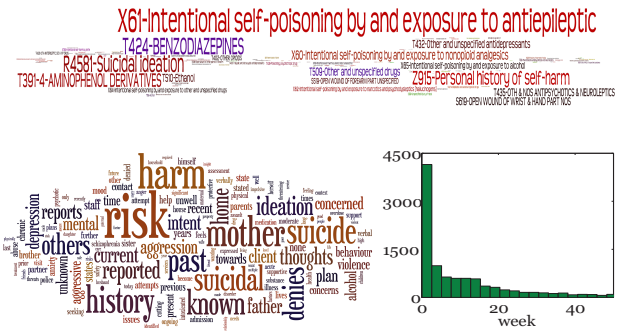

Figure 2: Some data distributions: (Top) codes suggesting attempts; (Bottom left) words in assessment notes; and (Bottom right) interval between risk assessments in weeks. Best viewed in colour.

the latest classification scheme, the ICD-10[1]. This is a hierarchy of diseases covering almost all known conditions. The codes start with a letter followed by several digits where the digits placed later in the sequence indicate more specific conditions. For example, injuries to the head are classified into 10 groups, from *S00* to *S09*. The group S01 would mean "open wound of head", the subgroup *S011* means the "wound in the eyelid and periocular area". Some ICD codes can be readily interpreted as suicide attempts, for example, the code *R4581* refers to suicide attempt or ideation. However, other codes must require expert interpretation about severity of incidents. Fig. 2 (top) depicts the distribution of codes which are associated with suicide risks.

Another feature of mental health data held at Barwon Health is that it contains *suicide risk assessments* for every mental patient under its management. The local clinical protocol requires a suicide risk assessment on admission, and then after every 91 days during the care episode, and finally on discharge. The assessment instrument used by Barwon Health was developed in-house in 1999. It has ordinal assessments for 18 items covering all mental aspects such as suicidal ideation, stressors, substance abuse, family support and psychiatric service history. See Fig. 2 (bottom left) for examples of keywords noted by the assessment staff.

The system recorded approximately $25K$ assessments on $10K$ patients in the period of 2009-2012. The majority of patients have only one assessment (62%), followed by two assessments (17%), but there are about 3% patients who have more than 10 assessments. For those with more than one assessments, the time between two successive assessments are: 30% within one week, 64% within 3 months (Fig. 2 (bottom right)).

Three other assessment tools are also required by the Australian government: the HoNOS (Health of the Nation Outcome Scales), the LSP (Life Skills Profile) and the BASIS-32 (Behaviour and Symptom Identification Scale). These provide different perspectives on the mental health of a patient. However, it appears that they are incomplete and noisy and thus our early experiments on these instruments did not result in positive findings.

We focus our study on those patients who have had a least one event prior to a risk assessment. The dataset then has $7,746$ patients and $17,771$ assessments. For each patient, we collect age, gender, spoken language, country of birth, religion, occupation, marital status, indigenous status, and the postcodes over time. Among

[1]International Statistical Classification of Diseases and Related Health Problems 10th Revision, available at: http://apps.who.int/classifications/icd10/browse/2010/en

Horizon (day)	**14**	**30**	**60**	**90**	**180**
C_1	16,985	16,525	15,952	15,471	14,490
C_2	536	836	1,174	1,440	19,29
C_3	250	410	645	8,60	1,352
Suicide	7	24	32	41	63

Table 1: Outcome class distribution following risk assessments.

patients considered, 48.7% are male and 48.6% are under 35 of age at the time of assessing.

The risk assessments are the evaluation points from which future prediction is required. Future outcomes are broadly classified into three levels of risk, based on expert at Barwon Health: class C_1 refers to low-risk outcomes, class C_2 refers to moderate-risk (non-deadly attempts), and class C_3 the high-risk (deadly outcomes). The classes are assigned using a look-up table from the diagnosis coded events. The convention is that among all events occurring within the prediction period, the class of the highest risk is chosen. For example, the ICD-10 coded event *S51* (open wound of forearm) is moderate-risk, while *S11* (open wound of neck) would be considered as high-risk. Typically the completed suicides are rare, and the class distributions are imbalanced. For example, for 2-week period following the risk assessment, there are only 7 suicides among 250 lethal attempts (1.4%), and 536 moderate-risk attempts (3.0%). Further class distributions are summarised in Table 1.

2.2 Medical Data Modelling

After data pooling, we obtain a temporal medical database where each patient has multiple time-indexed records. Each record specifies a particular event such as risk assessment, moving home, admission, diagnosis, lab test, or medicine prescription. In general, the data characteristics can be summarised as follows:

- ***Sparsity***. Only limited number of events are recorded.
- ***Irregularity of episodes***: Events are recorded at irregular intervals. An episode of events (such as diagnoses and interventions) may follow a doctor visit or an emergency attendance, but the trigger time is randomly distributed.
- ***Variable length***: Patient records vary greatly in length. Some chronic patients will have long longitudinal data.
- ***Shift-invariance***: It is of clinical importance to account the progression from a major event point, e.g., diagnosis. The absolute time point is not too relevant.
- ***Heterogeneity***: Patient records contain information of different types, some are continuous, such as blood pressure, but many are discrete. Some events are recorded only once (e.g., birth), but many others may be recorded in short-intervals (e.g., heart beats). Some event types change slowly, such as aging, but some others move fast.
- ***Distribution drifts***: New recording procedures, policies, findings and treatments are introduced at increasing pace, and thus creating drifts in event distributions.
- ***Contextual information***: Backgrounds (e.g., gender, education, religion, age) and primary care (GPs, insurances) play critical roles in clinical settings.

We note that similar observations have also been partly stated in [21], and these characteristics are common for other medical services as well.

The suicide risk analysis has been mostly carried out in the traditional medical research (e.g., see [4]), and it is well recognised that it is very hard to predict the actual suicidal outcomes (e.g., see [16]). The common feature in these studies is that the risk factors are manually designed based on expert knowledge, and thus each study can only handle a handful of such factors. The risk assessment instrument developed and practiced at Barwon Health, for example, is composed of 18 items. In data mining and machine learning, the problem of suicide risk prediction has largely been overlooked. Recent work of [17] proposes a Bayesian nonparametric approach to suicide attempt modelling. The main idea is to represent each patient by a set of binary features discovered from the data. However, the study has limited value in practice since it mainly involves interviews and does not contain real outcomes but ideation, which is known to be weakly associated with real attempts and suicide.

3. PREDICTIVE FRAMEWORK

Our ultimate goal is to predict attempts and associated lethality in the future, often at the point of risk assessments. We describe an integrated predictive framework which has the following components:

1. *Temporal feature extraction:* Most of the features are temporal, except for demographic variables like gender or country of birth. Some mental problems are long-term but suicidal episodes are often short (from few days to less than 6 months), thus it is necessary to take multiple time scales into account;
2. *Feature selection* using a surrogate task of detecting attempts embedded with ℓ_1-norm regularisers. The attempts are assumed to be triggered when the extremal hidden suicidal risk goes beyond a certain threshold;
3. *Risk classification* given the observed history. This is the main part of the model where the future risk is regressed again history (which is captured by the temporal feature extractors); and
4. *Risk calibration* for translating the probability of risk in to tunable prediction of outcomes to deal with the imbalanced data.

The second and third components are placed within the bootstrapping framework [5, 3, 15][3] for better stability and predictive performance.

3.1 Temporal Feature Extraction

Our problem is to construct a set of sensible features at a particular time in the patient history. It is desirable that the feature pool has a good coverage and is highly informative for the risk prediction tasks at multiple time-scales. In other words, the feature set should be insensitive to scales. The main conceptual insight is that *much of the clinical records can be represented as a sparse temporal image*, where at any given point of time, we can only look back to the recorded history. The key concept we introduce here is the *one-sided filter bank*[2] for detecting temporal features.

3.1.1 Representation of Patient History

Data includes demography, detailed clinical history and risk assessments. Clinical history includes a series of admissions and emergency visits. Each admission typically contains a subset of ICD diagnosis codes, procedure codes, diagnosis-related groups and discharge medications. To deal with the plethora of ICD codes, we preprocess to separate the rare codes, which we consider as one observation type.

[2]This is somewhat analogous to the concept of filter-bank in signal processing and computer vision.

Let t be the time point of interest, H be the maximum history length. Let $v_i(t)$ be the observation of the event of type i at time t and let there be D event types. If the event is not observed in the data, then $v_i(t) = 0$, otherwise, it is a real value if it is some measure, or 1 if it is an occurrence[3]. An event for an ICD code is the presence or absence of code. For demography, some events are fixed over time (like gender); for postcodes, we consider an event if a change of postcode has occurred; and for age, we discretise into bands, that is an event is recorded when the age reaches a particular band at the assessment time. For continuing events such as treatment episodes, $v_i(t)$ is the duration given that the entire episodes are in the history. Then a representation of the patient history is as depicted in Fig. 1.

3.1.2 One-Sided Filter Bank

Different events have different resolutions in time - an attempted suicide is time critical, whereas a Type I diabetic ICD code is not. To accommodate events having different time scales of evolution, we consider a multiscale temporal filter bank. For each event type i, we have a set of K filters over varying timescale. Each filter essentially evaluates the strength of the event type at that scale. Let $\mathcal{K}^k \in \mathbb{R}^{H+1}$ be the k-th one-sided filter (or kernel), the k-th feature evaluated at t for event i is

$$f_i^k(t) = \sum_{h=0}^{H} \mathcal{K}^k(h) v_i(t-h) \tag{1}$$

where $\mathcal{K}^k(h)$ is the convolution kernel evaluated at h. One useful kernel is the *truncated Gaussian*

$$\mathcal{K}^k(h) = \sqrt{\frac{2}{\pi\sigma_k^2}} \exp\left(-\frac{h^2}{2\sigma_k^2}\right) \tag{2}$$

where $\mathcal{K}^k(h) > 0$ for $h \geq 0$ and 0 otherwise. The hyper-parameter σ_k defines the effective width of the kernel, i.e., the response drops drastically as h goes beyond σ_k. The behaviour is similar to the *uniform kernel* with specified width σ_k

$$\mathcal{K}^k(h) = \frac{1}{\sigma_k} \mathbf{1}\left[h \in [0, \sigma_k]\right] \tag{3}$$

This kernel counts the normalised number of events falling within a given period of time.

Capturing temporal structure. A filter bank of multiple scales partly captures the temporal structure in the patient history. However, the nature of event aggregation using the kernels does not reveal temporal changes within the "medical image", for example the rise and fall of stress over time. We propose a simple way to do this by dividing the image into temporal fragments. Each fragment is then evaluated through filter responses and all fragment responses are concatenated. Indeed this can be captured using the same kernels as above but with a shifting operation, i.e., the convolution in Eq. (1) is modified as follows

$$f_i^k(t) = \sum_{h=0}^{H} \mathcal{K}^k(h - s_k) v_i(t-h) \tag{4}$$

where s_k denotes the delay from which the kernel operation has effect.

Finally, the design of filter bank is characterised by a set of pairs (σ_k, s_k). In this particular suicide application, we choose the pairs to be $(\sigma_k, s_k) \in \{(0.5, 0); (1, 0); (3, 0); (3, 3); (6, 6); (12, 12)\}$ (in months). That is, the history $H = 24$ months is considered. At the current evaluation point, three kernel widths are 0.5, 1 and 3 months reflecting the short-term scales of the mental risks. The delays of 3, 6 and 12 months are designed to capture the medium-term progression of the mental state and comorbidities.

[3]If an event is missing, it may due to the fact that nothing happens, or it is not recorded, or the time t is in the future.

3.2 Feature Selection

Given several risk factors, we need to find a *compact subset* that best explains suicide outcomes. Since suicides are rare, we look at the suicide attempts as the *first approximation*. Thus we are concerned with the setting where there are binary outcome of a suicide attempt) $y \in \{0, 1\}$, given the features. We choose the Generalised Linear Model (GLM) framework [13] with the complementary log-log link function, which is essentially the application of the Extreme Value Distribution (the Gumbel distribution) [6]. This link function is motivated by the fact that suicide attempts are at the extreme end of the risk spectrum.

Let $\mu(\boldsymbol{f}) = u_0 + \sum_i u_i f_i$ be the mean risk, where $u_0, u_1, ...u_n$ are feature weights. The probability of an suicide attempt is given as

$$P(y = 1 \mid \boldsymbol{f}) \quad = \quad 1 - \exp\left(-e^{\mu(\boldsymbol{f})}\right)$$

The model estimation and the feature selection can be carried out simultaneously by maximising the ℓ_1-regularised log-likelihood on training data $\mathcal{D}$

$$\mathcal{L}(\boldsymbol{u}) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \log P(y^d \mid \boldsymbol{u}, \boldsymbol{f}^d) - \lambda_1 \sum_i |u_i| \tag{5}$$

where $\lambda_1 > 0$ are regularisation parameters. In general, larger λ_1 would lead to sparser models (e.g., many features are not selected). This setting is essentially a variant of the *lasso* (the original problem was linear regression [19]). The output of this step is the list of features with non-zeros weights.

3.3 Risk Classifiers

Here we describe a set of models to deal with the ordinal nature of the suicide outcomes. Our goal is not only to come up with high performing classifiers but also to offer a reasonable interpretation of modelling choices. In particular, we assume that the observed outcomes are the discretised version of *underlying random risks* $x \in \mathbb{R}^m$. The probabilistic models are natural to estimate the probability of a particular risk class being observed. Let L be the number of discrete levels of lethality, in which level 1 refers to the normal state where no risk can be observed, and level L refers to the most fatal state or even death. The outcomes are regressed against the feature vector $\boldsymbol{f}$ evaluated at the time t.

3.3.1 k-Nearest Neighbours (k-NN)

k-NN makes no assumption about the underlying random risk, but it is based on the foundation that patients with similar recent history would assume similar risk in the near future. To this end, for each patient at any evaluation point t, we choose k similar history fragments from other patients with known outcomes and compute class probabilities of the outcomes in that neighbourhood. That is $P(r^d = l \mid \boldsymbol{f}^d) = \frac{1}{k} \sum_{p \in N(d)} \mathbf{1}[r^p = l]$, where $N(d)$ is the k-nearest neighbourhood of data d.

3.3.2 Linear Classifiers of Gaussian Risk

We assume that the underlying random risk is normally distributed, and resembles the lethality level $l-1$, i.e., we treat the discrete levels as real values and $x = r \sim \mathcal{N}(\mu(\boldsymbol{f}), \sigma)$. The distribution mean

is modelled as a linear function of features:$\mu(\boldsymbol{f}) = w_0 + \sum_i w_i f_i$. However, since we are mainly interested in the probabilities of the discrete outcomes, we need a way to convert from the continuous distributions $\mathcal{N}(\mu(\boldsymbol{f}), \sigma)$. We employ the following transforms $P(r = l \mid \boldsymbol{f}) = \frac{1}{Z(\boldsymbol{f})} \exp\left(-\frac{(l-\mu(\boldsymbol{f}))^2}{2\sigma^2}\right)$, where $Z(\boldsymbol{f})$ is the normalising constant. The standard deviation σ can be estimated from the set $\left\{e^d = r^d - \mu(\boldsymbol{f}^d)\right\}_{d\in\mathcal{D}}$ on the training data $\mathcal{D}$.

3.3.3 Cumulative Models of Risk Grouping

This model assumes that the discrete outcomes r are generated from the *one-dimensional* underlying random risk $x \in \mathbb{R}$ as follows [12]

$$r = \begin{cases} 1 & \text{if } x \le \tau_1 \\ l & \text{if } \tau_{l-1} < x \le \tau_l \\ L & \text{otherwise} \end{cases}$$

where $\tau_1 \le \tau_2 \le ...\tau_{L-1}$ are thresholds. This essentially says that the discrete outcome is a coarse version of the real-valued risk. Here the risk spectrum is the real line divided into intervals, each of which determines the corresponding outcome. In the form of probability distribution we have

$$\begin{aligned} P(r = l \mid \boldsymbol{f}) &= P(\tau_{l-1} \le x \le \tau_l \mid \boldsymbol{f}) \\ &= F(\tau_l \mid \boldsymbol{f}) - F(\tau_{l-1} \mid \boldsymbol{f}) \end{aligned}$$

where $F(\tau_l \mid \boldsymbol{f})$ is the cumulative distribution evaluated at τ_l. Choosing the form of $F(\tau_l \mid \boldsymbol{f})$ is usually the matter of practical convenience since x is unobserved and we do not know the true underlying distribution. For example, assume that the mean risk functional is linear in features, i.e., $\mu(\boldsymbol{f}) = \boldsymbol{w}'\boldsymbol{f}$, the logistic distribution $F(\tau_l \mid \boldsymbol{f}) = [1 + \exp(-(\tau_l - \mu(\boldsymbol{f})))]^{-1}$has an interesting interpretation:

$$\log\left(\frac{P(r \le l \mid \boldsymbol{f})}{P(r > l \mid \boldsymbol{f})}\right) = \tau_l - \boldsymbol{w}'\boldsymbol{f}$$

i.e., the log odds at the split level l is proportional to the risk factors. Another distribution is the Gumbel family studied in Sec. 3.2, and this can provide an interpretable model in terms of extremal risks.

This leaves a question of how to estimate $F(\tau \mid \boldsymbol{f})$ and the thresholds $\{\tau_l\}_{l=1}^{L-1}$. Since $\tau_1, \tau_2, ..., \tau_{L-1}$ is a monotonically increasing sequence, we enforce this monotonicity by using

$$\tau_l = \tau_{l-1} + e^{\gamma_l}$$

for $l = 2, 3, ..., L-1$, where $\gamma_l \in \mathbb{R}$, which is unconstrained. More details are left until Sec. 3.3.6.

3.3.4 Stagewise Models of Risk Progression

Cumulative models assume a single risk variable that can explain the ordinal outcomes. This assumption is quite limited and does not address the nature of the risk progression - for some patients, the risk may not reach a certain level immediately. It may, alternatively, start from a normal condition, and then progress upward. This suggests a stagewise model of outcomes: The next outcome level may be attained only if the lower levels have not been attained [1, 20].

Since there are several stages, we need not assume that there is only one underlying risk variable. Instead, the risks can be *multi-dimensional*, i.e., $\boldsymbol{x} \in \mathbb{R}^{L-1}$ and each stage $l \in \{1, 2, .., L-1\}$ assumes their own underlying risk variable $x_l \in \mathbb{R}$. The stagewise process can be formalised as follows

$$r = \begin{cases} 1 & \text{if } x_1 \le \tau_1 \\ l & \text{if } \{x_m \ge \tau_m\}_{m=1}^{l-1}, x_l \le \tau_l \\ L & \text{otherwise} \end{cases}$$

Here, the transition from level l to level $l+1$ is signified by the event that the risk value passes through the level-specific threshold τ_l. The probability that the outcome is the lowest is then

$$P(r = 1) = P(x_1 \le \tau_1) = F_1(\tau_1)$$

where $F_1(\tau_1)$ is the level-1 cumulative distribution. If the condition $x_1 \le \tau_1$ does not hold, then we consider level 2

$$P(r = 2 \mid r \ge 2) = P(x_2 \le \tau_2) = F_2(\tau_2)$$

This process continues until some level has been accepted, or we must accept the last level L. Thus the probability of having the highest level of risk, given all the lower levels have not been accepted, is

$$P(r = L \mid r > L-1) = 1 - F_{L-1}(\tau_{L-1})$$

Note that the probabilities above are *conditional*. The marginal probability of selecting a particular discrete outcome is

$$P(r = l) = \begin{cases} F_1(\tau_1) & \text{if } l = 1 \\ F_l(\tau_l)\prod_{m=1}^{l-1}(1 - F_m(\tau_m)) & \text{if } l \in \{2, .., L-1\} \\ \prod_{m=1}^{L-1}(1 - F_m(\tau_m)) & \text{otherwise} \end{cases}$$

With the choice $F_l(\tau_l)$ as a logistic distribution and the linear risk functional $\mu(\boldsymbol{f}) = \boldsymbol{w}'\boldsymbol{f}$ we have a nice interpretation

$$\log\left(\frac{P(r = l \mid \boldsymbol{f})}{P(r \ge l \mid \boldsymbol{f})}\right) = \tau_l - \boldsymbol{w}'\boldsymbol{f}$$

i.e., the log odds of the probability of choosing the next level, given the fact that all previous levels have failed, is proportional to the risk factors $\boldsymbol{f}$.

At this point, we are left with two choices: (i) using the same distribution across all levels, i.e., $F_l(x) = F_1(x)$ for all $l \in \{2, 3, .., L-1\}$, or (ii) using level-specific distributions. The later choice has more parameters, and thus more flexible.

3.3.5 Multinomial Models of Independent Choices

While the stagewise models greatly relax the assumption of the underlying random risks, the stagewise risk progression process is at best an approximation to the true process. Here we relax the assumption even further: (i) Outcomes are individual choices that are independent of other choices, (ii) An outcome is observed because it is the most likely choice among all choices given the situation.

Like the stagewise models, we assume that the underlying risk are multidimensional, i.e., $\boldsymbol{x} \in \mathbb{R}^L$, one dimension for a possible outcome. An outcome is observed if its underlying risk is the largest among all other underlying risks, i.e., $r = l$ if $x_l \ge \max_{m\ne l}\{x_m\}$. It has been proved that under the Gumbel distribution, this decision rule leads to the standard multinomial model $P(r = l \mid \boldsymbol{f}) \propto \exp(\mu_l(\boldsymbol{f}))$ [14]. Let $\mu_l^*(\boldsymbol{f}) = \mu_l(\boldsymbol{f}) - \mu_1(\boldsymbol{f})$, this simplifies to

$$\begin{aligned} P(r = 1 \mid \boldsymbol{f}) &= \frac{1}{1 + \sum_{m=2}^{L}\exp(\mu_m^*(\boldsymbol{f}))} \\ P(r = l \mid \boldsymbol{f}) &= \frac{\exp(\mu_l^*(\boldsymbol{f}))}{1 + \sum_{m=2}^{L}\exp(\mu_m^*(\boldsymbol{f}))} \end{aligned}$$

3.3.6 Model Estimation

The probabilistic models, except for the k-NNs, are estimated by maximising penalised likelihood over the training data $\mathcal{D}$

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\tau}) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \log P(r^d \mid \boldsymbol{f}^d; \boldsymbol{w}, \boldsymbol{\tau}) - \lambda_2 \sum_i |w_i| \quad (6)$$

where $\boldsymbol{\tau}$ are thresholds in the cumulative and stagewise models. The role of the ℓ_1-penalty is to further select strongest features for the predictive task. For the cumulative and stagewise models, we fix the first threshold $\tau_1 = 0$ and learn the others. For ease of interpretation, we employ the simple linear functional $\mu(\boldsymbol{f}) = w_0 + \boldsymbol{w}'\boldsymbol{f}$ if parameters are shared among all levels or $\mu_l(\boldsymbol{f}) = w_{0l} + \boldsymbol{w}_l'\boldsymbol{f}$ otherwise. For the multinomial model, we simply fix $\mu_1(\boldsymbol{f}) = 0$.

3.4 Risk Calibration

Let us now consider a specific situation at Barwon Health, where the outcomes are broadly classified into three levels of risk: class C_1 refers to low-risk outcomes, class C_2 refers to non-deadly attempts, and class C_3 the most deadly outcomes. Once trained, the classifiers described above produce the probabilities of future risk classes $P(r \mid \boldsymbol{f})$. However, there are two major problems with this setting. First, for everyday practice, it may create significant cognitive load for physicians to reason in terms of numerical probabilities. Second, the data collected here is highly imbalanced: For three-month horizon, only 8.1% data points belong to the class C_2 and 4.8% belong to C_3 (Sec. 2.1, Table 1). This leads unavoidable bias in estimation which is unfavourable towards the most important class, the C_3.

To mitigate the problems, we employ a simple calibration that first translates the risk class probabilities into a single, interpretable *risk index*, and derives a rule to assign the risk classes, in a manner similar to the cumulative model (Sec. 3.3.3). This translates into the following procedure:

1. Estimate the risk index, which is the expected risk, on each data point $\langle r_i \rangle = \sum_{m=1}^{3} (m-1) P(r = C_m \mid \boldsymbol{f}_i; \boldsymbol{\theta})$ for all training/test points i; This ensures that the risk index is a positive number bounded within $[0, 2]$.
2. For each test point j, predict the test classes by using the following decision rules: output C_1 if $\langle r_j \rangle \leq \tau_{low}$; C_2 if $\tau_{low} < \langle r_j \rangle \leq \tau_{high}$; and C_3 otherwise. The thresholds τ_{low} and τ_{high} are controlling parameters which determine the recall/precision trade-off. In practice, they are estimated from the percentiles of the training risk indices.

3.5 Bootstrapping

One potential problem with the pipeline we have just described is the instability of the model, especially the selected features, due to the data sampling. That is, a different data collection scheme may produce an entirely different model, leading to the interpretation problem and high variance in the classifiers. To achieve stability and potentially boost the prediction performance, we draw from the existing literature of bootstrapping [5], including bagging [3] and stability selection [15]. The overall training loop is as follows:

1. For each bootstrap $b = 1, 2, ..., B$
   (a) Draw a training sample of original size $|\mathcal{D}|$ with replacement.
   (b) Subsample the class C_1 so that its data size is at most twice the size of $C_2 + C_3$
   (c) Select features (Sec. 3.2)
   (d) Train a classifier (Sec. 3.3)
2. For every training data point, compute the averaged class probabilities over all bootstraps $P(r \mid \boldsymbol{f}) = \frac{1}{B} \sum_b P_b(r \mid \boldsymbol{f})$.
3. Estimate the decision thresholds τ_{low} and τ_{high} (Sec. 3.4).
4. Collect statistics for every feature: (i) the *mean feature weights*, (ii) *the probability of a feature being selected* in the similar spirit to what introduced in, (iii) the *stability score*, which is the ratio of the absolute mean of the feature weight and its standard deviation, and (iv) the *importance*, which is the product of the mean feature weight and the standard deviation of the feature values across the training data.

The Step 1(a) is essentially the well-known procedure called "sampling the majority class" for handing the class imbalance problem, but we are not aware of the use within the context of bootstrapping. Thus at the end of the training phase, we have collection of B classifiers and a list of stable and predictive features, as well as the fully specified class-assignment rule.

At test time, the class probability is estimated as in Step 2, and the class assignment is carried out using procedure in Sec. 3.4.

4. IMPLEMENTATION AND RESULTS

4.1 Implementation

For robustness we consider items (e.g., codes) with more than 100 occurrences and are in the top $2,000$ most popular items of a given type. Other items that do not satisfy these conditions are considered rare events. Such rare events, though statistically less important individually, are critical in identifying risks if combined. We empirically find that using diagnostic features at level 3 in the ICD-10 hierarchy gave the best result as they appears to balance generality and specificity. Whenever appropriate, we also map diagnostic codes into the mental health grouping scheme known as MHDG[4].

We implement several kernel types and report here the results for Gaussian kernel filters, as they seem to work better than others (but similar to uniform kernels). Filter responses are then normalised into the range $[0, 1]$ before transformed by the sqrt(f) operators. We then apply feature selection described in Sec. 3.2 with control parameter: $\lambda_1 = 3 \times 10^{-4}$ unless specified otherwise. For cumulative and stagewise classifiers (Sec. 3.3.3 and Sec. 3.3.4), logistic distributions for the underlying random risks are used. The number of bootstrap is set as $B = 100$. The decision thresholds used in the class assignment rule in Sec. 3.4 are set at the 78*th* percentile and the 93*th* percentile, respectively.

We use 10-fold cross-validation *in the patient space*, that is, the set of unique patients is divided in to subsets of equal size. Models are trained on data for 9 subsets and tested on the other. The results are reported for all validation subsets combined. Note that this can be a stronger test than the cross-validation in the data space because it removes any potential patient-specific correlation (also known as *random-effects*).

We employ several performance measures: For general model fitting, the likelihood evaluated on validation sets provide a measure of how the model generalises to unseen data. For each outcome class, we use *recall* R – the portion of groundtruth class that is correctly identified; the *precision* P – the portion of identified

[4]MHDG stands for Mental Health Diagnosis Group. The mapping is available at http://www.health.gov.au

	Suicide (out of 41)	C_2				C_3				Resource cost (↑ %)	FN (↓ %)
		Cases	R(%)	P(%)	F_1(%)	Cases	R(%)	P(%)	F_1(%)		
Clinician rating	*14*	*338*	*23.5*	*11.7*	*15.6*	*70*	*8.1*	*12.9*	*10.0*	***3,445 (0)***	*1,535 (0)*
k-NN ($k = 100$)	29	423	29.4	15.6	20.4	262	30.5	23.3	26.4	3,827 (11)	1,227 (20)
Linear classifier	30	421	29.2	15.8	20.5	297	34.5	24.3	28.5	3,893 (13)	1,133 (26)
Cumulative	31	449	31.2	16.4	21.5	297	34.5	25.5	29.3	3,907 (13)	1,110 (28)
Linear→Cumul	31	433	30.1	16.1	21.0	314	36.5	26.1	30.5	3,889 (13)	1,129 (26)
Stagewise (Shared)	32	432	30.0	15.9	20.8	**318**	**37.0**	**27.0**	**31.2**	3,890 (13)	1,148 (25)
Stagewise (Multi)	31	438	30.4	16.0	21.0	290	33.7	25.6	29.1	3,869 (12)	1,129 (26)
Multinomial	31	**473**	**32.8**	**17.2**	**22.6**	289	33.6	23.9	27.9	3,960 (15)	**1,077 (30)**

Table 2: Performance of calibrated classifiers for predicting 3-month risks. R = Recall, P = Precision, in percentages. *FN* = false negatives, which are the risky cases wrongly classified as low-risk. Resource cost is the total number of cases assigned as moderate/high-risk. *Linear→Cumul* means the outcome of the linear classifier is fed into a cumulative ordinal regression model to compute the correct class probabilities. The symbols ↑ and ↓ denote the amount increase or decrease relative to the reference figures by clinicians.

class that is actually correct; and the *F-score* – their harmonic mean $F_1 = 2RP/(R+P)$.

4.2 Results

4.2.1 Outcome Prediction

We first evaluate the predictive power of the mandatory risk assessments being performed by Barwon Health. Using the overall assessment (risk ratings of 3 and 4 are high-risk, 2 moderate-risk, and ratings of 1 and 0 are low-risk), the performance on the high-risk class for 3 month horizons is quite poor: $R = 8.1\%$, $P = 12.9\%$, $F_1 = 10.0\%$. There are 14 suicide cases (34%) detected from the C_2 and C_3 assignments. Tab. 2 lists more details. Machine learning algorithms significantly outperform the mental health professionals to a large margin. For moderate-risk prediction, the F_1-score by machines ranges from 20.4% to 22.6%, which are $31\% - 45\%$ improvement over the score by clinicians. The differentials are even better for the high-risk class: the improvement are between 164% to 212%. In terms of suicide detection, the machine detects $29 - 32$ cases, which are more than twice the number detected by human (14 cases).

The practical significance of the difference is remarkable. Assuming for simplicity that the management cost, on average, is similar for both the moderate and high risk cases, then the total cost when predicting by human is $3,445$ resource units. There are $1,535$ cases are misclassified as low-risk (they are false negative, and thus left untreated). The machine algorithms typically cost slightly higher than human but with less false negatives. For example, the stagewise model with shared parameters (Sec. 3.3.4) leads to $3,890$ resource units (13% higher than those by clinicians), but with $1,148$ false negatives (25% lower than those by clinicians). The significance may be amplified when considering that the social cost for false negatives is much more serious than hospital resources.

4.2.2 Risk Factors

Excepts for the k-NN classifiers which do not have built-in selection mechanism, all other classifiers are capable of fine-tuning the features selected from the previous step (Sec. 3.2).Under the ℓ_1-norm regularisation schemes within the bootstrap framework, only *few percents* of strong and stable features are kept.

Class-independent features. Linear, cumulative and stagewise models (with shared parameters) do not distinguish the parameters between classes, and thus we have a single list of features at the end of the training phase. Tab. 3 presents top 20 features ordered by their importance (see Sec. 3.5), as produced by the cumulative model (Sec. 3.3.3). Predictive features include: Recent emergency visits, recent high-risk attempts (C_3), moderate-risk attempts (C_2 & self-poisoning) within 12 months, recent history of mental problems and of drug abuse, socioeconomic problems (pensioner, frequent home moving). Although these risk factors are known [7, 4], our discovered factors are more precise in timing.

Class-specific features. Class-specific models such as the stagewise model with class-specific parameters and the multinomial model can offer re-ranking of features for C_2 and C_3 separately. Tabs. 4 list top-ranked class-specific features for C_2 and C_3, respectively, under the stagewise models. A noticeable aspect is the strong association between prior C_3 attempts with future C_3 outcomes.

5. DISCUSSION

We have proposed an integrated computational framework for suicide risk prediction. The framework has three components: temporal feature extraction, an ensemble loop for feature selection and ordinal classifiers, and risk calibration. The key innovative aspect of the paper lies in its representation of the patient clinical history as a temporal event image, from which time-dependent features are generated by applying a bank of multiscale one-sided convolutional filters. Risk-bearing features are then selected by training an extreme-value classifier equipped with ℓ_1-norm regularisations. Using the proposed framework, we have presented a thorough study on a cohort of mental health patients from a large regional hospital. The results demonstrate that the framework outperforms risk assessment instruments by medical practitioners in terms of predictive power.

This project started with the goal of predicting suicide. However, we soon realised that this was an impossible task due to the rarity of suicide while there are many possible risk factors, none of which are really strong. This difficulty actually resembles the long-standing conjecture in the mental health literature [11, 9]. While the existing literature focuses instead in predicting suicide attempts, for practitioners the high-risk attempts are those we should pay extra attention to. And thus one of the contributions of this study is the separation of the attempts into those moderate-risk (C_2) and high-risk (C_3).

As the time of this writing, the deployment is on-going. Since the data is readily available through Barwon Health's warehousing, a real-time clinician support system can be readily implemented with very minimal cost. There is no need for special hardware/software. As the cohort is relatively small by current machine learning standards, the feature extraction and model training are relatively fast. Our prototype implementation on a standard PC using SQL Sever and Perl typically takes a couple of minutes to extract features for

Feature	$(\sigma_k; s_k)$	*Importance*	*Stability*	*Sel.Pr.*
Number of EDs	$(0.5; 0)$	99.1	3.0	1.00
Number of EDs	$(3; 0)$	93.3	3.2	1.00
High-lethality attempts (C_3)	$(3; 0)$	85.3	2.5	0.94
ICD code: *Z29* (Need for other prophylactic measures)	$(3; 0)$	72.7	3.2	1.00
Number of EDs	$(6; 6)$	62.4	2.1	0.96
Number of postcode changes & Male	$(6; 0)$	60.0	1.9	1.00
Moderate-lethality attempts (C_2)	$(6; 6)$	56.9	2.9	0.96
Number of EDs	$(1; 0)$	52.4	3.6	1.00
Moderate-lethality attempts (C_2)	$(12; 12)$	48.4	2.3	0.96
ICD code: *F19* (Mental disorders due to drug abuse)	$(6; 6)$	46.6	2.2	0.96
Marital status: single/never married & Male	*NA*	42.1	1.2	0.82
ICD code: *F33* (Recurrent depressive disorder)	$(0.5; 0)$	41.6	1.6	0.80
ICD code: *F60* (Specific personality disorders)	$(3; 3)$	39.3	1.6	0.76
ICD code: *T43* (Poisoning by psychotropic drugs)	$(3, 0)$	38.5	1.3	0.82
ICD code: *U73* (Other activity)	$(3, 0)$	35.5	1.5	0.92
Occupation: pensioner & Male	*NA*	33.2	1.2	0.86
Number of postcode changes & Female	$(12, 12)$	27.9	1.5	0.92
ICD Code: *T50* (Poisoning)	$(3, 0)$	25.8	1.7	0.90
Marital status: single/never married & Female	*NA*	25.5	0.9	0.74
Number of EDs	$(12, 12)$	25.1	1.4	0.90

Table 3: Predictive and stable features associated with risky outcomes in the next 3 months, ranked by *importance*, as produced by cumulative models (Sec. 3.3.3). The Gaussian kernel width σ_k and the delay s_k are measured in months; *Sel. Pr.* = selection probability.

about 10 thousands patients. The same amount of time is needed for model building in Matlab, while prediction is unnoticeable by users. The model needs to be retrained periodically as new data flowing in, e.g., every month. The front-end that interacts with clinicians is being developed – this will offer easy browsing of risk history (through the predictive and stable risk factors which have been discovered by our models), alerting risk and predicting future outcomes.

The main challenge faced in deploying the solution would be earning trust from clinicians in their daily work-flow. We anticipate that the initial resistance will be significant as the implication of taking the advice from the machine is profound for professionals. The next phase of this research is consulting with physicians and psychologists on how to best present the results and explain the reasoning behind the prediction. Another issue is the interaction between the physicians and the system: If the physicians modify their treatment strategy based on the machine prediction, then the outcome will be altered, leading to the poorer match between the actual outcome and the predicted.

The framework introduced in this paper is generalisable as the information extracted from the data warehousing is standardised, e.g., using the ICD-10 coding system and Mental Health Diagnosis Group mapping, and the models make no use of local expertise (such as risk assessments). The main limitation is that the research is based on the cohort at Barwon Health alone, and thus local characteristics of the population and the practice may bias the prediction. Finally, the pipeline of feature extraction, selection and classifier is in fact general and thus can readily be applicable for many types of risks with very minimal effort. This has been validated on a series of other predictive problems: The risk of hospitalisation in diabetes, COPD, mental health, heart failure, heart attack and pneumonia, and of mortality in cancers, all at Barwon Health demonstrating the versatility.

Acknowledgments

We thank Ross Arblaster and Ann Larkins for helping data collections, Paul Cohen for providing management support for the project, and the three reviewers for helpful comments.

6. REFERENCES

[1] T. Amemiya. Qualitative response models. *Annals of Economic and Social Measurement*, 4(3):363–372, 1975.

[2] B. Bondy, A. Buettner, and P. Zill. Genetics of suicide. *Molecular psychiatry*, 11(4):336–351, 2006.

[3] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] G.K. Brown, A.T. Beck, R.A. Steer, and J.R. Grisham. Risk factors for suicide in psychiatric outpatients: A 20-year prospective study. *Journal of Consulting and Clinical Psychology*, 68(3):371, 2000.

[5] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, 1(1):54–75, 1986.

[6] EJ Gumbel. *Statistical of extremes*. Columbia University Press, New York, 1958.

[7] Keith Hawton, Daniel Zahl, and Rosamund Weatherall. Suicide following deliberate self-harm: long-term follow-up of patients who presented to a general hospital. *The British Journal of Psychiatry*, 182(6):537–542, 2003.

[8] A.K. Johnston, J.E. Pirkis, and P.M. Burgess. Suicidal thoughts and behaviours among Australian adults: findings from the 2007 National Survey of Mental Health and Wellbeing. *Australian and New Zealand Journal of Psychiatry*, 43(7):635–643, 2009.

[9] M. Large and O. Nielssen. Suicide is preventable but not predictable. *Australasian Psychiatry*, 20(6):532–533, 2012.

[10] M. Large, C. Ryan, and O. Nielssen. The validity and utility of risk assessment for inpatient suicide. *Australasian Psychiatry*, 19(6):507–512, 2011.

[11] M.M. Large and O.B. Nielssen. Suicide in Australia: meta-analysis of rates and methods of suicide between 1988 and 2007. *Medical Journal of Australia*, 192(8):432–437, 2010.

[12] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142, 1980.

Feature	$(\sigma_k; s_k)$	*Importance*	*Stability*	*Sel.Pr.*
Moderate-risk class (C_2)				
Number of EDs	$(0.5; 0)$	100	4.0	1.00
Number of EDs	$(3; 0)$	97.0	3.6	1.00
ICD code: *Z29* (Need for other prophylactic measures)	$(3; 0)$	93.3	4.2	1.00
Moderate-lethality attempts (C_2)	$(3; 0)$	68.4	3.5	1.00
Moderate-lethality attempts (C_2)	$(6; 6)$	63.6	3.9	1.00
Number of postcode changes & Male	$(6; 0)$	60.5	2.0	0.99
Number of EDs	$(1; 0)$	59.7	4.4	1.00
Moderate-lethality attempts (C_2)	$(12; 12)$	59.5	2.3	0.97
Number of EDs	$(6; 6)$	55.8	1.9	0.97
Marital status: single/never married & Female	*NA*	49.2	1.3	0.92
High-risk class (C_3)				
High-lethality attempts (C_3)	$(3; 0)$	100	2.9	0.93
ICD code: *T43* (Poisoning by psychotropic drugs)	$(3, 0)$	37.3	1.6	0.80
Occupation: student & Female	*NA*	30.9	1.2	0.77
MHDG: *042* - Depressive episodes; bipolar disorders	$(3, 0)$	28.1	1.1	0.64
ICD Code 2W: *F33* (Recurrent depressive disorder)	$(0.5; 0)$	22.0	1.9	0.90
Number of EDs	$(6; 6)$	21.7	1.5	0.96
ICD code: *F60* (Specific personality disorders)	$(3; 3)$	20.3	1.6	0.82
ICD code: *R45* (Symptoms and signs involving emotional state)	$(6; 6)$	16.8	1.2	0.72
Moderate-lethality attempts (C_2)	$(12; 12)$	15.6	1.3	0.94
ICD code: *U73* (Other activity)	$(3, 0)$	15.3	1.0	0.95

Table 4: Predictive and stable features associated with risk classes in the next 3 months, ranked by *importance*, as produced by the stagewise model *without* parameter sharing (Sec. 3.3.4). The Gaussian kernel width σ_k and the delay s_k are measured in months; *Sel. Pr.* = selection probability. *MHDG* = Mental Health Diagnosis Group.

[13] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & HallCRC, 1989.

[14] D. McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1973.

[15] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[16] G.E. Murphy. The prediction of suicide: Why is it so difficult? *American Journal of Psychotherapy*, 1984.

[17] F. Ruiz, I. Valera, C. Blanco, and F. Perez-Cruz. Bayesian nonparametric modeling of suicide attempts. In *NIPS*, 2012.

[18] C. Ryan, O. Nielssen, M. Paton, and M. Large. Clinical decisions in psychiatry should not be based on risk assessment. *Australasian Psychiatry*, 18(5):398–403, 2010.

[19] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[20] G. Tutz. Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11(3):275–295, 1991.

[21] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In *Proc. of the 18th SIGKDD*, pages 453–461. ACM, 2012.