Moderating the Influence of Current Intention to Improve Suicide Risk Prediction

Nawal A. Zaher, MSc[1], Christopher D. Buckingham, PhD[2]
[1]AASTMT, Cairo, Egypt; [2]Aston University, Birmingham, United Kingdom

Abstract

When assessors evaluate a person's risk of completing suicide, the person's expressed current intention is one of the most influential factors. However, if people say they have no intention, this may not be true for a number of reasons. This paper explores the reliability of negative intention in data provided by mental-health services using the GRiST decision support system in England. It identifies features within a risk assessment record that can classify a negative statement regarding current intention of suicide as being reliable or unreliable. The algorithm is tested on previously conducted assessments, where outcomes found in later assessments do or do not match the initially stated intention. Test results show significant separation between the two classes. It means suicide predictions could be made more accurate by modifying the assessment process and associated risk judgement in accordance with a better understanding of the person's true intention.

Keywords: Suicide Risk, Suicide Intention, Clinical Risk Judgment, GRiST, Decision Support System

Introduction

Intention is a mental state that represents a commitment to carrying out an action or actions in the future[1]. It is one of the most important components for assessing the risk of a suicide attempt by mental health practitioners[2-6] but it is also difficult to measure. It relies on people being assessed giving accurate self reports and there are many reasons why this may not happen. The problem is not so much with those who indicate a degree of positive intention because this alerts the assessor to the need for exploring intention further. Intentions can be evaluated as representations of possible actions or plans set to achieve a goal[7]; realistic and established plans with steps taken towards the goal indicate increased underlying intention.

While it is reasonable to claim that the presence of current intention of suicide is an indication of risk, the absence of intention does not necessarily dictate the opposite[8]. Patients may, knowingly or unknowingly, hide their intentions, which raises reasonable doubt regarding their statement of intention.

The difficulty with people who say they have no intention is that validating their statement depends on an absence of evidence and this is difficult for assessors to validate. In effect, it stops further investigation into the current intention, diverts attention elsewhere, and can mean important information is lost. The overall risk judgement is compromised as well as a comprehensive understanding of the underlying symptoms and causes. The assessor's evaluation plan and course of action are affected, which results in a mismatch between interventions and outcomes[8-10]. It happens with a "no" answer because of its categorical nature, as opposed to a "yes" answer, which can be further investigated to attenuate its influence on risk judgements.

This study uses data from a web-based decision support system (DSS) called the Galatean Risk and Safety Tool, GRiST[11] to determine the reliability of people's expressed absence of suicide intention. The goal is to identify cues that would enable assessors to change how the absence of current intention is conceived. If it is possible to detect unreliable absence of intention, the assessment process can be changed to explore it further or, at the very least, the influence of the stated absence of intention on the assessor's judgement of suicide risk can be reduced.

The next section provides an overview of the nature of mental health risk data and how it is collected by GRiST. The paper then introduces an algorithm to classify people's negative intentions of suicide as being reliable or unreliable. The results are presented and followed by a discussion about how the algorithm could be incorporated in risk evaluation tools to improve their accuracy.

Background

Questions about patients' intentions direct assessors and DSSs towards the cues that are most relevant to the risk being assessed[12]; some or all of these could be masked if people have stated an absence of intention. It leads to two major problems. The first is missing data[13], where relevant data is not collected because the assessor thinks it is not needed[9,10,14]. The second concerns outcomes because people wrongly assessed as low risk may not receive required

interventions and can go on to make a suicide attempt. Assessments are made more difficult by the volatile nature of stated intentions, which is compounded by factors such as impulsiveness and deception or denial. In short, we lack insight into the intention of others[15]. The GRiST DSS attempts to penetrate the gloom by collecting a large amount of contextual and historical data, in addition to the immediate risk history and behaviour, so that a holistic picture of a person can be drawn. This is the data set used for modeling and evaluating current intention.

1- GRiST Mental Health Data

GRiST is a tool for helping practitioners assess and manage multiple risks associated with mental-health problems, including suicide, self-harm, harm to others, self-neglect and vulnerability. The tool represents expert consensus, elicited by preliminary interviews conducted with mental health practitioners[16] and refined through feedback from using the tool in practice[13,17]. The level of suicide risk is measured in GRiST using membership grades (MGs). They are generated from the person's assessent data into a common value between zero and one that represents each cue's risk input[12]. The relationship between input MGs and associated risk judgements given by assessors is learned across all assessments in the GRiST database to provide a model for predicting the risk judgement for a new assessment[12,13].

2- Current Intention of Suicide

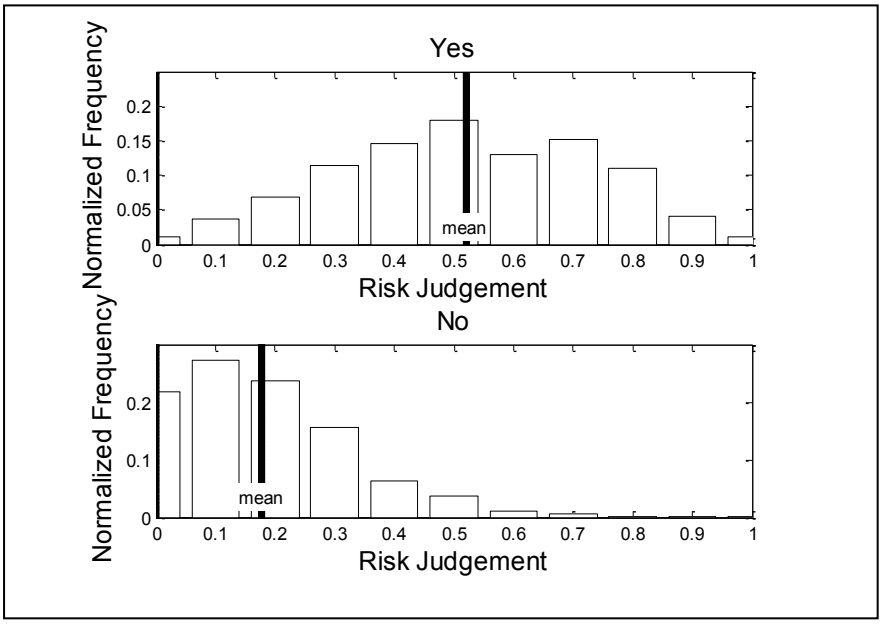

Figure 1. Distribution of clinical risk judgment for patients with current intention of "yes" (the top figure) and "no" (the bottom figure).

The clinical risk judgment distributions for GRiST assessments that have "yes" and "no" answers (Figure 1) show that the mean risk for the "yes" group is 0.5201 and for the "no" group is 0.1764. This is a significant clinical separation, with people being treated differently when having more than a three-point greater risk level, where 0 is minimum risk and 1 is maximum risk.

It could be argued that intention is more accurately predicted by attitude and behaviour[3,18], rather than by a simple YES/NO question but their ability to interpret the reliability of someone's lack of current intention has not been modeled. For GRiST, current intention is not binary for a "yes" answer because it opens up a number of additional questions to assess the degree of intention. In contrast, a "no" answer blocks off any further direct exploration of current intention and may also cause other factors to be masked because of a perceived reduction in risk. An indirect approach to evaluating the "no" answer is required and this paper pursues it by investigating whether any of the multiple GRiST cues can differentiate between the reliability of assessments' claimed lack of intention.

Proposed Algorithm

Although GRiST does not directly collect intervention and outcome information following an assessment, it does contain relevant data for those patients who have more than one assessment. An important piece of information recorded for all patients who have a history of suicide attempts is the date of the most recent attempt. If this has

changed between assessments then it means the person has carried out a new suicide attempt and is therefore an outcome we can use in our analysis.

This paper analyses all those assessments that had no current intention and are followed by a subsequent (repeat) assessment. The idea is that people with a new risk episode will have a less reliable absence of current intention than those who did not carry out an attempt between assessments. If the two groups of patients can be distinguished, it should be possible to determine a reliability measure for the negative answer and a principled way of adjusting the assessment process accordingly.

1- Defining the Classes

The parent population for the study sample is 71,024 assessments, consisting of 27,947 patients, of which 12,595 have repeated assessments. Of these, only 6,502 assessments have a current intention answer of "no" in an assessment which also has a subsequent one (i.e. it must not be the most recent assessment). The reliability of the answer is crudely evaluated by inspecting the subsequent assessment to find out whether this patient has had a repeat episode since the assessment, indicated by a new date for the most recent attempt. If so, the answer to current intention is marked as unreliable. This produces two classes: those people with a reliable negative intention, Class Rno, and those with an unreliable answer of "no", Class Uno. The size of Class Rno is 4,458 assessments, and the size of Class Uno is 2,044 assessments.

2- Feature Selection

Having established the two classes, the task is to learn the defining features that distinguish them. The Maximum Likelihood Estimate (MLE)[19] of the mean and variance of each feature in each class is calculated assuming a Gaussian distribution. MLE provides an interpretable parametric form and has a significantly low computational cost compared to other parametric techniques like Bayesian estimation[20]. Then the normalized Euclidean distance[21] between the two classes is calculated for each feature to produce the distance vector shown by Equation 1

$$\hat{d} = \sqrt{\frac{|(\hat{\mu}_{Uno} - \hat{\mu}_{Rno})|^2}{\hat{\sigma}^2_{Uno}\hat{\sigma}^2_{Rno}}} \quad (1)$$

where $\hat{\mu}_{Rno}$ and $\hat{\mu}_{Uno}$ are the average assessments of class Rno and class Uno respectively and $\hat{\sigma}^2_{Rno}$ and $\hat{\sigma}^2_{Uno}$ are the variances of Rno and Uno. Each element of the distance vector represents the distance between Rno and Uno for a particular cue.

The top 20 cues arranged in descending order of the distance measure, d (Table 1), are appealing candidates for the feature vector. Although these produce the largest class separation, some will be missing from assessments if the assessors do not ask about them due to absence of intention. The table gives the number of occurrences of each of the cues in the dataset and the percentage of assessments where they are missing. The cues with the least amount of missing data and the most influence are the first seven cues. If only these are used, the classifier will have a larger sample and be more reliable because only assessments with data for all the cues in the vector are used for the estimation of model parameters and for testing. These cues all happen to be related to previous history of suicide, which now becomes a selection criterion for the unreliability measure. It means a different measure will need to be calculated for other subsets of the population that do not have a history of suicide.

It is encouraging, but not surprising, that the most influential cues are also the most commonly collected. GRiST was built on the pooled expertise of mental-health practitioners[16] and one would expect them to ensure they identify and collect the most important data for assessments. The literature also confirms the importance of cues related to past history for predicting suicide risk[4,13,22,23].

Table 1. The top 20 cues listed in order of how well they separate the reliability classes, given by their distance measures. The number of occurrences of the cue is within the sample of 6,502 assessments, which is also given as a percentage of assessments with the cue missing.

No.	Cue	Distance	Occurrences	Missing (%)
1	Date of most recent suicide attempt	10.16	6276	3.50%
2	How much did the person want to succeed	2.70	5873	9.70%
3	Chance of discovery after suicide attempts	2.60	5824	10.4%

4	Potential lethality of suicide method	2.59	5713	12.3%
5	Regret about trying to commit suicide	2.55	6130	5.70%
6	Insight into lethality of previous suicide attempts	2.41	6015	7.50%
7	Suicide attempts escalating in frequency	1.78	5443	16.3%
8	How many suicide attempts	1.14	3965	39.0%
9	Likelihood of acting on delusions	1.08	3230	50.3%
10	Potential triggers match previous triggers	0.98	2314	64.4%
11	Stage of depression	0.95	1818	72.0%
12	Number of dependents sharing accommodation	0.93	6308	3.00%
13	Mania/hypomania	0.89	3390	47.8%
14	Life not worth living	0.84	3252	50.0%
15	Capacity to cope with major life stresses	0.83	3028	53.4%
16	Habitable accommodation	0.81	2046	68.5%
17	General motivation in life	0.79	1013	84.4%
18	Plans for the future	0.74	3287	49.4%
19	Potential triggers of suicide	0.73	2859	56.0%
20	Impulsiveness	0.72	2981	54.1%

3- Classifier Model

There are many approaches to binary classification, the most prominent of which is logistic regression[24]. Although it is powerful and appropriate, we decided against it for our approach because the heuristic method for choosing features was not designed to produce regression vectors. It could have been achieved using exhaustive search but it would have been limited to samples with all features present in the vector and over fitting may have occurred due to selecting features based solely on classification results.

Instead, features are independently chosen using a fully parametric probabilistic model to generate the conditional density functions on the assumption that each feature has a Gaussian distribution. Features do not all need to be in the same patient vector, which means selection can include any assessments where they individually occur, making the sample size for selecting each one larger than would be the case for regression.

The classes are represented by two Gaussian distributions, with means and variances calculated using MLE[19] for the top seven cues shown above (Table 1). The classification decisions are based on a minimum error rate threshold in each dimension[20]. An example for the classifier model for one dimension: the most recent suicide attempt date, which is the top cue, is shown below (Figure 2). In each dimension the classifier needs a threshold to separate the two classes. The optimum threshold ρ that minimizes the probability of error is the MG value at the point of intersection of the two Gaussian distributions representing the two classes[25] (Figure 1). Equating the two sides and substituting for ρ:

$$\frac{1}{\sqrt{2\pi\sigma_{Uno}^2}} e^{-\frac{(\rho-\mu_{Uno})^2}{2\sigma_{Uno}^2}} = \frac{1}{\sqrt{2\pi\sigma_{Rno}^2}} e^{-\frac{(\rho-\mu_{Rno})^2}{2\sigma_{Rno}^2}} \quad (2)$$

Simplifying the expression yields Equation 3 with the coefficients a_2, a_1 and a_0 given by Equations 4, 5, and 6 respectively. The threshold ρ that minimizes the probability of error in each dimension is calculated for each dimension individually by solving Equation 3.

$$a_2\rho^2 + a_1\rho + a_0 = 0 \quad (3)$$

$$a_2 = \sigma_{Rno}^2 - \sigma_{Uno}^2 \tag{4}$$

$$a_1 = 2(\mu_{Rno}\sigma_{Uno}^2 - \mu_{Uno}\sigma_{Rno}^2) \tag{5}$$

$$a_0 = \mu_{Uno}^2\sigma_{Rno}^2 - \mu_{Rno}^2\sigma_{Uno}^2 - 2\sigma_{Uno}^2\sigma_{Rno}^2 \ln\left(\frac{\sigma_{Rno}}{\sigma_{Uno}}\right) \tag{6}$$

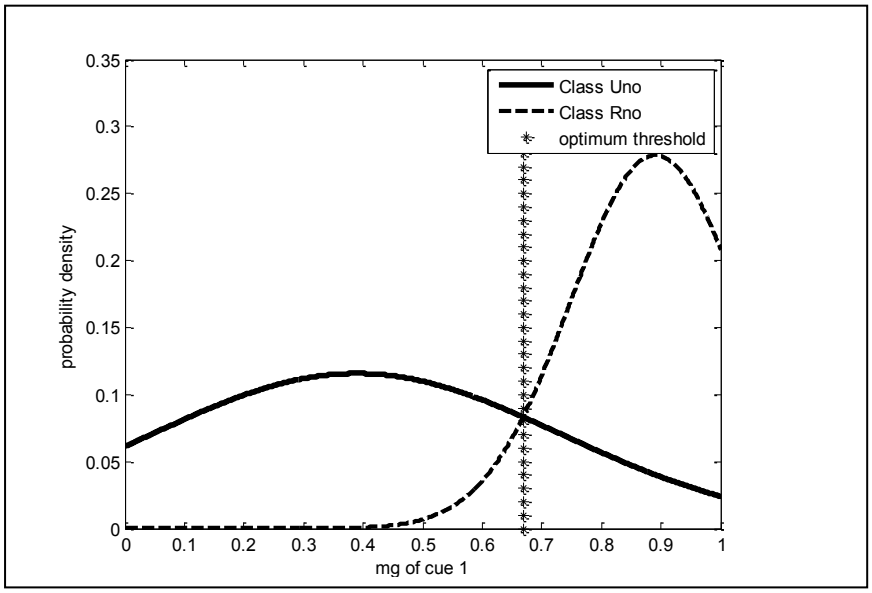

Figure 2. Classifier model for Cue 1, "date of most recent suicide attempt", showing the Gaussian distribution for each class, Uno and Rno (unreliable and reliable no intention respectively), and the decision threshold for this dimension, based on the cue's membership grade, MG.

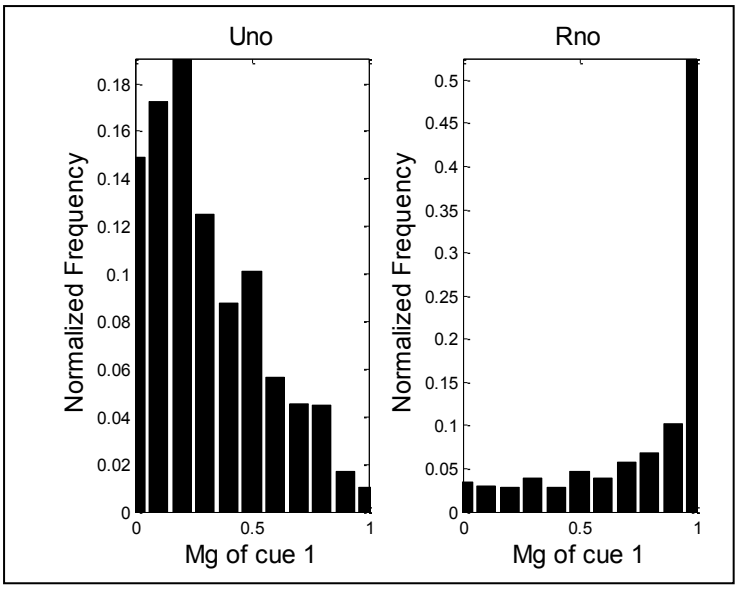

Figure 3. Distribution of MG values for cue 1, "most recent suicide attempt", RHS: MG values of cue 1 for class Uno, LHS: MG values of cue 1 for class Rno.

The idea is to find the input MG for each of the seven cues that split the classes most accurately. Each cue will independently indicate the most likely class by whether the assessment MG for the cue is above or below the threshold value.

The natural distribution of data shown (Figure 3) supports the Gaussian assumption adopted. Nevertheless, considering more skewed models may be of interest, but will add some complexity to parameters and threshold computations.

4- Decision Fusion

The last stage of the classification process is combining the decisions from all 7 cues (dimensions) into one decision as to whether an assessment belongs to Class Rno or Uno. This can be done by voting over the decisions of each cue or summing their probabilities that the assessment is in one class or the other[26]. This paper combines the two methods because a majority vote among the hard decisions taken by each dimension does not take into account the accuracy of each dimension.

First a hard decision is made for each cue by comparing the value of the cue's MG to the threshold. The cue's vote is then weighted based on how well it separates the classes. This is determined by calculating its positive predictive value (PPV) using Equation 7

$$ppv = tpr/(tpr + fpr) \tag{7}$$

where TPR is the true positive rate and FPR is the false positive rate. The PPV shows how powerful each dimension (cue) is individually, with a higher PPV signifying more separation of the two classes in that dimension.

The PPV value for each cue is used to weight the overall decision obtained from its corresponding dimension as in Equation 8

$$v = \sum_{i=1}^{7} ppv_i \, d_i \tag{8}$$

where ppv_i is the positive predictive power of dimension i and d_i is a Boolean variable that represents the decision from each cue, which is given a value of '1' for class Uno and a value of '0' for class Rno.

The output of the fusion process, v, is compared to a threshold v_{th} such that if $v > v_{th}$, then the decision is Rno, otherwise the decision is Uno. To choose v_{th}, the Receiver Operating Characteristics (ROC) curve is plotted using different values for v_{th} so that the point with the maximum accuracy is chosen. The figure below (Figure 4) shows the ROC curve of the best performing model and its distance from the line of no discrimination.

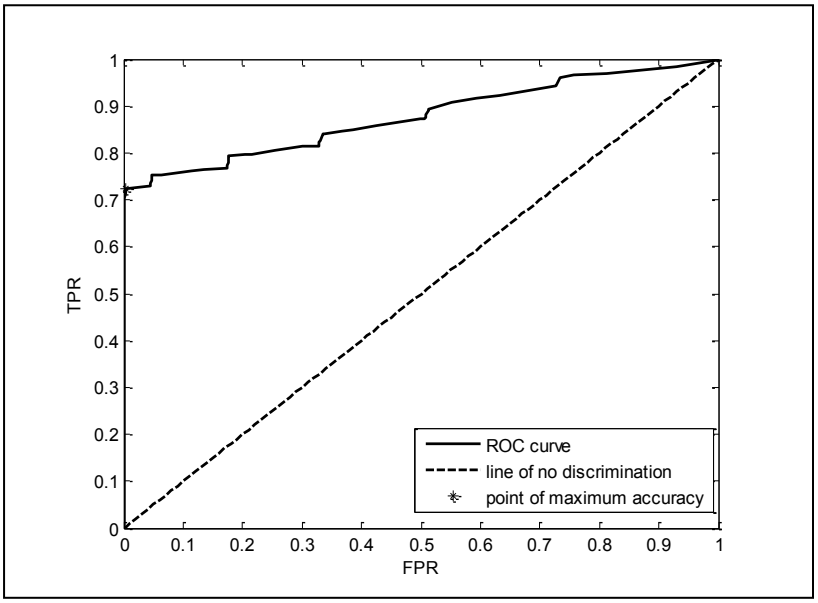

Figure 4. ROC curve showing TPR against FPR for different decision thresholds.

Results

After filtering out assessments with incomplete vectors (i.e. where one or more of the top seven cues in Table 1 are missing), the size of Class Rno is 1,154 and Class Uno is 1,903. The model was tested by combining the two classes into one group and then applying 10-fold cross validation[21], where the group is split into 10 equal-sized sets of randomly selected assessments. Each set is used once for testing with the remaining 9 sets used for training to give 10 sets of test results that were pooled.

1- Accuracy

The average accuracy of the 10 splits is 0.847 while the percentage values of TPR, FPR, false negative rate (FNR) and true negative rate (TNR) at different setups are shown below (Table 2). The point with the highest accuracy is not necessarily the best operating point. It depends on which of the cells needs to be maximized or minimized. For the sake of risk assessment, the TPR is probably the most important but the FPR is also significant because false alarms would erroneously impact on the assessment process as well as exaggerating the risk and potentially triggering unnecessary interventions.

Table 2. Classifier statistics showing average performance, best performing split and worst performing split for Class Rno. TPR is the true positive rate, or hits, FPR is the false positive rate or false alarms, TNR is the true negative rate, or hits for Class Uno, and FNR is the false negative rate or misses.

Performance	Accuracy	TPR	FPR	TNR	FNR
Average	84.07%	72.55%	4.41%	95.59%	27.45%
Best	89.98%	72.62%	2.02%	97.98%	27.38%
Worst	79.51%	60.01%	11.24%	88.76%	39.99%

2- Chi-square test

To verify the significance of the results a Chi-square test was performed to check if the classification decision (Rno or Uno) is dependent on whether the person subsequently repeated a suicide attempt or not. The sample sizes on which the test was performed and the results of the classification at the minimum FPR are shown below (Table 3). The test gives $\chi^2 = 886.5033$ which equates to a probability very close to 0: the classifier deviates from chance at the $p < 0.01$ significance level which means the decision classes are dependent on repeated episodes.

Table 3. Classifier results at maximum accuracy. Rno predicts no repeat episodes and Uno predicts repeat episodes.

	Uno	Rno	Total
Repeat	407	170	577
No Repeat	7	945	952
Total	414	1115	1529

The most encouraging feature is that the unreliable intention decision has very few errors: only seven out of 407, which means it is not generating false alarms that would dilute the effect of warning practitioners. Although the assessments classified as reliable have many more errors, where people go on to repeat, the status quo is to consider all assessments as reliable. Ideally, we would like to move as many as possible of these into the unreliable class but, for now, the real impact of the research is in demonstrating that unreliability can be meaningfully predicted so that practitioners take the warning seriously.

Conclusion

The results show that assessments where the answer to current intention of suicide is “no” can be divided into two classes, unreliable and reliable negative intention. A non-standard method was used due to the amount of missing data in the samples but future work would explore more conventional methods such as regression, support vector machines, and decision trees, for example. Either way, the classification is based on other cues collected during an assessment, which can be used to develop a reliability measure for current intention that smoothes the influence of intention on the clinical risk judgement. Instead of having a categorical “no intention”, a graded value for intention

could be generated where some negative intention statements could even be regarded as the opposite. Alternatively, the classification based on reliability may be incorporated as a risk factor on its own, or incorporated in some of the cues that are already included in the risk assessment.

Alerting assessors to unreliability of intention would, in itself, help improve the data collection process, by ensuring assessors pay attention to issues that would not normally be considered for low-risk people. The top seven cues (Table 1) used for our reliability measure have a significant percentage of missing data and our results strongly support the need for them to be collected as a matter of course. However, they are all related to previous history of suicide and the next step is to determine whether a suitable reliability measure can be found for people without a history. The non-history cues from Cue 9 (Table 1) suggest this will be possible because they are able to separate the classes; a reliability indicator is needed because the table shows the cues are missing from a high percentage of assessments and should certainly be collected for those patients without a history of suicide. The upshot will be a measure applicable for all populations that will increase assessors' understanding of suicide risk and help prevent future attempts.

Acknowledgement

This work was part supported by Grant SRG-0-060-11 awarded to C.D. Buckingham from the American Foundation for Suicide Prevention. The content is solely the responsibility of the authors and does not necessarily represent the official views of the American Foundation for Suicide Prevention.

References

1. Bratman M. Intention, plans, and practical reason. The Center for the Study of Language and Information Publications; 1999.
2. Raue PJ, Ghesquiere AR, Bruce ML. Suicide risk in primary care: identification and management in older adults. Current psychiatry reports. 2014;16(9):1–8.
3. World Health Organization. Preventing suicide: a resource for non-fatal suicidal behaviour case registration. 2014.
4. Rezaei-Yazdi A, Buckingham CD. Understanding data collection behaviour of mental health practitioners. Studies in Health Technology and Informatics. 2014;207:193–202.
5. Huang SF, Lu CH, Ju CL, Lan JT, Chang CW, Chang CL, et al. The benefit of clinical psychologists in prevention from the suicide in one hospital in Taiwan, Republic of China. Life Science Journal. 2015;12(3).
6. Purushothaman P, Premarajan KC, Sahu SK, Kattimani S, et al. Risk factors and reporting status for attempted Suicide: A hospital-based study. International Journal of Medicine and Public Health. 2015;5(1):45.
7. Cohen PR, Levesque HJ. Intention is choice with commitment. Artificial intelligence. 1990;42(2):213–261.
8. Anderson ME, Myhre MR, Suckow D, McCabe A. Screening and Assessment of Suicide Risk in Oncology. Handbook of Oncology Social Work: Psychosocial Care for People with Cancer. 2015;p. 147.
9. Buckingham CD, Adams AE. Classifying clinical decision making: a unifying approach. Journal of Advanced Nursing. 2000;32(4):981–989.
10. Buckingham CD, Adams AE. Classifying clinical decision making: interpreting nursing intuition, heuristics, and medical diagnosis. Journal of Advanced Nursing. 2000;32(4):990–998.
11. GRiST. Galatean Risk and Saftey Tool; 2016 (accessed June, 2016). www.egrist.org.
12. Buckingham CD. Psychological cue use and implications for a clinical decision support system. Medical Informatics and the Internet in Medicine. 2002;27(4):237–251.
13. Saleh SN, Buckingham CD. Handling varying amounts of missing data when classifying mental-health risk levels. Studies in Health Technology and Informatics. 2014;207:92–101.
14. Berner ES, La Lande TJ. Overview of clinical decision support systems. In: Clinical decision support systems. Springer; 2007. p. 3–22.
15. Sch¨achtele S, Gerstenberg T, Lagnado D. Beyond outcomes: The influence of intentions and deception. In: Proceedings of the 33nd annual conference of the cognitive science society. Cognitive Science Society; 2011. p. 1860–1865.
16. Buckingham CD, Adams AE, Mace C. Cues and knowledge structures used by mental-health professionals when makingrisk assessments. Journal of Mental Health. 2008;17(3):299–314.
17. Buckingham CD, Ahmed A, Adams A. Designing multiple user perspectives and functionality for clinical decision support systems. In: Ganzha M, Maciaszek L, Paprzycki M, editors. Proceedings of the 2013 Federated Conference on Computer Science and Information Systems. IEEE; 2013. p. 211–218.
18. Ajzen I. Theory of Planned Behavior. Handb Theor Soc Psychol Vol One. 2011;1:438.

19. Agresti A, Kateri M. Categorical Data Analysis. Springer; 2011.
20. Duda RO, Hart PE, Stork DG. Pattern Classification. John Wiley & Sons; 2012.
21. Kruskal JB. Nonmetric multidimensional scaling: a numerical method. Psychometrika. 1964;29(2):115–129.
22. Swann AC, Dougherty DM, Pazzaglia PJ, Pham M, Steinberg JL, Moeller FG. Increased impulsivity associated with severity of suicide attempt history in patients with bipolar disorder. American Journal of Psychiatry. 2005;162(9):1680–1687.
23. Trakhtenbrot R, Gvion Y, Levi-Belz Y, Horesh N, Fischel T, Weiser M, et al. Predictive value of psychological characteristics and suicide history on medical lethality of suicide attempts: A follow-up study of hospitalized patients. Journal of Affective Disorders. 2016;199:73 – 80.
24. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. vol. 398. John Wiley & Sons; 2013.
25. Huang K, Yang H, King I, Lyu MR, Chan L. The minimum error minimax probability machine. The Journal of Machine Learning Research. 2004;5:1253–1286
26. Kittler J, Alkoot FM. Sum versus vote fusion in multiple classifier systems. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2003;25(1):110–115.