Journal of Affective Disorders 196 (2016) 218–224

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

# Optimizing the assessment of suicidal behavior: The application of curtailment techniques

Derek P. de Beurs [a,*], Marjolein Fokkema [b], Rory C. O'Connor [c]

[a] Netherlands Institute of Health Services Research (NIVEL), Utrecht, The Netherlands
[b] Department of Methods and Statistics at the Faculty of Social Sciences, Leiden University, The Netherlands
[c] Suicidal Behavior Research Laboratory, Institute of Health & Wellbeing, University of Glasgow, United Kingdom

ARTICLE INFO

*Article history:*
Received 28 August 2015
Received in revised form
14 January 2016
Accepted 6 February 2016
Available online 23 February 2016

*Keywords:*
Suicide ideation
Curtailment
Cut-off
Screener
Hospital-treated suicide attempters

ABSTRACT

*Background:* Given their length, commonly used scales to assess suicide risk, such as the Beck Scale for Suicide Ideation (SSI) are of limited use as screening tools. In the current study we tested whether deterministic and stochastic curtailment can be applied to shorten the 19-item SSI, without compromising its accuracy.
*Methods:* Data from 366 patients, who were seen by a liaison psychiatry service in a general hospital in Scotland after a suicide attempt, were used. Within 24 h of admission, the SSI was administered; 15 months later, it was determined whether a patient was re-admitted to a hospital as the result of another suicide attempt. We fitted a Receiver Operating Characteristic curve to derive the best cut-off value of the SSI for predicting future suicidal behavior. Using this cut-off, both deterministic and stochastic curtailment were simulated on the item score patterns of the SSI.
*Results:* A cut-off value of SSI $\geq$ 6 provided the best classification accuracy for future suicidal behavior. Using this cut-off, we found that both deterministic and stochastic curtailment reduce the length of the SSI, without reducing the accuracy of the final classification decision. With stochastic curtailment, on average, less than 8 items are needed to assess whether administration of the full-length test will result in an SSI score below or above the cut-off value of 6.
*Limitations:* New studies using other datasets should re-validate the optimal cut-off for risk of repeated suicidal behavior after being treated in a hospital following an attempt.
*Conclusions:* Curtailment can be used to simplify the assessment of suicidal behavior, and should be considered as an alternative to the full scale.



## 1. Introduction

Suicidal behavior is a major public health problem, accounting for 804,000 deaths per year (World Health Organisation, 2014). Clinical guidelines on suicidal behavior highlight the importance of assessing the risk of suicidal behavior (Jacobs and Brewer, 2004; van Hemert et al., 2012; Wasserman et al., 2012). This also applies for patients treated at emergency departments after a suicide attempt. Although an earlier suicide attempt has been shown to be the best predictor of future suicidal behavior (Hawton and van Heeringen, 2009), the assessment of suicidal ideation after an attempt may help clinicians to better differentiate between patients with acute risk and a relatively low risk for future suicidal behavior. Given the stress, the great time pressure and the need for somatic treatment of patients at emergency departments, it can be difficult to assess suicidal behavior in these settings (Verwey et al., 2007). In the Netherlands it was found that of the 14,000 patients who presented at an emergency department after a suicide attempt, only 25% were seen by a hospital psychiatrist (Kerkhof et al., 2007). Also, more than half of the patients who were treated for self-harm in English hospitals left the hospital without any form of risk assessment (Friedman et al., 2006; Kapur et al., 2004).

A simple and efficient scale for risk assessment may help to improve the assessment of suicide risk in emergency departments. Using a short screener scale, patients with an elevated risk, for whom further, more thorough assessment is required, can be identified. Given their length, commonly used scales for suicide-risk assessment, such as the Beck Scale for Suicide Ideation (Beck et al., 1979) are of limited use as screening tools (De Beurs et al., 2014; Reeve et al., 2007; Smits et al., 2011; Spijker et al., 2014). However, with the application of modern psychometric techniques

*Abbreviations:* SSI, Beck Scale for Suicide Ideation; DC, deterministic curtailment; SC, stochastic curtailment; ROC, Receiver Operating Characteristic Curve; AUC, area under the curve; CAT, computerized adaptive testing; SD, standard deviation

* Corresponding author.
*E-mail address:* d.debeurs@nivel.nl (D.P. de Beurs).

http://dx.doi.org/10.1016/j.jad.2016.02.033
0165-0327/

such as computerized adaptive testing (CAT), it is possible to reduce the number of items to be administered, without reducing predictive accuracy (Reeve et al., 2007). Specifically, in a recent clinical study (De Beurs et al., 2014, 2015a, 2015b), it was shown that on average, four items from the SSI, instead of the full set of 19 items, were sufficient to classify patients as having an elevated risk for future suicidal behavior or not (i.e., with a cut-off value of SSI > 2). Although these findings for CAT are promising, it requires dedicated software that may not be readily available in clinical settings.

In the current paper, we use a technique to shorten tests that does not require a computer or dedicated software for its application, called curtailment (Finkelman et al., 2012; Fokkema et al., 2014). The rationale behind curtailment is somewhat more intuitive than that of CAT, because it depends on observed item scores only, and does not assume a latent variable underlying observed item scores (van der Linden and Hambleton, 1997).

Curtailment always needs a pre-established cut-off value for a scale. A cut-off value for a scale determines which score on that scale best classifies patients as at risk or not at risk. When the best possible cut-off value is established, curtailment can then be used to minimize the number of items that need to be administered, to decide whether a participant would score above or below the cut-off on the full scale, or in other words, should be classified as at risk or not at risk. Simply put, with curtailment, test administration is halted when responses to the remaining items can no longer change the final classification decision (at risk, or not at risk). By allowing for early stopping of item administration, curtailment shortens questionnaire administration. With deterministic curtailment (DC), item administration is stopped as soon as the responses to the remaining items cannot change the final classification decision. It is also possible to take a non-deterministic approach to curtailment, by deriving probabilities for each of the two classification outcomes, and stopping item administration as soon as the probability of one of the classification outcomes exceeds an a-priori selected threshold value. This is called stochastic curtailment (SC; Finkelman et al., 2011). Like DC, SC requires a cut-off value. In addition, for the stochastic part, a value for $\gamma$ (gamma) needs to be selected by the user. $\gamma$ is the threshold for the probability that classification under SC matches that of administration of the full-length test. Although SC requires some computing, it produces simple look-up tables, with stopping criteria for every item. So, in contrast to CAT, no software is needed for administration of the questionnaire, making the results much easier to implement in daily practice. Also, the order in which items are administered remains the same as in the original scale. As a result, any unforeseen effect of the order of items can be ruled out (Fokkema et al., 2014).

In the current study, we applied deterministic and stochastic curtailment to shorten one of the most commonly used questionnaires for suicide risk assessment, the Beck Scale for Suicide Ideation (SSI; Beck et al., 1979).

Data were collected for patients who were seen by a liaison psychiatry service in a general hospital in Scotland after a suicide attempt (O'Connor et al., 2015). To our knowledge, the only published cut-off score for the SSI comes from a 20-year prospective study among 6891 psychiatric outpatients (Brown et al., 2000). Outpatients with a baseline score of SSI ≥ 3 were seven times more likely to die by suicide than outpatients who scored less than 3 at baseline. However, this cut-off may be less appropriate in a population of patients treated for suicide attempts, as the baseline suicide ideation among those patients is likely to be higher than among psychiatric outpatients (Brown et al., 2000). Therefore, the first step in our study was to determine the best cut-off value in our sample. Next, we used the item responses and the selected cut-off value to assess the extent to which DC and SC allow for shortening of the SSI without reducing classification accuracy.

## 2. Methods

### 2.1. Participants

Data were used from a study on psychological predictors of repeat suicidal behavior in those who were admitted to a general hospital following a suicide attempt. Full details of the study are described elsewhere (O'Connor et al., 2015). In short, 432 patients who were seen by the liaison psychiatry service the morning after presenting to a single general hospital following a suicide attempt were invited to participate in the study. These patients also did not meet any of the exclusion criteria, namely: unfit to participate (e.g., actively psychotic), unable to give informed consent (e.g., medically unfit to give informed consent), participating in one of the other studies being conducted in the hospital, or who were unable to understand English. Approximately 10% of participants who were approached declined to take part (10.2%, N=44). At baseline, 388 patients were asked directly by a member of the research team whether they had intended to end their lives and they were only included in the sample if they confirmed this to be the case.

### 2.2. Suicidal ideation (Beck Scale for Suicide Ideation; SSI)

At baseline, suicidal thoughts were assessed via the Beck Scale for Suicide Ideation (Beck et al., 1979). The scale contains 19 items on suicidal thoughts and plans. Each item has three options, which are rated on a three-point scale from 0 to 2, a higher score indicating a higher level of suicidality. The total score of the SSI is determined by totaling the 19 items resulting in a range from 0 to 38. Cronbach's alpha for the current sample was .94 (O'Connor et al., 2015).

### 2.3. Repeated hospital admission at follow-up

At follow-up, using a national linkage database, we were able to determine whether a patient had been admitted to any Scottish hospital within 15 months of the index episode. The Information Services Division was able to link 96.4% of the initial sample (374/388). Two trained coders independently examined the extracts from the medical records to determine the presence or absence of suicide intent. They agreed that of 94 of the 101 patients who were admitted for self-harm, there was presence of suicide intent. So, in the original study, data was used for 367 participants who were linked, and for whom the researchers were able to find suicide intent data if they were readmitted to the hospital following an episode of self-harm.

### 2.4. Missing data

In the current study, data from 366 participants were used, who had < 5 missing values on the SSI. Among these 366 participants, 6 had 2 values missing, and 5 had 1 value missing. Those missing values were imputed by filling in participants' mean item score, rounded up to a whole number. All 94 patients that were admitted for self-harm and for whom a presence of suicide intent was found were part of the sample of 366 participants.

## 3. Statistical analyses

### 3.1. ROC analysis and cut-off value selection

To derive the best cut-off value of the SSI, we fitted a Receiver Operating Characteristic curve (ROC) with the package pROC (Robin et al., 2011) from the R environment. The ROC curve was fitted for the data from all 366 participants. The package calculates sensitivity (i.e. proportion of correctly classified positive observations) and specificity (i.e. the proportion of correctly classified negative observations) over the range of all possible cut-off values on a continuous test score (in this case, total SSI score). The area under the curve (AUC) calculation is central to the interpretation of an ROC curve, as it presents the predictive quality of the classifier. In general, the higher the area under the curve, the more accurately a test score can differentiate between classes. As the classification method is always binary (in our case: elevated risk for suicidal behavior vs. no elevated risk), the area under the curve needs to be $>0.5$ to add information above chance. The value of the AUC might improve if estimated according to potential moderating groups. Therefore, we compared the ROC curves estimated in subgroups of the total sample, as defined by the following moderators: gender, previous suicide attempts (no previous attempt or one versus two or more) and age (split by median age: age=36). We used the roc.test function from the pROC package to compare the roc curves.

To derive the optimal cut-off value on the SSI, we calculated the Youden index (Youden, 1950). The Youden index takes a value between 0 and 1 to summarize the performance of a diagnostic test (Youden, 1950). It is calculated as $J = \text{sensitivity} + \text{specificity} - 1$. In the current study, this means summing the proportion of repeated attempters who were correctly classified as having an elevated risk of future suicidal behavior (sensitivity) with the proportion of patients who did not make a repeated suicide attempt, and who were correctly classified as not having an elevated risk (specificity). A value of $J=0$ means the test is useless (i.e., the test does not perform better than random classification) and a value $J=1$ indicates that all participants were accurately classified. In calculating $J$, we set the cost of a false negative classification at twice the cost of a false positive classification. The cut-off value that maximized the value of $J$ was taken as the optimal cut-off value.

To construct 95% confidence intervals for the estimates of the AUC, sensitivity, specificity and J, we used the bootstrap method, with non-parametric stratified resampling and the percentile method, as described in Carpenter and Bithell (2000). The number of bootstrap replications was set to 10,000.

To examine the predictive value of each item of the SSI on the observed outcome (re-attempt), we estimated the odds ratio for each separate item using logistic regression. We also compared the predictive validity of the binary cut score versus full-length test for re-hospitalization, including when accounting for gender, baseline depression and hopelessness.

### 3.2. Simulation design

Both DC and the SC were simulated on the item score patterns of the 366 participants with a complete SSI. An R package was created, based on the descriptions of Finkelman et al. (2012) to simulate DC and SC. The code is available via the GitHub repository of the second author, https://github.com/marjoleinF/curtail. The $\gamma$-values for SC were set to 0.9, as Finkelman et al. (2012) have found that gamma values of .90 provided substantial reductions in test length, without reductions in classification accuracy. In addition, Fokkema et al. (2014) have found that gamma values as low as .75 do not reduce predictive accuracy.

### 3.3. Classification accuracy

As shown by for example, Hastie et al. (2009) and Stone (1974), employing the same dataset for both calibration and evaluation of a model results in overly optimistic estimates of performance. Therefore, after deriving the optimal cut-off value for the SSI, and deriving tables for DC and SC, using the full dataset, we assessed efficiency and predictive accuracy by means of 10-fold cross validation (CV). This means that we randomly partitioned the original dataset of 366 observations into 10 equal sized subsamples. The cross-validation process is repeated 10 times, where each subsample is retained once as the validation dataset for assessing accuracy and efficiency, and the remaining 9 subsamples are used for deriving the optimal cut-off and the curtailment tables. The cut-off and the curtailment tables are then used for simulating curtailment on the validation dataset. For each observation, a classification and a number of items administered are thus obtained. Using these values to assess accuracy and efficiency provides cross-validated estimates of performance (Hastie et al., 2009).

To assess the efficiency of both algorithms, means and standard deviations of the test lengths under both DC and SC were calculated. Classification accuracy was evaluated by calculating the concordance with the classification according to the full-length test, and concordance with actual readmission to a hospital. In addition, the Matthews' (1975) correlation coefficient (MCC) was calculated. The MCC is a measure of the accuracy of binary classifications, and is regarded as a measure of accuracy that can be used even when classes are unbalanced. The MCC can be interpreted as a correlation coefficient between the actual and predicted class of an observation. An MCC value of 1 indicates perfect predictions, and 0 indicates accuracy no better than random prediction.

## 4. Results

Data for 366 patients (84% of total sample) who had $<5$ missing values on the SSI were used. There were 158 males and 208 females and the mean ages of females and males were 33 years (SD=13.2) and 38 years (SD=13.8), respectively. Total scores on the SSI ranged from 0 to 38 and the mean score was 19 (SD=10.3). During the follow-up 94 patients (44 males and 50 females) were treated in hospital following a repeat suicide attempt. There were no significant differences between those who did and did not attempt suicide during the follow-up in terms of age (OR=1.02, 95% CI=1.00–1.03, p=.066), sex (OR=.82, 95% CI=.51–1.31, p=.409), employment status OR=1.49, 95% CI=.91–2.45, p=.113), marital status (OR=.1.81, 95% CI=.93–3.55, p=.083) or baseline suicidal intent (OR=1.04, 95% CI=.98–1.10, p=.240). However, those who did attempt suicide reported significantly higher levels of baseline depression (OR=1.04, 95% CI=1.02–1.06, $p<.0001$), hopelessness (OR=1.07, 95% CI=1.02–1.13, $p<.006$) and suicidal ideation (mean SSI 23(SD=8.4) versus mean SSI 18 (SD=10.6), $p<0.001$)) compared to those who did not. Full details of the measures are reported elsewhere (O'Connor et al., 2015). Table 1 shows the odds ratio of each SSI item on the observed outcome (re-attempt). Most items were significant predictors of the outcome. Items 4 and 10 showed the highest odds ratios.

The area under the curve was 0.63 (95% CI 0.57–0.69). When comparing the AUC of the full sample, with the AUC split on gender, number of suicide attempts or age, no significant differences were found. The values of the sensitivity, specificity and Youden index for all possible cut-off values on the SSI are presented in Table 2. According to the Youden index, a cut-off value of 6 provided the best classification accuracy. This means that

**Table 1**
Odds-ratio for every SSI item, with respect to the observed outcome (re-attempt).

| Items | Odds ratio | p-value |
|---|---|---|
| Item 1 | 1.07(1.01–1.14) | 0.03 |
| Item 2 | 1.07(1.01–1.14) | 0.02 |
| Item 3 | 1.07(1.02–1.14) | 0.01 |
| Item 4 | 1.13(1.06–1.19) | < 0.001 |
| Item 5 | 1.08(1.02–1.14) | 0.02 |
| Item 6 | 1.07(1.01–1.14) | 0.002 |
| Item 7 | 1.11(1.05–1.18) | < 0.001 |
| Item 8 | 1.04(1.04–1.17) | < 0.001 |
| Item 9 | 1.15(1.08–1.23) | < 0.001 |
| Item 10 | 1.13(1.07–1.20) | < 0.001 |
| Item 11 | 1.06(0.99–1.12) | 0.05 |
| Item 12 | 1.11(1.05–1.17) | < 0.001 |
| Item 13 | 1.07(1.02–1.13) | 0.007 |
| Item 14 | 1.06(0.99–1.12) | 0.07 |
| Item 15 | 1.12(1.06–1.19) | < 0.001 |
| Item 16 | 1.09(1.03–1.16) | 0.004 |
| Item 17 | 1.02(0.97–1.08) | 0.4 |
| Item 18 | 1.10(1.03–1.18) | 0.006 |
| Item 19 | 1.05(1.00–1.11) | 0.05 |

patients with test scores equal to or greater than 6 were identified as having an elevated risk for future suicidal behavior. Participants with an SSI score equal to or above 6 were classified as having an elevated risk of future suicidal behavior.

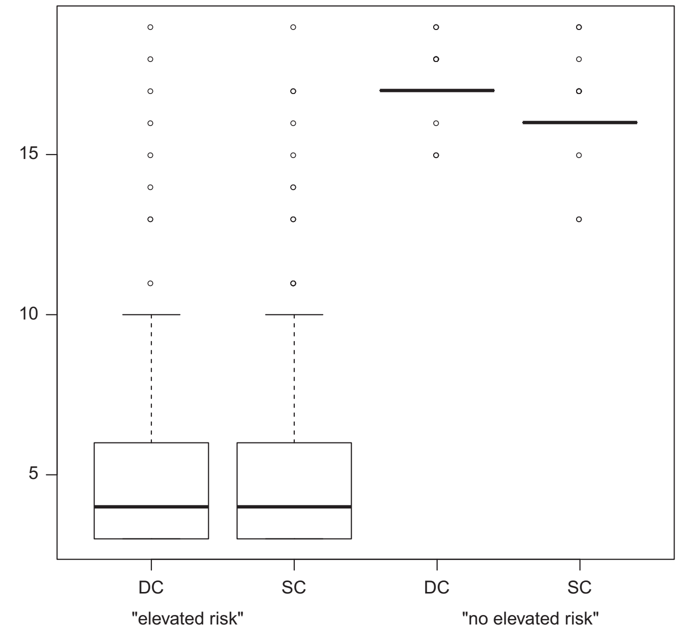


**Fig. 1.** Box plots of test-length distributions of deterministic curtailment (DC) and stochastic curtailment (SC). Split for patients at risk and patients not at risk.

**Table 2**
Sensitivity, specificity and (weighted) Youden's index for different BSSI cut-off values.

| BSSI cut-off value | Specificity (95% CI) | Sensitivity (95% CI) | Youden's index (95% CI) | Youden's index-weighted (95% CI) |
|---|---|---|---|---|
| 0 | 0.00 (0.00–0.00) | 1.00 (1.00–1.00) | 0.00 (0.00–0.00) | 0.50 (0.50–0.50) |
| 1 | 0.11 (0.08–0.15) | 0.98 (0.95–1.00) | 0.09 (0.05–0.14) | 0.54 (0.50–0.57) |
| 2 | 0.14 (0.1–0.18) | 0.97 (0.93–1.00) | 0.11 (0.05–0.16) | 0.54 (0.49–0.58) |
| 3 | 0.17 (0.13–0.22) | 0.96 (0.91–0.99) | 0.13 (0.07–0.19) | 0.55 (0.50–0.59) |
| 4 | 0.18 (0.13–0.22) | 0.96 (0.91–0.99) | 0.13 (0.07–0.19) | 0.55 (0.50–0.59) |
| 5 | 0.19 (0.14–0.24) | 0.96 (0.91–0.99) | 0.15 (0.08–0.21) | 0.55 (0.50–0.59) |
| 6 | 0.20 (0.15–0.25) | 0.96 (0.91–0.99) | 0.16 (0.09–0.22) | 0.56 (0.51–0.60) |
| 7 | 0.20 (0.15–0.25) | 0.95 (0.89–0.99) | 0.15 (0.08–0.21) | 0.55 (0.49–0.60) |
| 8 | 0.21 (0.16–0.26) | 0.94 (0.88–0.98) | 0.15 (0.07–0.21) | 0.54 (0.48–0.59) |
| 9 | 0.22 (0.17–0.27) | 0.93 (0.87–0.98) | 0.15 (0.07–0.22) | 0.54 (0.47–0.59) |
| 10 | 0.25 (0.20–0.30) | 0.93 (0.87–0.98) | 0.17 (0.10–0.24) | 0.55 (0.49–0.60) |
| 11 | 0.25 (0.20–0.30) | 0.91 (0.85–0.97) | 0.16 (0.09–0.24) | 0.54 (0.47–0.60) |
| 12 | 0.26 (0.21–0.31) | 0.89 (0.83–0.95) | 0.15 (0.07–0.23) | 0.52 (0.45–0.59) |
| 13 | 0.27 (0.22–0.33) | 0.89 (0.83–0.95) | 0.17 (0.08–0.25) | 0.53 (0.46–0.60) |
| 14 | 0.28 (0.23–0.33) | 0.88 (0.81–0.95) | 0.16 (0.08–0.25) | 0.52 (0.45–0.59) |
| 15 | 0.31 (0.26–0.37) | 0.87 (0.80–0.94) | 0.19 (0.10–0.27) | 0.53 (0.45–0.60) |
| 16 | 0.34 (0.29–0.40) | 0.83 (0.74–0.90) | 0.17 (0.07–0.26) | 0.50 (0.42–0.58) |
| 17 | 0.37 (0.31–0.43) | 0.80 (0.71–0.87) | 0.17 (0.06–0.26) | 0.48 (0.39–0.57) |
| 18 | 0.39 (0.33–0.44) | 0.79 (0.70–0.87) | 0.17 (0.07–0.27) | 0.48 (0.39–0.57) |
| 19 | 0.42 (0.36–0.48) | 0.73 (0.64–0.82) | 0.16 (0.05–0.26) | 0.45 (0.35–0.54) |
| 20 | 0.46 (0.40–0.52) | 0.70 (0.61–0.79) | 0.16 (0.05–0.27) | 0.43 (0.33–0.53) |
| 21 | 0.50 (0.44–0.56) | 0.67 (0.57–0.77) | 0.17 (0.06–0.28) | 0.42 (0.32–0.52) |
| 22 | 0.55 (0.49–0.61) | 0.64 (0.54–0.73) | 0.19 (0.08–0.30) | 0.42 (0.31–0.51) |
| 23 | 0.58 (0.53–0.64) | 0.57 (0.48–0.67) | 0.16 (0.04–0.27) | 0.37 (0.26–0.47) |
| 24 | 0.65 (0.59–0.70) | 0.51 (0.41–0.61) | 0.16 (0.04–0.27) | 0.33 (0.23–0.44) |
| 25 | 0.69 (0.63–0.74) | 0.48 (0.38–0.57) | 0.17 (0.05–0.28) | 0.32 (0.22–0.43) |
| 26 | 0.74 (0.68–0.79) | 0.44 (0.34–0.53) | 0.17 (0.06–0.28) | 0.30 (0.20–0.41) |
| 27 | 0.77 (0.72–0.82) | 0.36 (0.27–0.46) | 0.13 (0.03–0.24) | 0.25 (0.15–0.35) |
| 28 | 0.81 (0.76–0.86) | 0.3 (0.21–0.39) | 0.11 (0.01–0.21) | 0.20 (0.11–0.30) |
| 29 | 0.85 (0.80–0.89) | 0.24 (0.16–0.33) | 0.09 (−0.01–0.19) | 0.17 (0.08–0.26) |
| 30 | 0.88 (0.84–0.92) | 0.21 (0.13–0.30) | 0.09 (0.00–0.18) | 0.15 (0.07–0.24) |
| 31 | 0.91 (0.87–0.94) | 0.18 (0.11–0.26) | 0.09 (0.01–0.17) | 0.13 (0.06–0.22) |
| 32 | 0.93 (0.90–0.96) | 0.16 (0.09–0.23) | 0.09 (0.01–0.17) | 0.12 (0.05–0.20) |
| 33 | 0.94 (0.91–0.96) | 0.12 (0.05–0.18) | 0.05 (−0.01–0.13) | 0.08 (0.02–0.16) |
| 34 | 0.96 (0.93–0.98) | 0.06 (0.02–0.12) | 0.02 (−0.03–0.08) | 0.04 (0.00–0.10) |
| 34 | 0.97 (0.94–0.99) | 0.03 (0.00–0.07) | 0.00 (−0.04–0.04) | 0.01 (−0.02–0.06) |
| 36 | 0.99 (0.98–1.00) | 0.03 (0.00–0.07) | 0.02 (−0.01–0.07) | 0.03 (0.00–0.07) |
| 37 | 1.00 (1.00–1.00) | 0.02 (0.00–0.05) | 0.02 (0.00–0.05) | 0.02 (0.00–0.05) |
| 38 | 1.00 (1.00–1.00) | 0.00 (0.00–0.00) | 000 (0.00–0.00) | 0.00 (0.00–0.00) |

*Note*: Point estimates are median values from 10,000 bootstrapped samples' distributions; 95%.CIs are the 100*(α/2)th and 100*(1−α/2)th percentiles of the bootstrapped samples' distributions. Youden's index was calculated as (sensitivity+specificity−1); the weighted Youden's index was calculated as (sensitivity+.5*specificity−.5).

Both the binary cutscore and the full-length score significantly predicted suicidal behavior (odds ratio of 1.2; 95% CI: 1.1–1.4 versus 1.0; 95% CI: 1.0–1.1. When controlling for gender, baseline depression and hopelessness, the odd ratios remained the same. No significant differences between the predictive validity of binary versus the full length scale were found.

The value of MCC for classifying patients according to the cut-off value of 6, estimated using 10-fold CV, was 0.18, indicating a small to moderate relationship between the classification according to the full-length SSI, and future suicidal behavior.

When DC was applied to the dataset, average test length, assessed by means of 10-fold CV, was 7.30 items (SD=5.11; Fig. 1). More specifically, for the observations with an elevated risk (N=308), average test length was 5.45 (SD=3.03) items. For the observations with no elevated risk (N=58), average test length was 17.14 (SD=0.87).

SC was applied to the dataset, with gamma-values set to .90. The average test length for SC, assessed by means of 10-fold CV, was 7.15 (SD=4.82; Fig. 1). More specifically, among the observations with an elevated risk (N=308), the average number of items administered was 5.45 (SD=3.02). Among the observations with no elevated risk (N=58), the average number of items administered was 16.16 (SD=1.23).

All final 'elevated risk' and 'no elevated risk' decisions based on DC and SC were the same as when the full-length test was administered. Therefore, accuracy for DC and SC, assessed by means of 10-fold CV was the same as for the full-length test: 144 observations were correctly classified, and 222 were incorrectly classified. Of the 222 observations that were incorrectly classified, 217 (80%) were false positives, and only 5 (5%) were false negatives (Table 3). This yields a proportion of correct classifications of 0.57 and a proportion of incorrect classifications of 0.43 (Table 3).

Tables 4 and 5 represent look-up tables for deterministically and stochastically curtailed administration of the SSI. For both DC and SC, as soon as a patient has a sum score of 6, item administration can be stopped and the patient can be classified as having an 'elevated risk'. Also for both analyses, the earliest theoretical possibility for this is after administration of item 3. The items 1, 2 and 3 must then have been answered with response option 2. If the participant had, for example, a cumulative score of 5 after answering item 3, item administration is continued until the sum score of 6 is reached. Otherwise all 19 items are administered.

Also, with DC, item administration can be halted and a 'no elevated risk' classification decision can be made, after administration of item 17. This can only be done if the cumulative score of a patient is ≤1 at item 17. If, for example, the cumulative score of a patient is 2 at item 17, a total item score of 6 is still possible if the patient endorses items 18 and 19 with a 2.

Table 3 shows that with the application of SC, further gains in efficiency can be made. With SC, early stopping of item administration and making a no elevated risk classification decision is possible after administration of item 16. As stated above, early stopping of item administration and making an elevated-risk classification decision is possible after administration of item 3 with SC, which is the same as with DC.

**Table 3**
Classification accuracy for DC and SC, based on 10-fold CV.

| Test score ≥ 6 | Suicide attempt at follow-up | | Total |
|---|---|---|---|
| | No | Yes | |
| No | 55 (20%) | 5 (5%) | 60 (16%) |
| Yes | 217 (80%) | 89 (95%) | 306 (84%) |
| Total | 272 (74%) | 94 (26%) | 366 (100%) |

*Note*: The total proportion of correct classifications equals (55/272+89/94)/2=0.57; the total proportion of incorrect classifications equals (217/272+5/94)/2=0.43.

## 5. Discussion

In this study we demonstrated that curtailment could reduce the length of the Beck Scale for Suicide Ideation without reducing the accuracy of the final classification decision. Firstly, a cut-off value of SSI ≥ 6 was found to provide the best classification accuracy for future suicidal behavior. Using this cut-off value, stochastic curtailment resulted in the greatest reduction of items administered. On average, less than 8 items were needed to assess whether a patient would be classified as at risk or not at risk, according to the full-length test. In other words, stochastic curtailment resulted in an average reduction in test length of 62%. It should be noted that the average test length among patients with an elevated risk of suicidal behavior was substantially smaller (about 11 items), compared to the average test length among patients not having an elevated risk.

The main advantage of curtailment is its ease of application. Compared to computer adaptive testing (CAT) solutions, the implementation of curtailment is possible without much effort. No specialized computer software is needed. The look-up tables can be distributed easily in mental health settings. This reduces the burden for both interviewers and patients, and may therefore result in a more frequent assessment of risk for suicidal behavior. Also, curtailment does not involve assumptions about factor structure and the distribution of the data, as CAT does (van der Linden and Hambleton, 1997).

Several limitations must be noted. First of all, the scores on the SSI allowed for better prediction of future suicidal behavior than chance (i.e. 0.5), however the area under the curve was only 0.63 (95% CI 0.56–0.69). This is below the suggested threshold for a medium effect size (.64) (Rice and Harris, 2005). Ideally, a screener would have an ROC of ≥0.8 (Chaudron et al., 2010). In this study, based on the cross-validated accuracy findings, the best cut-off failed to identify 5 out of 94 (5%) participants who were readmitted to a hospital following a repeat suicide attempt. Although this is a reasonable performance, it is not an optimal performance for such a potentially lethal condition. Also, of the patients who were not re-admitted, 80% were wrongly classified as having an elevated risk. Although the Youden index indicated that the cut-off of ≥6 was the best cut-off, the differences between lower or higher cut-off points was rather small. More concretely, any other cut-off between 1 and 15 would only show a marginally different classification accuracy when compared to the chosen cut-off of ≥6. As stated in the introduction, the only published cut-off score for the SSI comes from a 20 year prospective study among psychiatric outpatients. We argued that this cut-off would be less

**Table 4**
Cut-off values for every SSI item, based on deterministic curtailment.

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No elevated risk | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | 1 | 3 | 5 |
| Elevated risk | – | – | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |

**Table 5**
Cut-off values for every SSI item, based on stochastic curtailment.

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No elevated risk | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | 1 | 2 | 3 | 5 |
| Elevated risk | – | – | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |

appropriate for our sample, as suicide ideation among patients treated in a hospital for a suicide attempt is expected to be higher than among psychiatric outpatients. Indeed, in our sample, only 51 patients (14%) of the patients scored $< 3$. Still, the Youden index for SSI $\geq 3$ was only marginally different from the Youden index of SSI $\geq 6$. In sum, it remains difficult to establish a single cut-off point for such complex behaviors such as suicide attempts, which are inherently difficult to predict. Therefore, one should not only rely on a single instrument or a single cut-off value for risk classification. Although the lack of predictive power is a noteworthy limitation in the clinical utility of the SSI, or any other suicide assessment scale, it does not detract from the overall objective of this study, which was to demonstrate the potential utility of curtailment in reducing the length of widely used screening/risk assessment tools.

Future studies should aim to improve the predictive validity of the scale by altering or rephrasing items of the SSI, as suggested by De Beurs et al. (2014). More items were needed to classify patients not having an elevated risk. This was also found in the previous study (De Beurs et al., 2014). The SSI seems to be less effective in identifying patients with lower suicidal ideation. Adding items that are more easily endorsed by patients with a lower trait might improve the predictive validity of the SSI.

It is important to note that the ROC analysis which was used to derive the optimal cut-off was calculated using the same sample that was employed to test the curtailment methods. Ideally, the validation of a new method is tested on a different sample from the one upon which the threshold score has been derived. In this study, items of the scale are administered in the original order. As we found that some items predicted future suicidal behavior better than others, future potential reductions in SSI screening length might be reached by changing item administration order in curtailment.

Future studies should re-validate the optimal cut-off for risk of repeat suicidal behavior following hospital treatment for an index suicide attempt. The curtailment outcomes should also be replicated within a different dataset with hospital-treated patients. Importantly, less serious suicidal behavior that did not reach clinical services or patients who, although they presented to emergency departments, were not hospitalized are not captured in the current dataset. Also, although the numbers will have been very low, any suicidal behavior that occurred outside of Scotland was not included (O'Connor et al., 2015).

In sum, this is the first study to apply curtailment to optimize the assessment of suicidal behavior. Curtailment can be used to simplify the assessment of suicidal behavior, and should be considered as an alternative for the full-scale assessment. Given that a psychiatric assessment with appropriate follow-up after a suicide attempt has been found to reduce future suicidal behavior (Hickey et al., 2001), the application of curtailment techniques might help care providers to more frequently assess patients treated in a hospital for a suicide attempt. Although curtailment can reduce the item count considerably, for routine clinical practice, we acknowledge that 8 items still remains a lot, however more sophisticated techniques such as CAT are likely to further reduce the numbers of items needed. The limitation of methods such as CAT is that they require specific software that is not available in most health care practices. On balance, therefore, we believe that curtailment will remain helpful for clinicians while other more accessible techniques are developed and implemented.

## Authors' contributions

ROC conducted the initial study among hospital treated suicide attempters. DdB suggested the idea and wrote the initial draft. MF performed all simulations and analyses. All authors contributed to the writing of the manuscript and approved the final version.

### Conflict of interest

None.

### Acknowledgements

The initial study of ROC was supported by funding from the Chief Scientist Office, Scottish Government (CZH/4/449). For the additional curtailment analysis, no additional funding was obtained.

## References

Beck, A.T., Kovacs, M., Weissman, A., 1979. Assessment of suicidal intention: the scale for suicide ideation. J. Consult. Clin. Psychol. 47, 343–352.
Brown, G.K., Beck, A.T., Steer, R.A., Grisham, J.R., 2000. Risk factors for suicide in psychiatric outpatients: a 20-year prospective study. J. Consult. Clin. Psychol. 68 (3), 371–377.
Carpenter, J., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. 19 (9), 1141–1164.
Chaudron, L.H., Szilagyi, P.G., Tang, W., Anson, E., Talbot, N.L., Wadkins, H.I., Tu, X., Wisner, K.L., 2010. Accuracy of depression screening tools for identifying postpartum depression among urban mothers. Pediatrics 125, e609–e617.
De Beurs, D.P., De Groot, M.H., De Keijser, J., Van Duijn, E., De Winter, R., Kerkhof, A. J.F.M., 2015a. Do patients benefit from the training of mental health professionals in suicide practice guidelines? A cluster randomized trial. Br. J. Psychiatry . http://dx.doi.org/10.1192/bjp.bp.114.156208 (Published online first: 17 December 2015).
De Beurs, D.P., de Vries, A.L., de Groot, M.H., de Keijser, J., Kerkhof, A.J., 2014. Applying computer adaptive testing to optimize online assessment of suicidal behavior: a simulation study. J. Med. Internet Res. 16, e207.
de Beurs, D.P., Fokkema, M., de Groot, M.H., de Keijser, J., Kerkhof, A.J., 2015b. Longitudinal measurement invariance of the beck scale for suicide ideation. Psychiatry Res. 225, 368–373.
Finkelman, M.D., He, Y., Kim, W., Lai, A.M., 2011. Stochastic curtailment of health questionnaires: a method to reduce respondent burden. Stat. Med. 30, 1989–2004.
Finkelman, M.D., Smits, N., Kim, W., Riley, B., 2012. Curtailment and stochastic curtailment to shorten the CES-D. Appl. Psychol. Meas., 632–658.
Fokkema, M., Smits, N., Finkelman, M.D., Kelderman, H., Cuijpers, P., 2014. Curtailment: a method to reduce the length of self-report questionnaires while maintaining diagnostic accuracy. Psychiatry Res. 215, 477–482.
Friedman, T., Newton, C., Coggan, C., Hooley, S., Patel, R., Pickard, M., Mitchell, A.J., 2006. Predictors of A&E staff attitudes to self-harm patients who use self-laceration: influence of previous training and experience. J. Psychosom. Res. 60, 273–277.
Hawton, K., van Heeringen, K., 2009. Suicide. Lancet 373, 1372–1381.
Hastie, Trevor, Robert, Tibshirani, Jerome, Friedman, 2009. Unsupervised learning. Springer, New York.
Hickey, L., Hawton, K., Fagg, J., Weitzel, H., 2001. Deliberate self-harm patients who leave the accident and emergency department without a psychiatric assessment: a neglected population at risk of suicide. J. Psychosom. Res. 50, 87–93.
Jacobs, D., Brewer, M., 2004. APA practice guideline. Psychiatr. Ann. 34, 373.
Kapur, N., Cooper, J., Hiroeh, U., May, C., Appleby, L., House, A., 2004. Emergency department management and outcome for self-poisoning: a cohort study. Gen. Hosp. Psychiatry 26, 36–41.
Kerkhof, A., Mulder, S., Draisma, J., 2007. Suïcidepogingen behandeld in Nederlandse ziekenhuizen. Tijdschr. Geneeskd. 85, 37–42.
Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.
O'Connor, R.C., Smyth, R., Williams, J.M.G., 2015. Intrapersonal positive future

thinking predicts repeat suicide attempts in Hospital-treated suicide attempters. J. Consult. Clin. Psychol. 83, 169–176.
Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., 2007. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information system (PROMIS). Med. Care 45, S22–S31.
Rice, M.E., Harris, G.T., 2005. Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law Hum. Behav. 29, 615.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B 36, 111–147.
van der Linden, W.J., Hambleton, R.K., 1997. Handbook of Modern Item Response Theory, first ed. Springer, New York.
Smits, N., Cuijpers, P., van Straten, A., 2011. Applying computerized adaptive testing to the CES-D scale: a simulation study. Psychiatry Res. 188, 147–155.
Spijker, B.A., Batterham, P.J., Calear, A.L., Farrer, L., Christensen, H., Reynolds, J., Kerkhof, A.J., 2014. The suicidal ideation attributes scale (SIDAS): community-based validation study of a new scale for the measurement of suicidal ideation. Suicide Life-Threat. Behav. 44, 408–419.
van Hemert, A.M., Kerkhof, A.F.J.M., de Keijser, J., verwey, B., 2012. Multidisciplinaire richtlijn voor diagnostiek en behandeling van suïcidaal gedrag. De Tijdstroom, Utrecht.
Verwey, B., Waarde, J.V., Rooij, I.V., Gerritsen, G., Zitman, F.G., 2007. Guidelines in Dutch mental health institutions for dealing with persons attempting suicide. Tijdschr. Psychiatr. 49, 157–165.
Wasserman, D., Rihmer, Z., Rujescu, D., Sarchiapone, M., Sokolowski, M., Titelman, D., Zalsman, G., Zemishlany, Z., Carli, V., 2012. The European Psychiatric Association (EPA) guidance on suicide treatment and prevention. Eur. Psychiatry 27, 129–141.
World Health Organisation, 2014. Preventing Suicide, a Global Perspective. Geneva.
Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35.