Quinlivan, L., Cooper, J., Davies, L., Hawton, K., Gunnell, D., & Kapur, N. (2016). Which are the most useful scales for predicting repeat self-harm? A systematic review evaluating risk scales using measures of diagnostic accuracy. BMJ Open, 6(2), [e009297]. DOI: 10.1136/bmjopen-2015-009292
BMJ Open  Which are the most useful scales for predicting repeat self-harm? 
A systematic review evaluating risk scales using measures of diagnostic accuracy

L Quinlivan,1 J Cooper,1 L Davies,2 K Hawton,3 D Gunnell,4 N Kapur1,5

ABSTRACT
Objectives: The aims of this review were to calculate the diagnostic accuracy statistics of risk scales following self-harm and consider which might be the most useful scales in clinical practice.

Design: Systematic review.

Methods: We based our search terms on those used in the systematic reviews carried out for the National Institute for Health and Care Excellence self-harm guidelines (2012) and evidence update (2013), and updated the searches through to February 2015 (CINAHL, EMBASE, MEDLINE, and PsychINFO). Methodological quality was assessed and three reviewers extracted data independently. We limited our analysis to cohort studies in adults using the outcome of repeat self-harm or attempted suicide. We calculated diagnostic accuracy statistics including measures of global accuracy. Statistical pooling was not possible due to heterogeneity.

Results: The eight papers included in the final analysis varied widely according to methodological quality and the content of scales employed. Overall, sensitivity of scales ranged from 6% (95% CI 5% to 6%) to 97% (CI 95% 94% to 98%). The positive predictive value (PPV) ranged from 5% (95% CI 3% to 9%) to 84% (95% CI 80% to 87%). The diagnostic OR ranged from 1.01 (95% CI 0.434 to 2.5) to 16.3 (95%CI 12.5 to 21.4). Scales with high sensitivity tended to have low PPVs.

Conclusions: It is difficult to be certain which, if any, are the most useful scales for self-harm risk assessment. No scales perform sufficiently well so as to be recommended for routine clinical use. Further robust prospective studies are warranted to evaluate risk scales following an episode of self-harm. Diagnostic accuracy statistics should be considered in relation to the specific service needs, and scales should only be used as an adjunct to assessment.

INTRODUCTION
Self-harm is a frequent clinical challenge and a strong predictor of future suicide.1,2 One in six individuals presenting to hospital with self-harm will repeat the behaviour within 1 year.3–4 Psychosocial assessment on presentation to hospital is a key component of recommended clinical management.5,6 Guidelines recommend that all patients presenting to the hospital services with self-harm should receive a preliminary psychosocial assessment to determine mental capacity and evaluate willingness to stay for further treatment.5 Mental health professionals should conduct a more comprehensive evaluation of risk and needs at a later stage, and risk scales are typically a core component of assessments despite limited evidence of their effectiveness.6,7 Some clinical guidelines advise against the use of scales to determine management, but suggest they can be used to help structure assessments.6 Other guidelines recommend that only scales that have undergone formal testing should be used as part of clinical assessments.8

Our recent study in 32 English hospitals found that at least 20 risk tools were in use, suggesting that there is a lack of consensus...
over which scales are best for evaluating risk of further self-harm. The uncertainty is perhaps due to methodological differences between studies and variable standards of reporting. There are a small number of reviews which consider the predictive ability of risk scales for repeat self-harm which may help clinicians to select the most helpful tools but the information provided is mostly limited to dual indicators such as sensitivity/specificity, positive/negative predictive values, and there is little practical guidance for clinicians in selecting the 'most useful tools'.

While these dual indicators are useful for determining the predictive validity of a scale, a broader range of diagnostic test criteria may be helpful when selecting an appropriate scale for clinical use given the inevitable trade-off between sensitivity (the proportion of individuals who repeat self-harm identified by the test as high risk) and specificity (proportion of people who did not repeat self-harm identified as low risk by the test). For example, a highly sensitive test might identify all patients at risk of future self-harm but could be over inclusive with cost and resource implications. Conversely, the higher threshold inherent in highly specific tests may result in false negatives and a host of deleterious consequences for patients and clinical services.

We have conducted a systematic review of existing research on risk scales to consider these issues. The objectives were to:
1. Investigate the performance of risk scales following self-harm or attempted suicide on a wide range of dual measures, as well as more global measures of accuracy.
2. Consider which might be the most useful scales following self-harm in clinical practice settings.

This information may be useful to clinicians, commissioners and hospital risk managers, who need to critically evaluate scales for use in clinical practice.

**METHOD**

This study extends the reviews carried out as part of the National Institute for Health and Care Excellence (NICE) self-harm guidelines and evidence update on the use of risk scales for repeat self-harm. We included recent evidence and considered a much broader range of diagnostic accuracy statistics than the original reviews.

**Literature search**

We identified studies evaluating the predictive validity of risk scales for repeat self-harm from the NICE review on the longer term management of self-harm and the evidence update. We used the same published search strategy (see online supplementary appendix 1) on CINAHL, EMBASE, MEDLINE and PsychINFO databases through to February 2015. Reference lists were also screened and related references reviewed.

**Inclusion and exclusion criteria**

Consistent with the NICE self-harm evidence update, studies were included if they used a cohort design—the optimal design for evaluating the diagnostic accuracy of scales as case-control studies can overestimate diagnostic accuracy. Although suicide is an extremely important outcome following self-harm, the low base-rate hinders predictive efforts even in high-risk populations. We focused on repeat self-harm or attempted suicide as an outcome, as the incidence rate is higher and the prediction of repetition may more feasible than predicting suicide. Studies were included if measures of diagnostic accuracy (such as sensitivity, specificity and positive predictive values) were reported.

Studies were excluded if the scales were validated on a specific or restricted samples (eg, veterans, prisoners or specialist mental healthcare population), or a sample which did not include people presenting with self-harm or attempted suicide. One study recruited a mixed sample of people (presenting with suicide ideation or self-harm), but since a majority of the sample (75%) had a history of self-harm and the study outcome was self-harm repetition, this study was included.

Some tools were validated in more than one setting and these were included once in the final analysis, using the original paper, if this met the inclusion criteria. We did this in order to gain an indication of the 'best-case' scenario for different instruments (the first study of a new screening tool in a setting where it was developed might be expected to give the most positive results) and because of the potential difficulty of combining measures of diagnostic accuracy from different settings. However, in order to contextualise results we did also examine the broader performance of scales which had been tested in multiple studies in a post hoc analysis.

**Assessment of bias and study quality**

Study bias was evaluated at the study level using the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) and STARD (Standards for Reporting of Diagnostic Accuracy) guidelines.

**Statistical analysis**

True positives, false positives, true negatives and false negatives were extracted from the papers by two researchers (LQ and JC) independently, and results discussed with the third author (NK). Authors were contacted where these data were unavailable.

We used a wide range of recommended diagnostic accuracy estimates to evaluate the predictive validity of the risk scales (box 1 and see online supplementary appendix 2), including sensitivity (proportion of individuals who repeat self-harm identified as high risk by the test); specificity (proportion of people who did not repeat self-harm identified as low risk by the test); positive predictive values (probability that a person identified by the test as high risk will actually go onto
self-harm); negative predictive values (probability that a person identified as low risk will not go onto self-harm).

Positive and negative likelihood ratios (how much more or less likely test results are to occur in patients who repeat self-harm vs those who do not) were also calculated. Likelihood ratios of 1 indicate no change in likelihood of disease or outcome (in this case repeat self-harm). Positive likelihood ratios between 1–2, 2–5, 5–10 and >10 indicate minimal, small, moderate and large increases in risk, respectively. Negative likelihood ratios of 0.5–1.0, 0.02–0.5, 0.1–0.2 and <0.1 indicate minimal, small, moderate and large decreases in risk.

We also calculated global diagnostic statistics that summarise the diagnostic performance of a test as a single indicator, including the ‘number allowed to diagnose’ (number of individuals who are correctly assigned as at high risk of repetition before one is misassigned), and the diagnostic OR (odds of positivity in repeater relative to the odds of non-repeater). Higher values indicate greater test discriminatory power.

CIs for sensitivity and specificity were calculated using the Wilson score method without correction. CIs for positive and negative likelihood ratios were produced using the method of Simel et al. The CI for the diagnostic OR was produced using the method published by Armitage and Berry. CIs for ‘number allowed to diagnose’ were constructed using the method based on constant \( \chi^2 \) boundaries from Press et al. Results are unpooled due to heterogeneity in the studies.

STATA V.13.0; StataCorp. Stata Statistical Software: Release 13, College Station, Texas: StataCorp LP, 2013) and RevMan V.5.1 (Cochrane Collaboration) were used for statistical analysis.

RESULTS

Search results

The NICE 2011 review on the longer term management of self-harm included seven cohort studies testing the predictive validity of risk scales for repeat self-harm. Four were excluded as they did not meet our inclusion criteria—they examined global measures rather than scales or used a restricted clinical population (figure 1). The NICE

Figure 1 Preferred Reporting for Systematic Reviews and Meta-Analyses flow diagram describing the search process for included studies. NICE, National Institute for Health and Care Excellence.
evidence update included one additional cohort study. The search strategy from January 2012 to February 2015 resulted in an additional 60 papers of which three were relevant prospective cohort studies, and one additional cohort study was retrieved from related references (see figure 1). We also reran the searches for the earlier time periods. No additional studies were identified. In total, there were eight studies examining 11 scales which were included in the final analysis (figure 1).

Description of studies

The methodological characteristics of the eight studies evaluating 11 scales are described in table 1. Further detailed information on bias and reporting is presented in online supplementary appendix 3. The studies were conducted between 2002 and 2014, and included hospital presentations of self-harm or attempted suicide. In terms of service context, the studies were generally carried out across multiple sites, the majority in publicly funded health services. Four studies were based on self-harm emergency department populations, and one study included a mix of patients presenting with self-harm or suicidal ideation. One study was based on patients treated for self-poisoning. Two studies were based on hospital presentations for suicide attempts with suicidal intent as an inclusion criterion, and one study was based on patients admitted to a medical bed after self-harm. The length of follow-up ranged between 3 and 36 months, and outcome data was mostly ascertained through hospital databases. The incidence of repeat self-harm across studies ranged from 3% to 37%, possibly suggesting differences in casemix.

Four studies involved developing a tool which was then validated on a split site or external data set. The remainder were validation studies of existing scales. The scales varied in length ranging from four items (Manchester Self-Harm Rule, ReACT Self-Harm Rule) to 53 items for the Global Severity Scale. Most scales included previous history of self-harm or suicide attempts or prior psychiatric treatment as items. Others scales included personality factors (Barratt Impulsivity Scale, clinical symptomology (eg, Global Severity Index), drug misuse (eg, Drug Abuse Screening Test) and variations in symptoms associated with suicidal thoughts and behaviours (eg, Suicide Assessment Scale).

None of the studies were explicitly formatted according to standard guidelines (eg, STARDi) and reporting varied across the studies. For example, there were variations across studies in the reporting of recruitment flow and patient characteristics, cross-tabulations of raw data, and use of thresholds (eg, Randall et al did not use any). The database studies were the most robustly reported according to STARDi.

Diagnostic accuracy statistics

The full range of diagnostic accuracy statistics are presented in table 2. Figures 2 and 3 show forest plots for sensitivity and positive predictive values, respectively. Sensitivity (how well the test identifies people who repeat self-harm) ranged from 5.6% for the Repeated Episodes of Self-Harm scale using the threshold for the highest risk to 97% for the Manchester self-harm rule. Positive predictive values for the latter high sensitivity scales were low (26%, 21% and 11%, respectively) and were highest for the Repeated Episodes of Self-Harm scale at the highest threshold (84%) followed by the Global Severity Index (73%), and the Drug Abuse Screening Test (figure 3). It should be noted that the Repeated Episodes of Self-Harm score was tested on inpatients admitted to hospital services for self-harm.

Positive predictive values for the latter high sensitivity scales were low (26%, 21% and 11%, respectively) and were highest for the Repeated Episodes of Self-Harm scale at the highest threshold (indicating a large increase in the likelihood of repetition) to 1.0 for Söderjukuset Self-harm Rule and the Suicide Assessment Scale (indicating no change in the likelihood of repetition) (table 2). The diagnostic OR which presents the accuracy of a test as a global single indicator ranged from 16.34 (Repeated Episodes of Self-Harm scale at the highest threshold) and 10.77 (Manchester Self-Harm Rule) to 1.01 for the Södersjukuset Self-harm Rule and the Suicide Assessment Scale (table 2).

Although the length of follow-up varied, there were no clear patterns in relation to the prediction of shorter versus longer term risk. As noted previously, there was a wide variation in the methodological characteristics of the studies and in the scales themselves.

Operational issues

Operational characteristics (ie, the time taken to do the scale, technical specifications, ease of use, cost, staff training, user acceptability) are important to the clinical use of a scale and are listed in detail in table 3. Scales with characteristics which may need to be considered before their use include the Global Severity Index (copyright protected, costs associated with use, a 53-item scale with training required prior to use). The Drug Abuse Screening test may also be limited for clinicians working with self-harm populations, as the test is designed to assess drug-related problems.

DISCUSSION

Main findings

Risk scales are in widespread use in health services managing self-harm patients. We examined the diagnostic accuracy of a number of scales after self-harm and found a wide variation in samples, follow-up, reporting, thresholds, as well as differences in the content of the scales.
<table>
<thead>
<tr>
<th>Study ID</th>
<th>Index test and comparator tests</th>
<th>Participants</th>
<th>Outcome events</th>
<th>Sampling and clinical population</th>
<th>How assessment was conducted</th>
<th>Context</th>
<th>Outcomes</th>
<th>Reference standard</th>
<th>Follow-up (months)</th>
<th>Estimates of 95% CI for diagnostic accuracy</th>
<th>Included raw data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooper et al&lt;sup&gt;27&lt;/sup&gt; (England)</td>
<td>MSHR (development and validation)</td>
<td>2095</td>
<td>373</td>
<td>Consecutive emergency department self-harm presentations</td>
<td>Variables gathered as part of routine assessment and extracted from a database</td>
<td>Multisite, within publicly funded National Health Service</td>
<td>Self-harm and suicide</td>
<td>Hospital database records, searched by definition</td>
<td>6</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Steeg et al&lt;sup&gt;26&lt;/sup&gt; (England)</td>
<td>ReACT rule (development and validation)</td>
<td>7039</td>
<td>2096</td>
<td>Consecutive assessed and non-assessed emergency department self-harm presentations (England)</td>
<td>Variables gathered as part of routine assessment and extracted from a database</td>
<td>Multisite, within publicly funded National Health Service</td>
<td>Repeat self-harm and suicide</td>
<td>Hospital database records, searched by definition</td>
<td>6</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Bilèn et al&lt;sup&gt;25&lt;/sup&gt; (Sweden)</td>
<td>SSHR, MSHR</td>
<td>325</td>
<td>80</td>
<td>Consecutive emergency department self-harm presentations</td>
<td>Scales completed by treating physician</td>
<td>Two large university hospitals with emergency departments, within publicly funded National Health Service</td>
<td>Repeat self-harm</td>
<td>Hospital database records, searched by definition</td>
<td>6</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Spittal et al&lt;sup&gt;14&lt;/sup&gt; (Australia)</td>
<td>RESH (development and validation)</td>
<td>84 659</td>
<td>21 672</td>
<td>Consecutive inpatients admitted for self-harm</td>
<td>Large linked data gathered as part of hospital admissions</td>
<td>Multisite, private and publicly funded hospitals</td>
<td>Repeat self-harm and suicide combined</td>
<td>Hospital database records, searched</td>
<td>6</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Study ID</td>
<td>Index test and comparator tests</td>
<td>Participants</td>
<td>Outcome events</td>
<td>Sampling and clinical population</td>
<td>How assessment was conducted</td>
<td>Context</td>
<td>Outcomes</td>
<td>Reference standard</td>
<td>Follow-up (months)</td>
<td>Estimates of 95% CI for diagnostic accuracy</td>
<td>Included raw data</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>---------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>----------------------------------------------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>--------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Carter et al^29 (Australia)</td>
<td>ERRS</td>
<td>1317</td>
<td>188</td>
<td>Consecutive self-poisoning patients presenting for hospital treatment at centralised referral centre</td>
<td>Data gathered by toxicology and psychiatric staff and rated by psychiatric staff (psychiatrist, psychiatric registrars, clinical nurse consultants) rated ERRS variables based on clinical interviews, patient self-report, and case notes</td>
<td>Tertiary specialist service for self-poisoning</td>
<td>Repeat self-poisoning</td>
<td>Hospital database records, searched by definition</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Randall et al^45 (Canada)</td>
<td>GSI, BIS DAST BHI, CAGE, MSHR</td>
<td>157</td>
<td>34</td>
<td>Emergency department presentations with self-harm or suicidal ideation</td>
<td>Trained researcher administered standardised interview and conducted chart reviews</td>
<td>Two teaching hospitals with largest emergency departments in Edmonton</td>
<td>Repeat self-harm</td>
<td>Hospital records and telephone call</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Waern^33 (Sweden)</td>
<td>SUAS</td>
<td>162</td>
<td>56</td>
<td>Unclear sampling, patients admitted to ED</td>
<td>Face-to-face interviews carried out by three</td>
<td>Publically funded University hospital,</td>
<td>Repeat suicide attempts and suicide</td>
<td>Hospital database records, search</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th>Study ID</th>
<th>Index test and comparator tests</th>
<th>Participants</th>
<th>Outcome events</th>
<th>Sampling and clinical population</th>
<th>How assessment was conducted</th>
<th>Context</th>
<th>Outcomes</th>
<th>Reference standard</th>
<th>Follow-up (months)</th>
<th>Estimates of 95% CI for diagnostic accuracy</th>
<th>Included raw data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolton et al (Canada)</td>
<td>SPS, MSPS</td>
<td>2846</td>
<td>80</td>
<td>wards after a suicide attempt with at least some intent to die</td>
<td>Consecutive adult referrals to psychiatric services from the emergency department</td>
<td>which is the only hospital to provide emergency services in the study area</td>
<td>Future suicide attempts</td>
<td>Unclear</td>
<td>6</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

BHI, Beck Hopelessness Scale; BIS, Barratt Impulsivity Scale; C-CASA, Columbia Classification Algorithm of Suicide Assessment; CAGE, Cut down, Annoyed, Guilt, Eye-opener; DAST, Drug Abuse Screening Test; ED, emergency department; ERRS, Edinburgh Risk of Repetition; GSI, Global Severity Index; MSHR, Manchester Self-Harm Rule; MSPS, Modified SAD PERSONS Scale; ReACT, ReACT Self-Harm Rule; RESH, Repeated Episodes of Self-Harm score; ROC, receiver operating characteristic; SPS, SAD PERSONS Scale; SSHR, Söderjukuset Self-harm Rule; SUAS, Suicide Assessment Scale.
Diagnostic accuracy statistics with 95% CIs

<table>
<thead>
<tr>
<th>Scale</th>
<th>Prevalence</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>LR+</th>
<th>LR−</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIS</td>
<td>0.22</td>
<td>0.20 (0.10 to 0.36)</td>
<td>0.97 (0.94 to 0.99)</td>
<td>0.22</td>
<td>0.07 (0.03 to 0.14)</td>
<td>0.22</td>
<td>0.96 (0.93 to 0.99)</td>
</tr>
<tr>
<td>DAST</td>
<td>0.37</td>
<td>0.21 (0.12 to 0.31)</td>
<td>0.84 (0.83 to 0.85)</td>
<td>0.19</td>
<td>0.10 (0.06 to 0.15)</td>
<td>1.19</td>
<td>0.94 (0.92 to 0.96)</td>
</tr>
<tr>
<td>ERRS</td>
<td>0.30</td>
<td>0.20 (0.10 to 0.30)</td>
<td>0.97 (0.93 to 0.99)</td>
<td>0.30</td>
<td>0.10 (0.06 to 0.15)</td>
<td>1.30</td>
<td>0.97 (0.95 to 0.98)</td>
</tr>
</tbody>
</table>

*Numbers may be different from reported in original studies as the statistics were calculated using raw data where possible.

Methodological limitations

We did not conduct any meta-analyses due to the heterogeneity of the studies, nor did we calculate the receiver operating characteristics of the scales as we did not have the raw interval data. However, we provided a range of diagnostic accuracy statistics and associated CIs, which are useful in the critical evaluation of risk scales following self-harm. Some scales were tested in several settings, and we made no attempt to pool accuracy statistics across studies. Instead, we focused on a single study for each scale. This was the original study where this met inclusion criteria. We did this in order to gain an indication of the scale performance under potentially optimal conditions and because of the difficulty in pooling results from different settings.

Two scales in particular had been tested in multiple studies and settings (Edinburgh Risk of Repetition Scale and the Manchester Self-Harm Rule). Sensitivities for the Edinburgh Risk of Repetition Scale ranged from 26% to 41%, and specificities ranged from 84% to 91% in an early study. A further validation study conducted in Australia provided similar results (sensitivity: 26%, specificity: 84%). Broadly similar results were found in Oxford, for the Edinburgh Risk of Repetition Scale, but sensitivities were lower when tested on a 12-month rather than a 6-month follow-up, and ranged from 3% to 16%.

The Manchester Self-Harm Rule was validated in Sweden, Manchester and Canada. The results were similar to those of Cooper et al in demonstrating the high sensitivity (94%, 94%, 98% and 95.1%) for the studies, respectively) and low specificity (18%, 26%, 17% and 14.7%, respectively) of the scale.

We were keen to replicate the searches carried out as part of UK national guidance as far as possible. In some senses, the current paper was intended as an update of the review carried out as part of the NICE self-harm (longer term management guidelines), and we were constrained by the original methodology. Some well-known scales were not included in the NICE review on the basis of the prespecified inclusion criteria, for example, because they did not explicitly report themselves. This heterogeneity was reflected in the variation in predictive accuracy across scales. For example, the Manchester Self-Harm Rule was high in sensitivity (97%) but had low positive predictive value (22%). Conversely, the Drug Abuse Screening Test had low sensitivity (15%) but high positive predictive value (98%). Scales which scored highly on global measures of diagnostic accuracy included the Repeated Episodes of Self-Harm scale at the highest threshold (16.34), the Manchester Self-Harm Rule (10.77), the Drug Abuse Screening Test (8.66) and the Barratt Impulsivity Scale (8.25), but even these scales varied markedly in their sensitivity from 6% for the Repeated Episodes of Self-Harm scale to 97% for the Manchester Self-Harm Scale.
diagnostic accuracy outcomes), and therefore did not find their way into the current paper. Data in the papers (sensitivities ranging from 5.3% to 14.6% and specificities ranging from 93% to 97%) and from subsequent reviews indicate that in any case these older studies and scales did not have superior results to those described in our study. Inclusion of these additional scales would not have changed our findings. Although we used a published search strategy, there is a possibility that additional scales were excluded due to the search criteria and of publication bias in the included studies as some studies with negative results may not be widely accessible.

We considered the performance of these scales only in relation to people who self-harmed rather than the wider general or clinical population. However, this is an important clinical group, and in many settings risk scales are an intrinsic part of their management. Our main outcome was repeated self-harm or attempted suicide rather than suicide. While suicide is extremely important, because it is a relatively low-frequency event, it is much harder to predict. This is reflected in the poorer performance of scales in relation to suicide than repeat self-harm as outlined in UK guidance. Only two of the studies included in this review also reported suicide outcomes. The Manchester Self-Harm Rule identified 100% of the 22 suicide deaths that occurred within the 6-month follow-up period. The ReACT Rule identified 60 of the 66 suicide deaths (91%) in the derivation data set and 23 of 26 (88%) in the test data set within 6 months of the index episode. These results indicate high sensitivity, but this is once again at the expense of low specificity and poor positive predictive value. Two other studies combined suicide and repeat self-harm as an outcome, and deaths by suicide were not included in the remaining studies.

Clinical implications
What is the most useful scale following self-harm?

The use of scales is dependent on multiple factors. The scales are not directly comparable due to differences in the incidence of repeat self-harm across studies and methodological quality. Many of the studies were conducted in high-income countries in centrally funded health services, and so the findings may not be applicable to different settings. The Repeated Episodes of Self-Harm Scale was developed on an inpatient sample which is unlikely to be transferable to emergency department services. The performance of the scales may be additionally influenced by cultural contexts. For example, the Barratt Impulsivity scale was developed in the USA, and the terminology of some of the items may reduce the performance of the scale in other cultures (e.g., ‘I squirm at plays or lectures’). There is also a challenging balance when selecting scales based on diagnostic accuracy statistics, and no scale performed well across all indices.

Global indicators such as the diagnostic OR provide the strength of the association between the exposure and the disease and are readily interpreted by clinicians. False-positive and false-negative results are equally weighted, which is advantageous for research and meta-analyses, but may limit clinical use as clinicians cannot evaluate the scale on the basis of sensitivity and specificity. The scales which had the highest global diagnostic ORs were the Repeated Episodes of Self-Harm scale at the highest threshold (16.34) and the Manchester Self-Harm Rule (10.77).
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Purpose of instrument</th>
<th>Cost</th>
<th>Description</th>
<th>Time (min)</th>
<th>Copyright</th>
<th>Administration</th>
<th>Training</th>
<th>Study reference</th>
<th>Original scale reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Manchester Self-Harm Rule</td>
<td>Risk-stratification model for use with ED staff in the assessment of self-harm to discriminate between patients and higher vs lower risk of repetition or subsequent suicide by 6 months</td>
<td>Free</td>
<td>4 screening items, dichotomous answers (history of self-harm, prior psychiatric treatment, benzodiazepine overdose, current psychiatric treatment) (1=present, 0=absent), positive answer is a positive result</td>
<td>5</td>
<td>No</td>
<td>Paper and pen</td>
<td>Limited training necessary</td>
<td>Cooper et al</td>
<td>Cooper et al</td>
</tr>
<tr>
<td>The ReACT Self-Harm Rule</td>
<td>Screening tool to identify patients at higher risk of repeat self-harm suicide within 6 months of ED self-harm presentation</td>
<td>Free</td>
<td>4 items (recent self-harm (in the past year), Alone of homeless (living status), cutting used as a method of harm, and treatment for a current psychiatric disorder), presence of one or more of these items classifies patient as at higher risk of repeat self-harm/suicide within 6 months</td>
<td>5</td>
<td>No</td>
<td>Paper and pen</td>
<td>Limited training necessary</td>
<td>Steeg et al</td>
<td>Steeg et al</td>
</tr>
<tr>
<td>The RESH Self-Harm tool</td>
<td>Designed to assist clinicians in clinical management of self-harm patients</td>
<td>Free</td>
<td>4 main items with an assigned weight: number of prior episodes (0, 1, 2, 3, 4, 5, 6 or more), time between episodes (1–60 days, 61 days to 12 months, &gt; 12 months), psychiatric diagnosis in the last 12 months (substance misuse disorder, depression, anxiety, eating disorder, personality)</td>
<td>Unknown</td>
<td>No</td>
<td>Paper and Pen</td>
<td>Unknown</td>
<td>Spittal et al</td>
<td>Spittal et al</td>
</tr>
<tr>
<td>Instrument</td>
<td>Purpose of instrument</td>
<td>Cost</td>
<td>Description</td>
<td>Time (min)</td>
<td>Copyright</td>
<td>Administration</td>
<td>Training</td>
<td>Study reference</td>
<td>Original scale reference</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
<td>-----------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>The Global Severity Index (GSI)</td>
<td>The GSI symptom scale is a component of the Brief Symptom Inventory (BSI). Also a global indicator, the Symptom Checklist-90-Revised (SCL-90-R) is ‘designed to help quantify a patients severity-of-illness and provides a single composite score for measuring the outcome of a treatment programme based on reducing symptom severity’ (Pearson Assessments)</td>
<td>£118.32 per 50 answer sheets with test items, 50 profile forms and 2 worksheets (£935+VAT for 500)</td>
<td>53-item self-report on 5-point rating scale. 9 symptom dimensions (somatisation, Obsessive-compulsive, Interpersonal sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, Psychoticism) Global Indices (Global severity index, Positive Symptom Distress Index, Positive Symptom total). Also a component of the 90-item SCL-90-R (12–15 min to complete)</td>
<td>8–10</td>
<td>Pearson Assessments</td>
<td>Q Local Software, Mail-in scoring service, Hand scoring, or optimal Scan Scoring</td>
<td>B, Q1, Q2 level</td>
<td>Randall et al, Derogitis and Melisaratos</td>
<td></td>
</tr>
<tr>
<td>The SAD PERSONS scale</td>
<td>Educational tool for medical students to determine suicide risk</td>
<td>Free</td>
<td>10-item mnemonic consisting of risk factors based on literature review (male sex, age, depression, previous attempt, excess alcohol or</td>
<td>5–10</td>
<td>No</td>
<td>Paper and pen</td>
<td>Limited training necessary</td>
<td>Bolton et al, Patterson et al</td>
<td></td>
</tr>
</tbody>
</table>
Table 3  Continued

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Purpose of instrument</th>
<th>Cost</th>
<th>Description</th>
<th>Time (min)</th>
<th>Copyright</th>
<th>Administration</th>
<th>Training</th>
<th>Study reference</th>
<th>Original scale reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Modified SAD PERSONS</td>
<td>Suicide assessment in the ED</td>
<td>10-item scale. Modified SAD PERSONS by adding five additional criteria (feelings of hopelessness, history of psychiatric care, drug addiction, a 'serious' attempt, and affirmative or ambivalent answers when questioned about future intent regarding suicide. Four scale items are weighted with scores of 2 to give a total possible score of 14. Cut-off points: low (0–5), moderate (6–8), and high (9–14)</td>
<td>5–10</td>
<td>No</td>
<td>Paper and pen</td>
<td>Limited training necessary</td>
<td>Bolton et al&lt;sup&gt;24&lt;/sup&gt;</td>
<td>Hockberger and Rothstein&lt;sup&gt;39&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>The Barratt Impulsivity Scale</td>
<td>Designed to assess the personality trait of impulsiveness</td>
<td>Free</td>
<td>30 items based on personality. Self-report. Responses scored on a 4-point likert scale. Responses summed to total score</td>
<td>10</td>
<td>No</td>
<td>Paper and pen</td>
<td>None</td>
<td>Randall et al&lt;sup&gt;15&lt;/sup&gt;</td>
<td>Patton et al&lt;sup&gt;40&lt;/sup&gt;</td>
</tr>
<tr>
<td>The Drug Abuse Screening Test</td>
<td>Designed to identify patients who are abusing drugs, also</td>
<td>Free</td>
<td>28-items self-report. Responses are dichotomous (1=yes, 0=no)</td>
<td>5</td>
<td>No</td>
<td>Paper and Pen</td>
<td>Adherence to instructions for</td>
<td>Randall et al&lt;sup&gt;15&lt;/sup&gt;</td>
<td>Skinner&lt;sup&gt;41&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
Table 3  Continued

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Purpose of instrument</th>
<th>Cost</th>
<th>Description</th>
<th>Time (min)</th>
<th>Copyright</th>
<th>Administration</th>
<th>Training</th>
<th>Study reference</th>
<th>Original scale reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>to provide a</td>
<td></td>
<td>0=no, except for items 4, 5, and 7 which are reverse coded. The scale is unidimensional with a total score calculated from summing all the items that are positive in relation to increased drug use. Cut-off scores of six to 11 are used for identifying patients with drug abuse/misuse problems, whereas of 16 or above is considered as severe drug abuse or dependency</td>
<td>&gt;30=high risk</td>
<td>No</td>
<td>Paper and pen</td>
<td>Instructions available</td>
<td>Waern et al&lt;sup&gt;33&lt;/sup&gt;</td>
<td>Niméus et al&lt;sup&gt;43&lt;/sup&gt;</td>
</tr>
<tr>
<td>The Suicide Assessment Scale</td>
<td>Designed to be sensitive to change in suicidality over time and in treatment</td>
<td>Free</td>
<td>20 items, on a 0–4 likert scale summed to arrive at a maximum score of 80. Five main areas covered: (affect (5 items), bodily states (5 items), control and coping (5 items), emotional reactivity, suicidal thoughts and behaviour (5 items). Clinician and self-report versions available</td>
<td></td>
<td></td>
<td></td>
<td>Waern et al&lt;sup&gt;33&lt;/sup&gt;</td>
<td>Waern et al&lt;sup&gt;33&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Edinburgh Risk of Repetition Scale</td>
<td>Designed to identify patients at risk of repeat self-harm</td>
<td>Free</td>
<td>11 items (previous self-harm, personality disorder, alcohol, previous psychiatric care, unemployment, social class, drug</td>
<td>Men &gt;8=high risk</td>
<td>No</td>
<td>Paper and Pen</td>
<td>None</td>
<td>Carter et al&lt;sup&gt;29&lt;/sup&gt;</td>
<td>Kreitman and Foster&lt;sup&gt;44&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Continued
The balance between sensitivity and specificity is dependent on various factors such as resources, purpose of the test and stage of treatment. Clinicians may prefer a test high in sensitivity to capture as many repeat self-harm episodes as possible, for example, the Manchester Self-Harm Rule or the ReACT Self-Harm rule. Highly sensitive tests are sometimes used to screen patients or can assist in ‘ruling out’ patients as the possibility of a false negative is relatively low. The Manchester Self-Harm rule was also validated in other prospective cohort studies and similar sensitivities and specificities were reported. However, the ReACT Self-Harm Rule and Manchester Self-Harm Rule have poor specificities and positive predictive values, and there is a possibility that many patients could be false positives (ie, incorrectly labelled as at high risk), which has cost and resource implications.

Scales high in specificity, such as the Repeated Episodes of Self-Harm scale at the highest threshold, may be useful for a later stage of assessment or if treatment outcomes are expensive, medically invasive or burdensome to the patients. Scales high in specificity can also be used to ‘rule in’ patients, as the number of false positives is low (so people labelled as at high risk are quite likely to be at high risk). However, the clinical utility of high specificity scales may be limited because of the small numbers of patients who screen positive, and the fact that the high risk of the patients who reach the threshold is already fairly obvious on the basis of conventional clinical risk factors (eg, for the Repeated Episodes of Self-Harm Scale at the highest threshold, the small number of patients who have multiple prior episodes of self-harm, psychiatric diagnosis and recent psychiatric hospitalisation are clearly at elevated risk). The sensitivities of such scales in this study were poor, and there is a possibility of false negatives (people being labelled as at low risk when they are actually at high risk).

Clinicians might consider scales with high positive predictive values such as the Repeated Episodes of Self-Harm scale at the highest threshold, as positive predictive values are a measure of the probability that an individual at high risk actually goes on to repeat self-harm. However, positive predictive values are affected by how common the outcome is, which affects their transferability to clinical settings with a different incidence of repeat self-harm. The scales with high positive predictive values (eg, Repeated Episodes of Self-Harm scale at high threshold, and Global Severity Index) were also low in sensitivity, which is a further consideration when the evaluating the usefulness of scales for clinical practice.

Scales can be evaluated using likelihood ratios (probability of a specific result among people who repeat self-harm divided by the probability of a given result among people who do not repeat self-harm), and they are widely used in evidence-based medicine. They are advantageous in evaluating scales, as information from both sensitivity and specificity is used, they are not affected by prevalence, and they are fairly easy to...
CONCLUSION

On the basis of our review, it is clear that no scale appears to perform sufficiently well to be used routinely. The limitations of risk scales in clinical practice are well documented, and it is suggested that the clinical focus should be on ‘conducting comprehensive clinical assessments of each patient’s situation and needs’ rather than the categorisation of patients into high-risk and low-risk categories (p.463).50–52 The focus on risk assessment can detract from the therapeutic relationship,53 and studies have reported that patients and staff can find assessments with scales an adverse experience.8 However, risk scales continue to be widely used in self-harm services with hospitals commonly developing local instruments.7 Traditional paradigms which simply aim to balance sensitivity versus specificity may be of limited usefulness in the development of risk scales for use following self-harm. Future research should involve head-to-head comparisons. This may have more validity than comparing scales used in different patient groups across different settings. Studies need to determine the effectiveness of risk scales using robust predictive accuracy cohort studies that are clearly reported according to STARD criteria.54 Until then, it is difficult to evaluate what the most useful instruments are and, in line with clinical guidance, scales should not be used in isolation to determine management or to predict risk of future self-harm.5

Acknowledgements The authors would like to thank Dr David While for his statistical assistance and the authors of papers who provided us with additional data.

Contributors NK and JC designed the study with input from DG, KH and LD. LQ analysed the data with assistance from NK and JC. LQ, NK and JC interpreted the results and wrote the first draft. All authors contributed to subsequent drafts and have approved the final version of the manuscript.

Funding This paper presents independent research funded by the National Institute of Health Research (NIHR) under its Programme Grants for Applied Research Programme (grant reference number RP-PG-0610-10026).

Disclaimer The views expressed are those of the authors and not necessarily those of the NHS, the National Institute of Health Research or the Department of Health.

Competing interests DG, KH and NK are members of the Department of Health’s (England) National Suicide Prevention Advisory Group. NK chaired the NICE guideline development group for the longer term management of self-harm, the NICE Topic Expert Group (which developed the quality standards for self-harm services), the NICE evidence update for self-harm, and is a chair of the NICE guideline for Depression. KH and DG are NIHR Senior Investigators. KH is also supported by the Oxford Health NHS Foundation Trust and NK by the Manchester Mental Health and Social Care Trust. NK, KH and JC are authors of some of the papers and scales included in this review.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

Open Access This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/

REFERENCES


Box 1 Measures of diagnostic accuracy for scales following self-harm

<table>
<thead>
<tr>
<th>Key terms</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (Sens): Proportion of individuals who repeat self-harm identified as high risk by the test, that is, how well the test identifies patients who repeat self-harm</td>
<td>Positive likelihood ratio (LR+): How much more likely a positive test result is to occur in a patient who repeats self-harm versus one who does not repeat</td>
<td>Number allowed to diagnosis (NAD): The number of people correctly classified as having a repeat self-harm episode before an misclassification occurs</td>
</tr>
<tr>
<td>Specificity (Spec): Proportion of people who did not repeat self-harm identified as low risk by the test, that is, how well the test identifies patients who will not repeat self-harm</td>
<td>Negative likelihood ratio (LR−): How much less likely a negative test result is to occur in a patient who repeats self-harm compared to a patient who does not repeat</td>
<td></td>
</tr>
<tr>
<td>Positive predictive value (PPV): The probability that the person identified as at risk for repeat self-harm will actually repeat self-harm</td>
<td>Diagnostic OR (DOR): Overall global measure of test performance and represents the strength of the association between the test result and repeat self-harm (interpreted the same as an OR)</td>
<td></td>
</tr>
<tr>
<td>Negative predictive value (NPV): The probability that the person identified as low risk for repeat self-harm will not actually repeat self-harm</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Measures of diagnostic accuracy for scales following self-harm |
|---|---|---|
| Sensitivity (Sens): Proportion of individuals who repeat self-harm identified as high risk by the test, that is, how well the test identifies patients who repeat self-harm | Positive likelihood ratio (LR+): How much more likely a positive test result is to occur in a patient who repeats self-harm versus one who does not repeat | Number allowed to diagnosis (NAD): The number of people correctly classified as having a repeat self-harm episode before an misclassification occurs |
| Specificity (Spec): Proportion of people who did not repeat self-harm identified as low risk by the test, that is, how well the test identifies patients who will not repeat self-harm | Negative likelihood ratio (LR−): How much less likely a negative test result is to occur in a patient who repeats self-harm compared to a patient who does not repeat | |
| Positive predictive value (PPV): The probability that the person identified as at risk for repeat self-harm will actually repeat self-harm | Diagnostic OR (DOR): Overall global measure of test performance and represents the strength of the association between the test result and repeat self-harm (interpreted the same as an OR) | |
| Negative predictive value (NPV): The probability that the person identified as low risk for repeat self-harm will not actually repeat self-harm | | |

Clinicians may prefer to use a scale for predicting completed suicide, but scales which do so are perhaps more likely to have high sensitivity and be over inclusive (eg, the Manchester Self-Harm Rule27 and the ReACT Self-Harm Rule.36 Only two of the studies in this review27, 36 evaluated suicide separately as an outcome, and the predictive utility of the scales for suicide needs to be investigated further.

We were unable to examine the predictive usefulness of the scales in predicting shorter versus longer term risk of self-harm repetition due to the heterogeneity of the scales and methodological characteristics. The use of scales in predicting shorter vs longer term risk is clinically important and should be investigated further using prospective cohort studies.

Box 1 Measures of diagnostic accuracy for scales following self-harm

| Box 1 Measures of diagnostic accuracy for scales following self-harm |
|---|---|---|
| Key terms | | |
| Sensitivity (Sens): Proportion of individuals who repeat self-harm identified as high risk by the test, that is, how well the test identifies patients who repeat self-harm | Positive likelihood ratio (LR+): How much more likely a positive test result is to occur in a patient who repeats self-harm versus one who does not repeat | Number allowed to diagnosis (NAD): The number of people correctly classified as having a repeat self-harm episode before an misclassification occurs |
| Specificity (Spec): Proportion of people who did not repeat self-harm identified as low risk by the test, that is, how well the test identifies patients who will not repeat self-harm | Negative likelihood ratio (LR−): How much less likely a negative test result is to occur in a patient who repeats self-harm compared to a patient who does not repeat | |
| Positive predictive value (PPV): The probability that the person identified as at risk for repeat self-harm will actually repeat self-harm | Diagnostic OR (DOR): Overall global measure of test performance and represents the strength of the association between the test result and repeat self-harm (interpreted the same as an OR) | |
| Negative predictive value (NPV): The probability that the person identified as low risk for repeat self-harm will not actually repeat self-harm | | |

Clinicians may prefer to use a scale for predicting completed suicide, but scales which do so are perhaps more likely to have high sensitivity and be over inclusive (eg, the Manchester Self-Harm Rule27 and the ReACT Self-Harm Rule.36 Only two of the studies in this review27, 36 evaluated suicide separately as an outcome, and the predictive utility of the scales for suicide needs to be investigated further.

We were unable to examine the predictive usefulness of the scales in predicting shorter versus longer term risk of self-harm repetition due to the heterogeneity of the scales and methodological characteristics. The use of scales in predicting shorter vs longer term risk is clinically important and should be investigated further using prospective cohort studies.

Box 1 Measures of diagnostic accuracy for scales following self-harm

| Box 1 Measures of diagnostic accuracy for scales following self-harm |
|---|---|---|
| Key terms | | |
| Sensitivity (Sens): Proportion of individuals who repeat self-harm identified as high risk by the test, that is, how well the test identifies patients who repeat self-harm | Positive likelihood ratio (LR+): How much more likely a positive test result is to occur in a patient who repeats self-harm versus one who does not repeat | Number allowed to diagnosis (NAD): The number of people correctly classified as having a repeat self-harm episode before an misclassification occurs |
| Specificity (Spec): Proportion of people who did not repeat self-harm identified as low risk by the test, that is, how well the test identifies patients who will not repeat self-harm | Negative likelihood ratio (LR−): How much less likely a negative test result is to occur in a patient who repeats self-harm compared to a patient who does not repeat | |
| Positive predictive value (PPV): The probability that the person identified as at risk for repeat self-harm will actually repeat self-harm | Diagnostic OR (DOR): Overall global measure of test performance and represents the strength of the association between the test result and repeat self-harm (interpreted the same as an OR) | |
| Negative predictive value (NPV): The probability that the person identified as low risk for repeat self-harm will not actually repeat self-harm | | |

Clinicians may prefer to use a scale for predicting completed suicide, but scales which do so are perhaps more likely to have high sensitivity and be over inclusive (eg, the Manchester Self-Harm Rule27 and the ReACT Self-Harm Rule.36 Only two of the studies in this review27, 36 evaluated suicide separately as an outcome, and the predictive utility of the scales for suicide needs to be investigated further.

We were unable to examine the predictive usefulness of the scales in predicting shorter versus longer term risk of self-harm repetition due to the heterogeneity of the scales and methodological characteristics. The use of scales in predicting shorter vs longer term risk is clinically important and should be investigated further using prospective cohort studies.


Which are the most useful scales for predicting repeat self-harm? A systematic review evaluating risk scales using measures of diagnostic accuracy

L Quinlivan, J Cooper, L Davies, K Hawton, D Gunnell and N Kapur

BMJ Open 2016 6:
doi: 10.1136/bmjopen-2015-009297

Updated information and services can be found at:
http://bmjopen.bmj.com/content/6/2/e009297

These include:

Supplementary Material
Supplementary material can be found at:
http://bmjopen.bmj.com/content/suppl/2016/02/12/bmjopen-2015-009297.DC1.html

References
This article cites 44 articles, 8 of which you can access for free at:
http://bmjopen.bmj.com/content/6/2/e009297#BIBL

Open Access
This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See:
http://creativecommons.org/licenses/by/4.0/

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections

Emergency medicine (166)
Health services research (886)
Medical management (151)
Mental health (437)
Public health (1433)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/