Psychological Medicine, 2007, **37**, 769–778. © 2007 Cambridge University Press
doi:10.1017/S0033291706009664 First published online 4 January 2007 Printed in the United Kingdom

REVIEW ARTICLE

Can we predict suicide and non-fatal self-harm with the Beck Hopelessness Scale? A meta-analysis

DEAN McMILLAN[1]*, SIMON GILBODY[2], EMMA BERESFORD[1] AND LIZ NEILLY[3]

[1] *Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds, UK*; [2] *Department of Health Sciences, University of York, UK*; [3] *Health Sciences Library, University of Leeds, UK*

ABSTRACT

Background. Hopelessness is considered a pre-eminent risk factor for suicide and non-fatal self-harm. We aimed to quantify the ability of the Beck Hopelessness Scale (BHS) to predict these two outcomes.

Method. Medline, Embase, PsycINFO and Cinahl were searched to January 2006. We included cohort studies in which the BHS was applied and patients were followed-up to establish subsequent suicide or non-fatal self-harm. Four studies provided usable data on suicide, and six studies provided data on non-fatal self-harm. Summary sensitivity, specificity, likelihood ratios and diagnostic odds ratios (DORs) were calculated for each study. Random effects meta-analytic pooling across studies at the standard cut-off point ($\geqslant 9$) was undertaken and summary receiver operating characteristic (ROC) curves constructed.

Results. For suicide, pooled sensitivity was 0·80 [95% confidence interval (CI) 0·68–0·90], pooled specificity was 0·42 (95% CI 0·41–0·44), and the pooled DOR was 3·39 (95% CI 1·29–8·88). For non-fatal self-harm, pooled sensitivity was 0·78 (95% CI 0·74–0·82), pooled specificity was 0·42 (95% CI 0·38–0·45), and the pooled DOR was 2·27 (95% CI 1·53–3·37).

Conclusion. The standard cut-off point on the BHS identifies a high-risk group for potential suicide, but the magnitude of the risk is lower than previously reported estimates. The standard cut-off point is also capable of identifying those who are at risk of future self-harm, but the low specificity rate means it is unlikely to be of use in targeting treatment designed to lower the rate of repetition.

INTRODUCTION

Hopelessness, defined as generalized negative expectations of the future (Beck *et al*. 1974), has been described as one of the most important long-term risk factors for suicide in clinical populations (Joiner *et al*. 2005). In a study of clinical decision making, psychiatrists gave hopelessness the greatest weighting of any single risk factor, greater even than a previous suicide attempt or depressed mood (Truant *et al*. 1991). Current practice guidelines suggest that hopelessness should be examined when assessing suicide risk and, if present, should be targeted as part of a comprehensive treatment plan (American Psychiatric Association, 2003).

The Beck Hopelessness Scale (BHS; Beck *et al*. 1974) is the only widely used standardized measure of hopelessness (Glanz *et al*. 1995). It is a 20-item scale with a true–false response format (nine items are keyed false, 11 true); each response is summed to give a severity rating ranging from 0 to 20, with high scores indicating the presence of hopelessness. Its originators,

* Address for correspondence: Dr Dean McMillan, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds, 15 Hyde Terrace, Leeds, West Yorkshire LS2 9LT, UK.
(Email: d.mcmillan@leeds.ac.uk)

Beck and colleagues, report that it successfully predicts suicide in a sample of out-patients (Brown *et al*. 2000) and in-patients (Beck *et al*. 1985). The authors acknowledge that the low base rate of suicide means that any scale is likely to identify a large number of false positives, but they argue that the BHS may be useful in identifying a group who are at an increased risk of that outcome (Beck *et al*. 1990). In support of this they found that people scoring above the standard cut-off point on the scale (a score of 9 or more) were 11 times more likely to die by suicide than those scoring below that point.

Ten per cent of people who self-harm will eventually die by suicide, and about a quarter of suicides will have been preceded by an episode of non-fatal self-harm in the previous year (Owens & House, 1994). Although non-fatal self-harm and suicide can be distinguished on some risk factors, such as the patterning of psychiatric disorders, many of the risk factors appear to be common to both (Beautrais, 2001). The link between non-fatal self-harm and suicide may indicate that the BHS is also a predictor of non-fatal self-harm, a suggestion supported by a number of studies (Evans *et al*. 2004).

Although earlier reviews found only limited evidence for the effectiveness of interventions designed to reduce the rate of self-harm repetition (Hawton *et al*. 1998), recent reviews are more optimistic (Hepp *et al*. 2004; Mann *et al*. 2005). There are now a number of treatment approaches that appear to reduce repetition, including cognitive behaviour therapy (CBT; Brown *et al*. 2005), brief psychodynamic-interpersonal therapy (Guthrie *et al*. 2001), and individual and group psychoanalytic psychotherapy (Bateman & Fonagy, 1999). The base rate of self-harm is substantially higher than suicide, particularly among some clinical groups such as those who have had a previous episode of non-fatal self-harm. This higher base rate and the emergence of effective interventions raises the possibility that standardized scales, such as the BHS, may have a role in targeting interventions at those who are most at risk of repeating self-harm.

Some doubts, however, have been raised about the utility of the BHS. Not all studies have found that the scale successfully predicts suicide or self-harm. Beck *et al*. (1989), for example, found that the scale did not predict suicide in an alcohol-abuse group. Hawton *et al*. (1999) used the BHS in a group of adolescents who had self-harmed and found that it did not predict repetition. In those studies that have found it to predict suicide or self-harm, the information needed to comprehensively assess the performance of the test is often not reported. Although a number of studies report statistical significance, few report details of sensitivity and specificity analyses, the recommended method for assessing test performance (Sackett *et al*. 1991). Sensitivity is the proportion of the group with the outcome (e.g. suicide, non-fatal self-harm) that the predictive instrument correctly classifies as at risk. Specificity is the proportion of the group without the outcome correctly classified as not at risk. Moreover, studies of suicide that report these data have found the BHS to have moderate to low specificity at high levels of sensitivity (Glanz *et al*. 1995), a particular problem when the base rate of the outcome is likely to be low, as it is with suicide.

Over the past decade the medical literature has seen an increased use of meta-analytic strategies to assess the diagnostic and predictive accuracy of tests (Deville *et al*. 2002), but as yet examples in the psychiatric literature are rare. This is the first systematic review to use these techniques to assess the accuracy of the BHS as a predictor of non-fatal self-harm and suicide. The review aimed to answer three questions. First, to what extent does the standard cut-off point on the BHS identify a high-risk group for suicide? Second, does this standard cut-off point identify a high-risk group for non-fatal self-harm? Third, is the BHS a useful method of identifying those people who are most at risk for self-harm so that they can be targeted for intensive treatments designed to lower that risk?

METHOD

Data sources

The search for primary studies used four electronic databases: Medline, Cinahl, Embase and PsycINFO. Each database was searched from its earliest date to January 2006. MESH terms, individualized for each database, and a list of keywords were developed for suicide and non-fatal self-harm (e.g. overdose, parasuicide, self-injury, self-poisoning, suicide) and cohort design (e.g. longitudinal, follow-up, prospective). An adapted version of an optimal search

strategy for diagnostic tests (Deville *et al*. 2002) was also used. 'Hopelessness' was not included as a search term because during piloting we identified a number of studies that met inclusion criteria but made no reference to hopelessness or synonym in the title or abstract. Instead, the initial set of retrieved abstracts was examined to assess whether the study used a cohort design and had suicide or self-harm as one of the outcomes.

Study selection

Of studies that met this initial inclusion criteria, we obtained copies of the full article if either the title or abstract referred to hopelessness or the reference list for the study, accessed through Web of Knowledge, included one of the two standard methods of referencing the BHS (BHS manual or original published description of the measure; Beck *et al*. 1974). The reference list of included studies was also examined to search for additional papers that met inclusion criteria.

In cases where more than one paper referred to the same sample, we selected the one reporting data on the largest number of people. Data from studies with smaller samples were used only if it was not possible to obtain the information necessary to quantify predictive accuracy on a larger dataset.

The full articles were examined independently by two reviewers and were included if the study used a cohort design, the BHS was administered at time 1, data were available on suicide or self-harm at time 2, and the number of people with the outcome (suicide or self-harm) was at least 10. The definition of non-fatal self-harm used in the selection of studies was based on the definition of parasuicide developed for the World Health Organization/European Study on Parasuicide (Platt *et al*. 1992). No inclusion criteria were imposed in terms of patient population or length of follow-up. The weighted κ for the decision to include a study was 1·0.

Data extraction

Two reviewers independently undertook data extraction and quality assessment using a standardized form. Information extracted included details of the referral source, sampling procedure, sample size, sample characteristics, length of follow-up, and the assessment of predictive accuracy of the BHS. Intra-class correlation coefficients (ICCs) were used to assess the reliability. ICCs ranged between 0·79 and 1·0, and the majority (73%) were above 0·90.

We constructed 2×2 tables for a cut-off point of 9 and used these to calculate sensitivity, specificity and likelihood ratios (positive and negative) (Sackett *et al*. 1991). We summarized likelihood ratios in preference to predictive value because these are fundamental attributes of an instrument that do not vary according to baseline risk of an outcome. We also calculated the diagnostic odds ratio (DOR), that is the ratio of the odds of a positive test among those with subsequent self-harm/suicide to the odds of a positive result among those without subsequent self-harm/suicide (Deville *et al*. 2002). If the study met inclusion criteria but did not provide sufficient information to quantify predictive accuracy, we contacted authors to request this additional information.

Quantitative data synthesis

We undertook a random effects meta-analysis (DerSimonian & Laird, 1986) to obtain pooled estimates of sensitivity, specificity, positive and negative likelihood ratios, and a summary DOR. Between-study heterogeneity was assessed using the I^2 statistic (Higgins *et al*. 2003), which describes the percentage of total variation across studies that is due to heterogeneity rather than chance. The I^2 statistic has several advantages over other measures of heterogeneity, including greater statistical power to detect clinical heterogeneity when fewer studies are available. As a guide, I^2 values of 25% may be considered 'low', 50% 'moderate' and 75% 'high'.

Publication and small study bias was examined by constructing Begg funnel plots (Begg, 1994) of the logit DOR, and by testing for funnel plot asymmetry using Egger's weighted regression test (Egger *et al*. 1997).

Receiver operating characteristic (ROC) curves represent the most informative way of describing the inherent trade-offs between sensitivity and specificity for a diagnostic test or predictive instrument (Knottnerus, 2002). Single plots of sensitivity and specificity were therefore calculated in ROC space for all studies. We then constructed summary ROC (sROC) curves (Deeks, 2000) within the range of data, using the method of Moses *et al*. (1993). Inverse variance

Table 1. *Characteristics of included studies*

Study	Country, population and sample size	No. (%) with outcome	Maximum follow-up (years)	Sensitivity	Specificity	Likelihood ratio +ve	Likelihood ratio −ve
Outcome suicide							
Beck *et al.* (1985)	USA In-patient suicidal ideation ($n=165$)	11 (6·7)	5	0·91 (0·59–1·0)	0·51 (0·43–0·59)	1·84 (1·44–2·36)	0·18 (0·03–1·17)
Beck *et al.* (1990)	USA Out-patient ($n=1958$)	17 (0·9)	7·5	0·94 (0·71–1·0)	0·41 (0·39–0·43)	1·60 (1·41–1·81)	0·14 (0·02–0·96)
Nimeus *et al.* (1997)	Sweden Previous self-harm ($n=212$)	13 (6·1)	8	0·77 (0·46–0·95)	0·42 (0·35–0·49)	1·32 (0·96–1·82)	0·55 (0·20–1·51)
Suominen *et al.* (2004)[a]	Finland Previous self-harm ($n=224$)	17 (7·6)	12	0·60 (0·32–0·84)	0·52 (0·44–0·59)	1·24 (0·80–1·92)	0·78 (0·41–1·46)
Outcome self-harm							
Colman *et al.* (2004)[a]	Canada Previous self-harm ($n=369$)	92 (24·9)	1	0·71 (0·62–0·80)	0·54 (0·49–0·60)	1·56 (1·31–1·86)	0·53 (0·39–0·72)
Goldston *et al.* (2001)	USA In-patient adolescent ($n=180$)	45 (25·0)	4	0·84 (0·71–0·94)	0·49 (0·40–0·58)	1·65 (1·34–2·03)	0·31 (0·16–0·64)
Hawton *et al.* (2003)[a]	UK Previous self-harm ($n=111$)	34 (30·6)	1·7	0·77 (0·59–0·90)	0·30 (0·20–0·41)	1·10 (0·87–1·40)	0·76 (0·36–1·58)
Keller & Wolfersdorf (1993)	Germany In-patient depression ($n=61$)	10 (16·4)	1	0·80 (0·44–0·98)	0·29 (0·18–0·44)	1·13 (0·79–1·62)	0·68 (0·18–2·52)
Sidley *et al.* (1999)[a]	UK Previous self-harm ($n=65$)	25 (38·5)	1	0·75 (0·55–0·89)	0·36 (0·19–0·56)	1·17 (0·83–1·65)	0·70 (0·31–1·58)
Tyrer *et al.* (2003)[a]	UK Previous self-harm ($n=430$)	183 (42·6)	1	0·81 (0·74–0·86)	0·28 (0·22–0·34)	1·12 (1·01–1·25)	0·69 (0·48–0·99)

[a] Data for sensitivity, specificity and likelihood ratios provided by authors.

of logits of the DOR were used to assign weights to each individual study. We also calculated a pooled area under the curve (AUC) with 95% confidence intervals (CIs) from our modelled sROC curve. An AUC value approaching 1·0 represents a perfect test and 0·5 is an uninformative test (Hanley & McNeil, 1982).

Where there was significant between-study heterogeneity, we sought to explore its causes. We first visually inspected forest plots to look for obvious outliers and we constructed Galbraith plots of the logit of the DOR to identify studies lying outside of ± 2 pooled standard deviations (Galbraith, 1988). We also attempted to undertake a meta-regression analysis of our logit sROC model using potential predictive covariates (Thompson & Higgins, 2002), and examined the effect of these in reducing between-study heterogeneity, evidenced by reductions in the I^2 statistic. *A priori* causes of heterogeneity examined in these analyses included the study population, length of follow-up and baseline risk of self-harm/suicide. All meta-regression analyses were conducted using a permutation test to minimize type 1 errors, with 1000 Monte Carlo replications to generate p values (Higgins & Thompson, 2004).

We conducted meta-analysis using STATA version 8 (StataCorp, College Station, TX, USA) and the user-written commands: metan, metareg, metabias and galbr.

RESULTS

Description of included studies

The initial broad search of abstracts identified over 28 000 references, of which 1600 met initial inclusion criteria (cohort design, suicide or self-harm as an outcome). Sixty-eight studies consisting of 51 independent samples either mentioned hopelessness in the abstract or included a reference to the BHS in the reference list. Of these, 19 met full inclusion criteria (cohort study, BHS measured at time 1, suicide or self-harm measured at time 2, n with outcome $\geqslant 10$). Five of the 19 provided sufficient information to quantify predictive accuracy; for a further five we were able to obtain these data by contacting the authors.

Table 1 summarizes the characteristics of the 10 studies. Four studies ($n=2559$) provided data on suicide. Three had a similar rate of

https://doi.org/10.1017/S0033291706009664 Published online by Cambridge University Press

suicide in the range 6–8% (Beck *et al*. 1985; Nimeus *et al*. 1997; Suominen *et al*. 2004). Participants in these studies had been admitted to hospital because of severe suicidal ideation or an episode of self-harm and so represent a particularly high-risk group for suicide. The one study with a substantially lower rate (0·9%) used a sample recruited from an out-patient setting (Beck *et al*. 1990). The length of follow-up varied substantially within three of the four studies depending on when a person had been recruited. Each of the four samples had more females than males; the proportion of females in the samples ranged from 54% (Beck *et al*. 1985) to 58% (Beck *et al*. 1990). All of the studies used adult samples.

The remaining six studies ($n=1216$) provided data on the prediction of self-harm. Four of the samples were recruited from attendances at an emergency department after an episode of self-harm (Sidley *et al*. 1999; Hawton *et al*. 2003; Tyrer *et al*. 2003; Colman *et al*. 2004); in these studies the BHS was used to predict repetition. The two remaining studies (Keller & Wolfersdorf, 1993; Goldston *et al*. 2001) used in-patient psychiatric samples not recruited on the basis of previous self-harm. Most studies used a follow-up period of 1 year; only one study used a substantially longer follow-up (Goldston *et al*. 2001). The proportion of females in the studies ranged from 45% (Sidley *et al*. 1999) to 68% (Tyrer *et al*. 2003). With the exception of one study, which used an adolescent sample (Goldston *et al*. 2001), all examined self-harm in adults. Three of the six studies also provided information on people who had died by suicide during the follow-up period. In all of the studies the number was small (range 2–7), and these were included as part of the outcome group in the analyses reported here.

Quantitative analysis of suicide prediction

When we combined psychometric attributes across the four suicide studies, we found low to moderate between-study heterogeneity, except in the case of specificity, and no evidence of publication or small study bias ($p=0·12$). Pooled sensitivity was 0·80 (95% CI 0·68–0·90, $I^2=57\%$) and specificity was 0·42 (95% CI 0·41–0·44, $I^2=76\%$). The likelihood ratio for a positive test was 1·55 (95% CI 1·31–1·83, $I^2=44\%$) and the likelihood ratio for a negative test was 0·45 (95% CI 0·20–1·03, $I^2=49\%$). The pooled DOR was 3·39 (95% CI 1·29–8·88, $I^2=37\%$) (see Fig. 1 and Table 2).

When we summarized individual studies within ROC space, we found tests were gathered around the top-middle section of the graph, and a summary ROC curve provides a line of best fit, with 95% confidence limits, between studies for the cut-off point of 9 or over (see online Appendix). The pooled AUC was 0·70 (95% CI 0·59–0·85). The small number of studies meant that we were unable to explore sources of between-study heterogeneity using meta-regression.

Quantitative analysis of non-fatal self-harm prediction

For the self-harm studies, we found low to moderate between-study heterogeneity, again except for specificity, and no evidence of publication or small study bias ($p=0·93$). Pooled sensitivity was 0·78 (95% CI 0·74–0·82, $I^2=0\%$) and specificity was 0·42 (95% CI 0·38–0·45, $I^2=90\%$). The pooled likelihood ratio for a positive test was 1·29 (95% CI 1·09–1·52, $I^2=74\%$) and the likelihood ratio for a negative test was 0·58 (95% CI 0·47–0·71, $I^2=0\%$). The pooled DOR was 2·27 (95% CI 1·53–3·37, $I^2=35\%$) (see Fig. 2 and Table 2). We conducted a sensitivity analysis by removing the results of Goldston *et al*. (2001), which used an adolescent population; the results were not substantially altered by the removal of this study (sensitivity = 0·77, 95% CI 0·72–0·81; specificity = 0·40, 95% CI 0·37–0·44; positive likelihood ratio = 1·22, 95% CI 1·04–1·44; negative likelihood ratio = 0·61, 95% CI 0·49–0·76). We also conducted a further analysis in which we removed data from both Goldston *et al*. (2001) and Keller & Wolfersdorf (1993), which, unlike the other four studies, were not predicting the *repetition* of non-fatal self-harm. The removal of these two studies also produced similar results to the full analysis (sensitivity = 0·77; 95% CI 0·72–0·81; specificity = 0·41, 95% CI 0·37–0·45; positive likelihood ratio = 1·23, 95% CI 1·02–1·49; negative likelihood ratio = 0·61, 95% CI 0·49–0·76).

When we summarized individual studies within ROC space, we found tests were gathered within the top-left corner of the graph, and a summary ROC curve provides a line of best fit,

https://doi.org/10.1017/S0033291706009664 Published online by Cambridge University Press

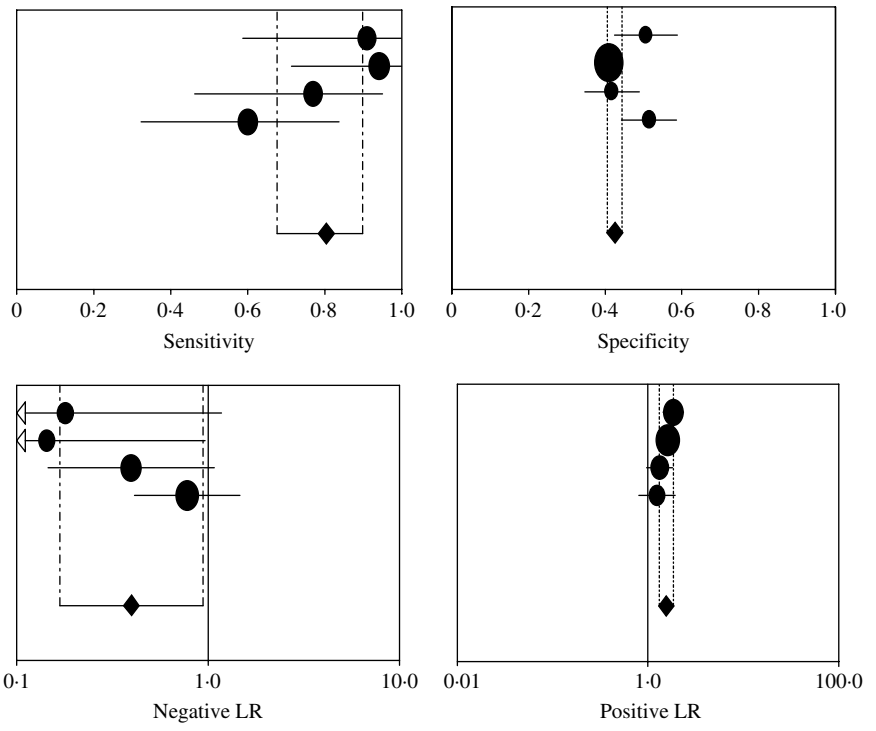

FIG. 1. Random effects pooled sensitivity, specificity and likelihood ratios (LR) for suicide studies. (Data from the following studies: Beck *et al*. 1985, 1990; Nimeus *et al*. 1997; Suominen *et al*. 2004.)

with 95% confidence limits, between studies for the cut-off point of 9 or over (see online Appendix). The pooled AUC was 0·63 (95% CI 0·57–0·70).

When we conducted a meta-regression, we found that study setting (emergency department *versus* in-patient ($\beta = 0{\cdot}74$, 95% CI $-0{\cdot}51$ to 0·99, $p = 0{\cdot}18$, $I^2 = 93\%$), length of follow-up ($\beta = 0{\cdot}31$, 95% CI $-0{\cdot}16$ to 0·79, $p = 0{\cdot}14$, $I^2 = 93\%$) and baseline risk ($\beta = -0{\cdot}04$, 95% CI $-0{\cdot}10$ to 0·03, $p = 0{\cdot}19$, $I^2 = 71\%$) were not significantly related to the DOR.

DISCUSSION

This is the first systematic review to provide a quantitative summary of the performance of the BHS in predicting suicide and self-harm. We identified four cohort studies that examined suicide (total $n = 2559$) and six that examined non-fatal self-harm ($n = 1216$).

The accurate prediction of suicide is inevitably compromised by the low base rate of the event, which leads to substantial over-inclusion even for scales with high sensitivity and specificity. However, as Beck *et al*. (1990) argue, the BHS is designed to identify a *potential* for suicide, rather than behaviour itself. In support of this, they found that people scoring 9 or above on the measure were 11 times more likely to kill themselves than those scoring less than 9, which corresponds to the DOR of 11·12 (95% CI 1·45–84·04) calculated in our analysis for that study. The other paper by Beck *et al*. (1985) reported a similarly high DOR (10·26; 95% CI 1·28–82·13). However, our main finding for the suicide studies is that the capacity of the BHS to identify this suicide potential is less than that reported in the original validation studies. Two subsequent studies (Nimeus *et al*. 1997; Suominen *et al*. 2004) found a notably lower DOR and the pooled DOR (3·39; 95% CI 1·29–8·88) was also substantially lower than that quoted in the research by Beck and colleagues. Although a score of 9 or above on the BHS does confer an increased risk of suicide, and therefore

https://doi.org/10.1017/S0033291706009664 Published online by Cambridge University Press

Table 2. *Random-effects meta-analyses of predictive accuracy*

Outcome	Psychometric attribute	Random effects pooled estimate (95% CI)	I^2 between study variance (%)
Suicide (four studies)	Sensitivity	0·80 (0·68–0·90)	57
	Specificity	0·42 (0·41–0·44)	76
	+ve likelihood ratio	1·55 (1·31–1·83)	44
	−ve likelihood ratio	0·45 (0·20–1·03)	49
	Diagnostic odds ratio	3·39 (1·29–8·88)	37
Self-harm (six studies)	Sensitivity	0·78 (0·74–0·82)	0
	Specificity	0·42 (0·38–0·45)	90
	+ve likelihood ratio	1·29 (1·09–1·52)	74
	−ve likelihood ratio	0·58 (0·47–0·71)	0
	Diagnostic odds ratio	2·27 (1·53–3·37)	35

CI, Confidence interval.

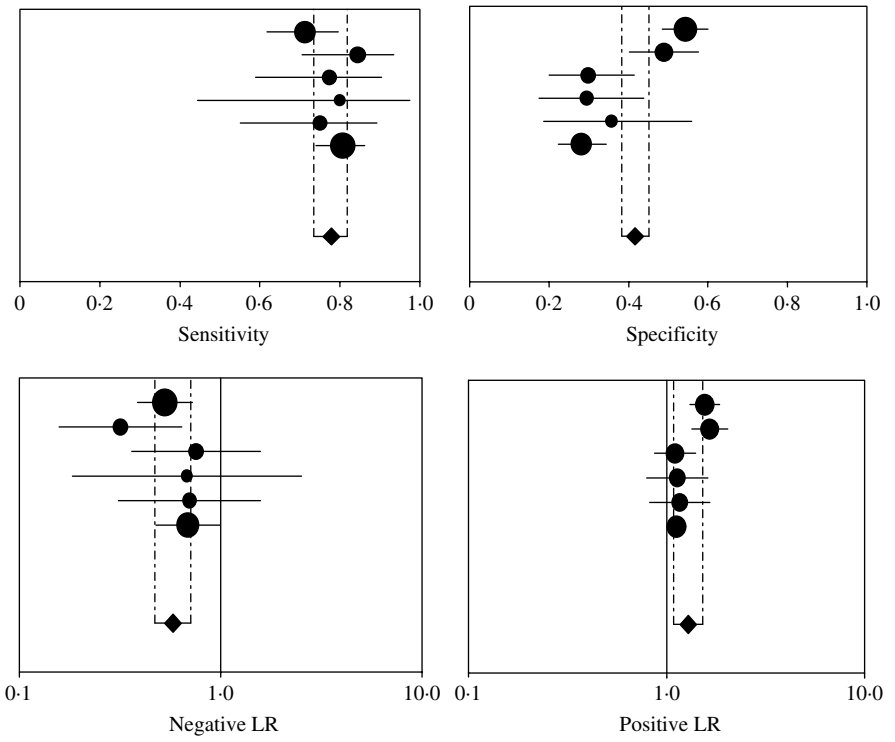

FIG. 2. Random effects pooled sensitivity, specificity and likelihood ratios (LR) for non-fatal self-harm studies. (Data from the following studies: Keller & Wolfersdorf, 1993; Sidley *et al.* 1999; Goldston *et al.* 2001; Hawton *et al.* 2003; Tyrer *et al.* 2003; Colman *et al.* 2004.)

identifies a suicide-potential group, the magnitude of that risk appears to be less than previously reported. There were too few studies to explore how methodological differences between the studies could account for this difference. However, the two later studies were conducted in Scandinavia and used translated versions of the BHS. It is possible that the psychometric properties of the translations are substantially different from the English version.

The pooled DOR for the self-harm studies is 2·27 (95% CI 1·53–3·37), which suggests that the standard cut-off point on the BHS does identify a group that is at increased risk of future

https://doi.org/10.1017/S0033291706009664 Published online by Cambridge University Press

self-harm. This does not necessarily suggest, however, that the scale would be of use in identifying a group at high risk of repetition so that it can be targeted for treatment. Previous systematic reviews of treatments for self-harm have concluded that psychological and pharmacological interventions have limited effects on the likelihood of future self-harm (Hawton *et al.* 1998), but a number of more recent studies and reviews suggest that psychological interventions can significantly reduce the repetition rate (Bateman & Fonagy, 1999; Guthrie *et al.* 2001; Hepp *et al.* 2004; Brown *et al.* 2005; Mann *et al.* 2005). These treatments are costly and intensive, so it is likely that only a small proportion of a group who may be at risk for self-harm, such as those attending an emergency setting after an initial episode of self-harm, could be provided with these interventions. Effective methods for identifying those at highest risk of repetition may be required to target treatment, and the data from our analysis can be used to assess whether the BHS would be a useful aid to decision making. The six self-harm studies and the four that examine repetition of self-harm in accident and emergency samples produce very similar sensitivity (0·78 and 0·77) and specificity estimates (0·42 and 0·41), but we use the values for the four repetition studies, because these more closely match the prediction situation under discussion. The pooled sensitivity of 0·77 would identify the majority of people who would repeat self-harm in the absence of intensive treatment, but the low pooled specificity rate (0·41) means that the BHS would identify more people as at risk than could be feasibly offered treatment. This can be illustrated by using data on the rate of repetition for people seen in an emergency setting after an episode of self-harm. A systematic review of these data found a repetition rate of 16% over the course of a year (Owens *et al.* 2002). If we combine this rate with the pooled sensitivity and specificity values, then over half (62%) of people would be identified as at risk of repetition according to a cut-off point of 9 on the BHS, a higher figure than could be offered an intensive treatment such as CBT or psychodynamic-interpersonal therapy. It remains possible that the BHS may be of use if a higher cut-off point were to be selected, but this cannot be assessed from the data reported here.

It is also unclear whether the BHS at the standard cut-off point would be of use as a screening tool to identify people for a more intensive risk assessment. Screening tests with high sensitivity may be acceptable even when accompanied by low or moderate specificity, because the subsequent stage can be used to more clearly differentiate those at risk from those who are not. However, this may not apply in this decision-making context. The screening test typically needs to produce a higher level of sensitivity than that reported for the BHS at a cut-off point of 9. At this cut-off point, about one-quarter of people who would subsequently self-harm would be classified as not at risk, a miss rate that may be considered unacceptably high. Furthermore, in the typical multi-stage screening process, the subsequent stages of assessment are expected to deliver a much more accurate assessment of risk than the initial screen, but it is unclear whether such a test exists for non-fatal self-harm. No other psychometric measure has markedly outperformed the BHS and the limited evidence on clinical prediction for this type of risk suggests that a clinical interview may not improve accuracy (Janofsky *et al.* 1988).

We note that nine of the 19 studies identified as containing potentially useful data could not be included in the analysis because, despite our best efforts in corresponding with authors, we were unable to obtain the necessary information required to construct 2×2 tables for the meta-analysis. As may be expected, we found it difficult to obtain data from studies conducted some time ago, either because we were unable to contact the authors or because the authors stated that the original data were no longer available. Although the absence of these nine studies may have introduced a selection bias, it is difficult to speculate on whether this would have improved or worsened the pooled sensitivity and specificity values. Specific guidelines on the publication of diagnostic and predictive studies have been published to improve the quality of data that are made available (Bossuyt *et al.* 2003). Adherence to these guidelines will make it easier to carry out future reviews of the performance of the BHS.

The small number of studies included in our analyses meant that we were unable to conduct a meta-regression for suicide and had to be conservative in our use of predictive covariates

in our meta-regression for non-fatal self-harm. We probably lacked sufficient statistical power to detect a substantial effect of population and study setting characteristics in predicting the psychometric performance of the BHS instrument. For example, we did not find a relationship between length of follow-up and psychometric properties, when one might logically have been expected to exist. A useful future exercise will be to repeat the analytic methods outlined in this paper as and when further research emerges relating to the predictive properties of the BHS. A greater number of individual studies with a greater range of settings, populations and periods of follow-up will allow the causes of underlying heterogeneity of the performance characteristics to be explored in more detail using techniques such as meta-regression.

This systematic review confirms that hopelessness, as measured by the BHS, is a risk factor for both suicide and non-fatal self-harm. However, our findings suggest two important caveats to this general conclusion. First, the extent to which a score of 9 or above on the BHS indicates an increased risk of suicide is notably lower than that reported in the original validation studies. Second, the standard cut-off point on the BHS is unlikely to be of use in identifying a group for treatment designed to reduce the rate of repetition of non-fatal self-harm, because that cut-off point would identify a higher proportion of people as at risk than could be feasibly offered treatment.

ACKNOWLEDGEMENTS

We thank the following researchers for their help in making data available to us: Ian Colman, Keith Hawton, Andrew Macleod, Stephen Newman, Gary Sidley, Julia Sinclair, Kirsi Suominen and Peter Tyrer.

DECLARATION OF INTEREST

None.

NOTE

Supplementary information accompanies this paper on the Journal's website (http://journals.cambridge.org).

REFERENCES

American Psychiatric Association (2003). Practice guideline for the assessment and treatment of patients with suicidal behaviors. *American Journal of Psychiatry* **160**, 1–60.

Bateman, A & Fonagy, P. (1999). Effects of partial hospitalization in the treatment of borderline personality disorder: a randomized controlled trial. *American Journal of Psychiatry* **156**, 1563–1569.

Beautrais, A. L. (2001). Suicides and serious suicide attempts: two populations or one? *Psychological Medicine* **31**, 837–845.

Beck, A. T., Brown, G., Berchick, R. J., Stewart, B. L. & Steer, R. A. (1990). Relationship between hopelessness and ultimate suicide: a replication with psychiatric outpatients. *American Journal of Psychiatry* **147**, 190–195.

Beck, A. T., Steer, R. A., Kovacs, M. & Garrison, B. (1985). Hopelessness and eventual suicide: a 10-year prospective study of patients hospitalized with suicidal ideation. *American Journal of Psychiatry* **142**, 559–563.

Beck, A. T., Steer, R. A. & Trexler, L. D. (1989). Alcohol abuse and eventual suicide: a 5- to 10-year prospective study of alcohol-abusing suicide attempters. *Journal of Studies on Alcohol* **50**, 202–209.

Beck, A. T., Weissman, A., Lester, D. & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology* **42**, 861–865.

Begg, C. B. (1994). Publication bias. In *The Handbook of Research Synthesis* (ed. L. V. Hedges), pp. 399–409. Russell Sage Foundation: New York.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Lijmer, J. G., Moher, D., Rennie, D. & de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *British Medical Journal* **326**, 41–44.

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. *Journal of Consulting and Clinical Psychology* **68**, 371–377.

Brown, G. K., Have, T. T., Henriques, G. R., Xie, S. X., Hollander, J. E. & Beck, A. T. (2005). Cognitive therapy for the prevention of suicide attempts: a randomized controlled trial. *Journal of the American Medical Association* **294**, 563–570.

Colman, I., Newman, S. C., Schopflocher, D., Bland, R. C. & Dyck, R. J. (2004). A multivariate study of predictors of repeat parasuicide. *Acta Psychiatrica Scandinavica* **109**, 306–312.

Deeks, J. (2000). Evaluations of diagnostic and screening tests. In *Systematic Reviews in Health Care* (ed. D. G. Altman), pp. 248–282. BMJ Books: London.

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

Deville, W. L., Buntinx, F., Bouter, L. M., Montori, V. M., de Vet, H. C., van der Windt, D. A. & Bezemer, P. D. (2002). Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Medical Research Methodology* **2**, 9.

Egger, M., Davey-Smith, G., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629–634.

Evans, E., Hawton, K. & Rodham, K. (2004). Factors associated with suicidal phenomena in adolescents: a systematic review of population-based studies. *Clinical Psychology Review* **24**, 957–979.

Galbraith, R. F. (1988). A note on the graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine* **7**, 889–884.

Glanz, L. M., Haas, G. L. & Sweeney, J. A. (1995). Assessment of hopelessness in suicidal patients. *Clinical Psychology Review* **15**, 49–64.

Goldston, D. B., Daniel, S. S., Reboussin, B. A., Reboussin, D. M., Frazier, P. H. & Harris, A. E. (2001). Cognitive risk factors and suicide attempts among formerly hospitalized adolescents: a prospective naturalistic study. *Journal of the American Academy of Child and Adolescent Psychiatry* **40**, 91–99.

Guthrie, E., Kapur, N., Mackway-Jones, K., Chew-Graham, C., Moorey, J., Mendel, E., Marino-Francis, F., Sanderson, S., Turpin, C., Boddy, G. & Tomenson, B. (2001). Randomised control trial of brief psychological intervention after deliberate self poisoning. *British Medical Journal* **323**, 135–138.

Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Hawton, K., Arensman, E., Townsend, E., Bremner, S., Feldman, E., Goldney, R., Gunnell, D., Hazell, P., van Heeringen, K., House, A., Owens, D., Sakinofsky, I. & Traskman-Bendz, L. (1998). Deliberate self harm: systematic review of efficacy of psychosocial and pharmacological treatments in preventing repetition. *British Medical Journal* **317**, 441–447.

Hawton, K., Houston, K., Haw, C., Townsend, E. & Harriss, L. (2003). Comorbidity of axis I and axis II disorders in patients who attempted suicide. *American Journal of Psychiatry* **160**, 1494–1500.

Hawton, K., Kingsbury, S., Steinhardt, K., James, A. & Fagg, J. (1999). Repetition of deliberate self-harm by adolescents: the role of psychological factors. *Journal of Adolescence* **22**, 369–378.

Hepp, U., Wittmann, L., Schnyder, U. & Michel, K. (2004). Psychological and psychosocial interventions after suicide: an overview of treatment studies. *Crisis* **25**, 108–117.

Higgins, J. P., Thompson, S. G., Deeks, J. J. & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal* **327**, 557–560.

Higgins, J. P. T. & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine* **23**, 1663–1682.

Janofsky, J. S., Spears, S. & Neubauer, D. N. (1988). Psychiatrist's accuracy in predicting violent behavior on an inpatient unit. *Hospital and Community Psychiatry* **39**, 1090–1094.

Joiner, T. E., Jr., Brown, J. S. & Wingate, L. R. (2005). The psychology and neurobiology of suicidal behavior. *Annual Review of Psychology* **56**, 287–314.

Keller, F. & Wolfersdorf, M. (1993). Hopelessness and the tendency to commit suicide in the course of depressive disorders. *Crisis: Journal of Crisis Intervention and Suicide* **14**, 173–177.

Knottnerus, J. A. (ed.) (2002). *The Evidence Base of Clinical Diagnosis*. BMJ Publishing: London.

Mann, J. J., Apter, A., Bertolote, J., Beautrais, A., Currier, D., Haas, A., Hegerl, U., Lonnqvist, J., Malone, K., Marusic, A., Mehlum, L., Patton, G., Phillips, M., Rutz, W., Rihmer, Z., Schmidtke, A., Shaffer, D., Silverman, M., Takahashi, Y., Varnik, A., Wasserman, D., Yip, P. & Hendin, H. (2005). Suicide prevention strategies: a systematic review. *Journal of the American Medical Association* **294**, 2064–2074.

Moses, L., Shapiro, D. & Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* **12**, 1293–1316.

Nimeus, A., Traskman-Bendz, L. & Alsen, M. (1997). Hopelessness and suicidal behavior. *Journal of Affective Disorders* **42**, 137–144.

Owens, D., Horrocks, J. & House, A. (2002). Fatal and non-fatal repetition of self-harm. Systematic review. *British Journal of Psychiatry* **181**, 193–199.

Owens, D. & House, A. (1994). General hospital services for deliberate self-harm. *Journal of the Royal College of Physicians of London* **28**, 370–371.

Platt, S., Bille-Brahe, U., Kerkhof, A., Schmidtke, A., Bjerke, T., Crepet, P., De Leo, D., Haring, C., Lonnqvist, J., Michel, K., Philippe, A., Pommereau, X., Querejeta, I., Salander-Renberg, E., Temesvary, B., Wasserman, D. & Sampaio Faria, J. (1992). Parasuicide in Europe: the WHO/EURO multicentre study on parasuicide. I. Introduction and preliminary analysis for 1989. *Acta Psychiatrica Scandinavica* **85**, 97–104.

Sackett, D. L., Haynes, R. B., Guyatt, G. H. & Tugwell, P. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Little, Brown and Company: Boston, MA.

Sidley, G. L., Calam, R., Wells, A., Hughes, T. & Whitaker, K. (1999). The prediction of parasuicide repetition in a high-risk group. *British Journal of Clinical Psychology* **38**, 375–386.

Suominen, K., Isometsa, E., Ostamo, A. & Lonnqvist, J. (2004). Level of suicidal intent predicts overall mortality and suicide after attempted suicide: a 12-year follow-up study. *BMC Psychiatry* **4**, 11.

Thompson, S. G. & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* **21**, 1559–1573.

Truant, G., O'Reilly, R. & Donaldson, L. (1991). How psychiatrists weigh risk factors when assessing suicide risk. *Suicide and Life-Threatening Behavior* **21**, 106–114.

Tyrer, P., Thompson, S., Schmidt, U., Jones, V., Knapp, M., Davidson, K., Catalan, J., Airlie, J., Baxter, S., Byford, S., Byrne, G., Cameron, S., Caplan, R., Cooper, S., Ferguson, B., Freeman, C., Frost, S., Godley, J., Greenshields, J., Henderson, J., Holden, N., Keech, P., Kim, L., Logan, K., Manley, C., MacLeod, A., Murphy, R., Patience, L., Ramsay, L., De Munroz, S., Scott, J., Seivewright, H., Sivakumar, K., Tata, P., Thornton, S., Ukoumunne, O. C. & Wessely, S. (2003). Randomized controlled trial of brief cognitive behaviour therapy versus treatment as usual in recurrent deliberate self-harm: the POPMACT study. *Psychological Medicine* **33**, 969–976.