Psychiatria Danubina, 2011; Vol. 23, Suppl. 1, pp 198–202 *Conference paper*

CLASSIFICATION OF RISK IN PSYCHIATRY

Tim M. Gale[1,2], Chris J. Hawley[1,2] & Thanusha Sivakumaran[1]

[1]Department of Psychiatry, QEII Hospital, Welwyn Garden City, UK
[2]Department of Psychology, University of Hertfordshire, Hatfield, UK

SUMMARY

Psychiatric risk-assessments generally quantify risk using broad, categorical, indicators (e.g., high-risk, low-risk). We examined reliability of such indicators when applied by mental-health professionals. Four versions of a questionnaire were used, each specifying a different clinical outcome along with a range of different probabilities at which that outcome might occur. Respondents classified each probability, allowing a comparison of the level of likelihood at which different professionals would apply the terms 'high-risk', 'medium-risk' and 'low-risk'. We found little consistency among professionals who assessed risk for the same outcomes. Moreover, there were also large and unpredicted differences in response-profiles between the 4 clinical outcomes. These findings raise concerns about the communication value of current risk-assessment terminology.

Key words: *risk-assessment - quantification of risk*

* * * * *

INTRODUCTION

Risk-assessment is a core feature of psychiatry, typically involving completion of proformas to record relevant demographic and biographic information about patients. It is generally assumed this focuses attention on salient issues of risk or, at least, minimises neglect of these issues.

A key component of risk-assessment is quantification of risk. In some areas of medicine risk can be quantified precisely: clinical trials are designed and phased to produce reliable estimates of side-effect risks for medications. Similarly, epidemiological studies support reasonably precise quantification of risk for given illnesses in a population, based upon prevalence within a representative sample.

By contrast, psychiatric risk-assessments usually quantify risk using broad, categorical, terminology. In a recent review of risk-assessment proformas used by mental-health trusts (Hawley 2006), the majority coded risk using nominal (present/absent) or ordinal categories (high/medium/low-risk). Indeed this was the most consistent property to emerge from an otherwise diverse set of assessment tools. Why should this be the case? We might argue that probabilistic estimates of rare and tragic outcomes tell us only what proportion of a population will be affected rather than who the unfortunate individuals will be: attempting to give precise risk estimates for individuals is, therefore, misguided. But one might then ask what level of information is conveyed by generalised risk terminology and whether different professionals share a similar understanding of it? This study explores how professionals in psychiatry use risk terminology when making judgements about patients.

METHOD

We used an independent-groups design with 4 conditions: each used a different questionnaire in which respondents were invited to apply the terms 'high-risk', 'medium-risk' or 'low-risk' to 12 different probabilities, each specifying the risk of an unpleasant outcome for a fictitious patient. Each questionnaire/condition focused on a different unpleasant outcome but, in other respects, they were identical. The aims were: (i) to explore use of risk terminology by professionals assessing the same outcomes (within-conditions comparison); (ii) to compare professionals assessing different outcomes (between-conditions comparison). The dependent-variable was the probability value used to indicate the boundaries between: (i) high/medium-risk; and (ii) medium/low-risk.

Materials

Questionnaires were worded as follows: "Imagine that a highly sophisticated computer-programme has been developed that can predict the risk of a psychiatric patient committing a fatal assault on another person very accurately from information provided by the patient's clinician. This programme has been applied to 12 patients to calculate the risk of such an assault within the next year. According to your own judgement, indicate for each patient which term, high, medium or low, you would apply to the risk. Please tick one box for each patient." The italicised words varied across the 4 conditions (Table 1) describing outcomes with decreasingly unpleasant consequences.

Probability values were then presented for each the 12 fictitious patients and these ranged from the smallest (0.0003%) to the largest (20%). The respondent

indicated, for each value, whether they considered it indicative of high, medium or low-risk in that particular context. To counteract response-bias, presentation order of risk-terms was randomly determined for each of the 12 values, as was presentation order of the values. The questionnaire was extensively pilot tested. Table 1 illustrates the response structure.

Table 1. The 4 outcomes: each was phrased to allow transposition into the questionnaire without altering other wording

Condition	Outcome	Order of Unpleasantness
A	committing a fatal assault on another person	1
B	committing medically serious, but non-fatal, deliberate self-harm	2
C	losing his/her job	3
D	suffering transient nausea from medication	4

Table 2. Example of a completed questionnaire

Patient	Predicted Risk of Outcome (%)	Rank Score	Please indicate the category that applies		
#1	0.9% (9 in 1000)	6	Low	High	**X** Medium
#2	10% (1 in 10)	2	**X** High	Medium	Low
#3	0.5% (1 in 200)	7	High	Low	**X** Medium
#4	4% (1 in 25)	3	Medium	**X** High	Low
#5	0.005% (1 in 20,000)	11	High	Medium	**X** Low
#6	2% (1 in 50)	4	High	Low	**X** Medium
#7	0.1% (1 in 1000)	8	Medium	**X** Low	High
#8	20% (1 in 5)	1	Low	Medium	**X** High
#9	0.0003% (3 in 1 million)	12	High	**X** Low	Medium
#10	1% (1 in 100)		High	**X** Medium	Low
#11	0.08% (8 in 10,000)	9	**X** Low	High	Medium
#12	0.01% (1 in 10,000)	10	Medium	**X** Low	High

This replicates the layout used, except for the 'Rank Score' column, which did not appear. This column simply indicates the numerical order of the 12 probabilities ranging from highest to lowest. For this particular respondent, the boundary for high/medium is 3 and the boundary for medium/low is 8. In cases where the largest probability was not classified as 'high-risk', the high/medium boundary was scored as zero. Similarly, in cases where the smallest probability was not classified as 'low-risk', the medium/low boundary was scored as 13. Patients #10 and #1 have very similar values and were included to examine response consistency.

Participants

252 professionals (87 males, 165 females; aged 22-62 years) were allocated to the 4 conditions (63/condition). Psychiatrists, nurses, social-workers, and occupational-therapists were included. Participants did not differ between conditions on the following variables: age (ANOVA $F_{[3,247]}<1$, p=0.82); sex M:F (Chi-square$_{[3]}$=0.49, p=0.92); professional-group (Chi-square$_{[15]}$=2, p=1); level-of-education (Chi-square$_{[3]}$=0.74, p=0.99). These variables are known to predict statistical knowledge in MHPs (Gale 2003) so it was important to match them across conditions. All participants were recruited opportunistically and voluntarily at training events, being included only if they undertook risk-assessments regularly. Ethical approval was given by the Hertfordshire REC.

Procedure

Participants were told this was a questionnaire to assess opinion about risk. They were discouraged from spending longer than 15-minutes completing it, or conferring with colleagues. Responders provided background details and an anonymity-code, which was later used to exclude duplicate returns from the same individuals. Each participant noted the time taken to complete the questionnaire.

Data coding, analyses and predictions

Although probability values are continuously distributed, the small number of discrete values used here (i.e., 12) cannot be assumed to approximate continuous data. Therefore we ranked probabilities from highest to lowest (e.g., 20%=R1, 10%=R2, 4%=R3, etc.) and used non-parametric techniques.

The dependent-measures in this study were the boundaries between (i) high/medium-risk and, (ii) medium/low-risk, as conceived by respondents. These were calculated, respectively, as: (i) the lowest rank at which 'high-risk' was used; (ii) the highest rank at which 'low-risk' was used. Table 1 includes an illustrative example.

This method assumes that respondents apply terminology consistently: e.g., if 2% was classified as 'medium-risk', it would violate orderliness if a smaller

probability, say 1%, was classified as 'high-risk'. We therefore screened all completed questionnaires, excluding cases where such violations were present. Questionnaires were classified as: 'sequentially-consistent' if the use of risk-terms did not violate the probability rank order; and 'sequentially-inconsistent' if the probability rank order was violated, as in the aforementioned example.

Boundaries for risk-terms were compared between conditions using Kruskall-Wallace tests. We predicted that the boundaries chosen would be affected by the unpleasantness of the outcome. More formally, we predicted that 'high-risk' would be applied at a lower probability for more unpleasant outcomes, and at a higher probability for less unpleasant outcomes; and 'low-risk' would be applied at a lower probability for more unpleasant outcomes and at a higher probability for less unpleasant outcomes. In short, the risk-term chosen would be contingent on the outcome itself rather than the probability of that outcome.

RESULTS

The number of sequentially-consistent responders (i.e., those whose risk terminology did not violate the probability rank order) was 130/252 (52%). In 69 (27%) questionnaires there was a single violation and in 53 (21%) there were ≥2 violations. Completion time was not associated with sequential-consistency (means for sequentially-consistent and inconsistent responders were 4.67 and 4.75 minutes respectively, ANOVA $F_{[1,210]}<1$, NS). Educational-level predicted sequential-consistency: of those with a degree or higher-qualification, 59% used risk-terms consistently while, for those educated to 18-or-less, the figure was 35% (Chi-square$_{[1]}$=12.7, p<0.0005). Professional-group also predicted sequential-consistency (Chi-square$_{[5]}$=24.36, p<0.0002: psychiatrists 78%; social-workers 61%; CPNs 50%; inpatient-nurses 36%; occupational-therapists 35%).

All reported analyses are based on sequentially-consistent responders only: we can be assured this group fully understood the task whereas, for the sequentially-inconsistent responders, conceptual difficulties may have undermined questionnaire completion. We re-ran matching comparisons between conditions, confirming that they did not differ on any relevant variables after excluding sequentially-inconsistent responders. Table 3 displays mean and median ranks at which the high/medium and medium/low boundaries were placed.

Table 3. Mean and median ranks for high/medium and medium/low boundaries

Condition	Rank score of High/Medium boundary		Rank score of Medium/Low boundary	
	Mean (±SD)	Median (range)	Mean (±SD)	Median (range)
A	5.61 (±3.04)	6 (1-11)	9.39 (±2.57)	9 (5-13)
B	3.11 (±1.75)	3 (0-7)	7.05 (±2.51)	8 (2-10)
C	3.54 (±2.06)	3 (0-8)	7.62 (±1.86)	8 (4-13)
D	3.33 (±1.80)	4 (0-7)	7.24 (±2.45)	8 (3-11)

Comparing within condition

In table 3, the smallest range, i.e. the best agreement between responders, was 8 intervals while the largest, i.e., worst agreement, was 11 intervals (In cases where the highest probability was not classified as 'high-risk', the high/medium boundary was scored as zero. Similarly, in cases where the lowest probability was not classified as 'low-risk', the medium/low boundary was scored as 13). Working with one of the better cases, e.g., the medium/low boundary for B, where the range was 9 intervals, the two probability-values indicated by this range differ by a factor of 1000, so one professional's cut-off value for 'low-risk' was 1000-times smaller than another's.

Figure 1 indicates the percentage of professionals under each condition applying each of the 3 risk categories at each rank (see Table 3 for the value corresponding with each rank). As would be expected, agreement was good at extremes of the scale, most likely due to an anchoring effect, i.e., the tendency to use 'high-risk' for the largest probability and 'low-risk' for the smallest. The points of interest, however, are towards the centre of the graphs. Here there is consistently poor agreement: in some cases the votes are split almost equally between the 3 risk-terms.

Comparing between conditions

A significant difference emerged for placement of the high/medium boundary (Kruskall-Wallis H corrected for ties$_{[3]}$=11.9, p=0.008) and the medium-low boundary (Kruskall-Wallis H corrected for ties$_{[3]}$=13.9, p=0.003; table 3). Post-hoc comparisons revealed that condition-A differed from all others. So professionals who saw A, the most unpleasant scenario, applied the term 'high-risk' to smaller probabilities than professionals in any other condition. Similarly, the level of probability at which they applied the term 'low-risk' was also much smaller than for the other 3 conditions (N.B., a higher mean rank in table 3 is associated with smaller probabilities).

Response consistency

We compared responses for patients #10 and #1 (with respective probabilities of 1% and 0.9%). Since these values are almost identical, differing by just

1/1000, it is useful to examine the consistency of risk-terms applied to them. The prediction is that professionals should use the same risk-term for both probabilities but only 117 did so. However, sequentially-consistent responders applied the same term to these values in 76% cases compared with 26% of sequentially-inconsistent responders (Chi-square$_{[1]}$=60.7, p<0.0001). This may not be surprising but does demonstrate that sequentially-inconsistent responders probably lacked requisite statistical knowledge to complete the task properly; and further supports excluding this group.

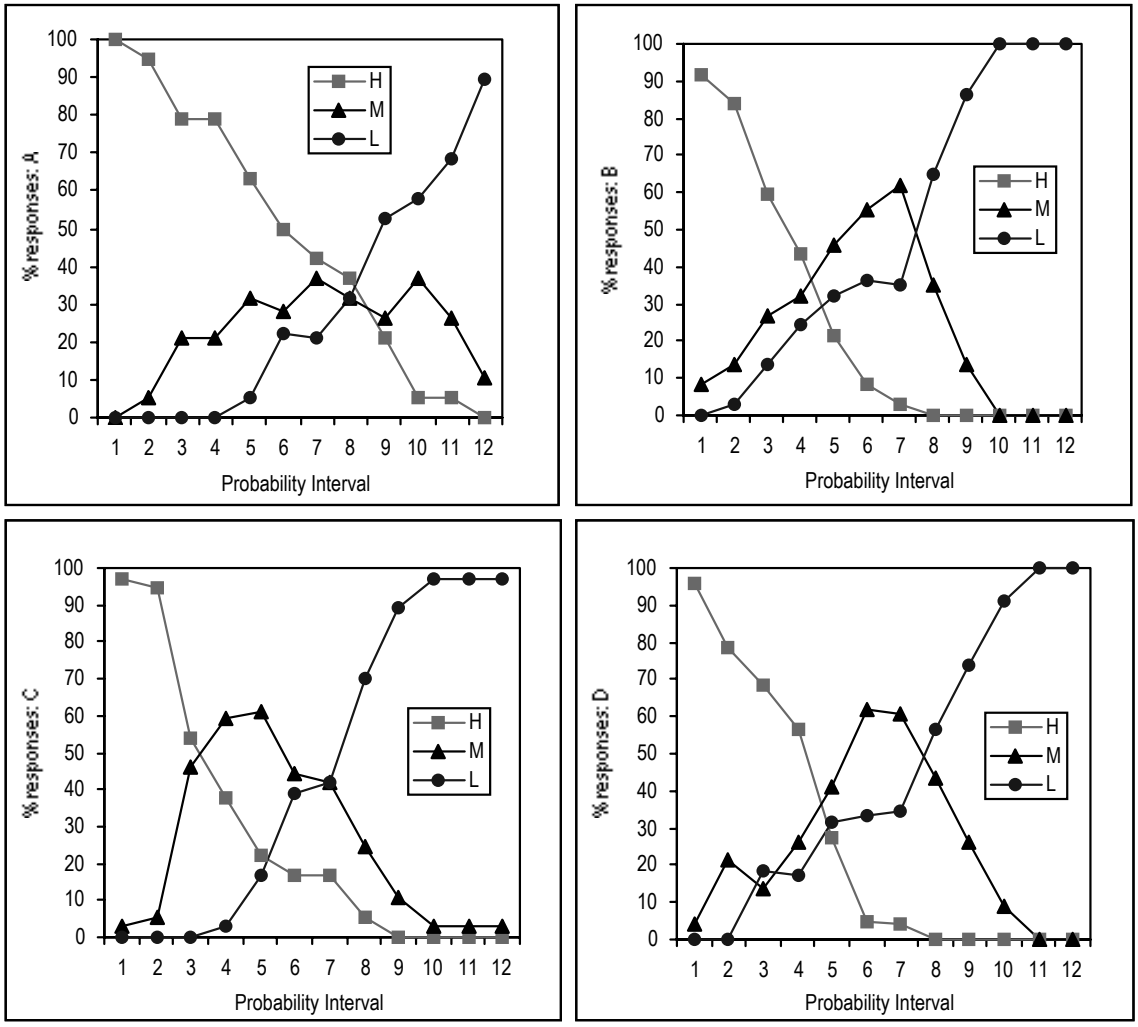

Figure 1. The percentage of professionals under each condition applying each of the 3 risk categories at each rank

DISCUSSION

We investigated use of verbal labels to denote levels of risk by asking professionals to quantify the terms 'high-risk', 'medium-risk' and 'low-risk'. A similar approach has been used previously, albeit in a different context (Biehl 2001). Before discussing the results, we first acknowledge the level of artifice in this task, and the contrived nature of the scenarios. In day-to-day risk-assessments, professionals would not have access to probabilities for patient outcomes. Nevertheless, they are expected to classify their patients according to risk so, on one-hand, the probabilities provided information over-and-above that usually available. On the other-hand, we provided no background information on patients, as might be available during a real risk-assessment. This was intentional since our aim was to keep the task as simple as possible, thereby excluding potentially distracting information. Nonetheless we acknowledge that this study was a laboratory-based, rather than field-based, investigation of risk-assessment.

Just over half the participants (52%) completed questionnaires logically and coherently, i.e., their use of risk-terms produced no anomalies. That so many participants failed to apply risk-terms consistently is unsurprising: mental-health professionals, like the wider

population, have conceptual difficulties with statistics Galle 2003) so the task presented here was undoubtedly challenging for individuals lacking familiarity with probability. We therefore included, in our analysis, only those respondents who used risk-terms consistently. Of those assigning risk-terms inconsistently, the majority made a single error so we can infer that most participants (79%) paid due-attention to the task, making a genuine attempt to complete it systematically and logically. But, by focusing on those who produced no inconsistent responses, we are assured that the main findings are not undermined by participant mis-understanding or lack of effort.

We look first at professionals who assessed identical patient outcomes, i.e., within-condition comparisons. There was marked variation in the probability threshold above which risk was denoted as high rather than medium, and similar variability for the transition from medium to low-risk. Some variability is expected: it would be remarkable if respondents demonstrated perfect agreement in positioning the boundaries associated with each term. What is notable, however, is the extent of variability: an extreme example was a 4000-fold difference for the probability at which 'high-risk' was applied to the same outcome. It is possible that professionals were constrained in their choice, having just 12 values to quantify the 3 risk-terms. However, the probabilities offered did span a broad range, so lack of choice could not fully account for the sheer range of estimates given.

Turning to quantification of risk-terms when applied to different outcomes, the pattern of results was not entirely as predicted. A biasing effect of outcome-severity was apparent for the most homicide scenario, but did not decrease linearly across the other conditions. So, on one-hand, there is evidence that the gravity of the outcome, and not merely its probability, affects the chosen risk-term. But on the other-hand, the results suggest this only happens when the outcome is particularly tragic. Either way, this finding raises concerns about reliability and validity of risk-assessment. If a latent expectation is that professionals should adjust terminology according to outcome severity, this would have to be evident across different outcomes for such an approach to be valid. It is likely that a death-related outcome elicits an emotional response, leading to greater caution being exerted. However, no such difference emerged when comparing medically serious self-harm with transient nausea. So, if risk-assessment is assumed to be contingent on both the likelihood and severity of an outcome, our results suggest that this assumption is unwarranted. It is possible that the non-fatal outcomes in our study failed to exceed respondents' thresholds of concern. However, real risk-assessment concerns not only rare, tragic outcomes, but also less severe, but still potentially damaging, outcomes: the latter are commonplace in the treatment of mental-illness.

In the UK, providers of mental-health services are required to produce their own risk-assessment tools. There is, as yet, no explicit guidance in the design and implementation of such tools, resulting in considerable variability between proformas currently in use (Hawley 2006, Higgins 2005). One commonality, however, is that level of risk is nearly always recorded by broad, categorical terms, rarely more expansive than the 3 terms investigated here. For this reason, we argue that our study has important implications for the design of risk-assessment tools and processes, and also for the expectations we should have of risk-assessment more generally. Accepting that our data was collected in an artificial context, rather than in clinical practice, the findings nonetheless point to the poor communication value of broad, categorical risk-terms in the context of patient assessments. Risk assessment is an increasingly central part of mental-healthcare, yet there is a surprising lack of good evidence to support current methods of assessment.

REFERENCES

1. *Biehl M, Halpern-Felsher BL Adolescents' and adults' understanding of probability expressions. (2001) Journal of Adolescent Health, 28, 30-35.*
2. *Gale TM, Hawley C, Sivakumaran, T. Do mental health professionals really understand probability? Implications for risk assessment and evidence-based practice. (2003) Journal of Mental Health, 12(4), 417-430.*
3. *Hawley CJ, Littlechild B, Sivakumaran T, Sender H, Gale TM, Wilson KJ Structure and content of risk assessment proformas in mental healthcare. (2006) Journal of Mental Health, 15(4), 437-448.*
4. *Higgins N, Watts D, Bindman J, Slade M, Thornicroft G. Assessing violence risk in general adult psychiatry. (2005) Psychiatric Bulletin, 29, 131-133.*

Acknowledgements

We thank: all participants for giving up their time voluntarily; Anne Farrow, Paul St. John Smith, Al Williamson, for assistance with data collection; Kerry Foley for help with participant matching.

Correspondence:
Dr Tim M. Gale
Department of Psychiatry, QEII Hospital
Howlands, Welwyn-Garden-City, Herts, AL7 4HL, UK
E-mail: t.gale@herts.ac.uk