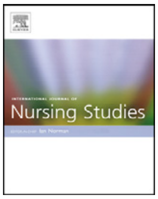
Contents lists available at SciVerse ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns



Effect of improving the realism of simulated clinical judgement tasks on nurses' overconfidence and underconfidence: Evidence from a comparative confidence calibration analysis

Huiqin Yang[a,*], Carl Thompson[b], Martin Bland[b]

[a] *Centre for Reviews and Dissemination, University of York, YO10 5DD, UK*
[b] *Department of Health Sciences, University of York, YO10 5DD, UK*

ARTICLE INFO

Article history:
Received 1 December 2011
Received in revised form 24 July 2012
Accepted 7 August 2012

Keywords:
Overconfidence
Underconfidence
Confidence calibration
Clinical simulation
Paper cases
Clinical judgement
Risk assessment

ABSTRACT

Background: Apparent overconfidence and underconfidence in clinicians making clinical judgements could be a feature of evaluative research designs that fail to accurately represent clinical environments.
Objectives: To test the effect of improved realism of clinical judgement tasks on confidence calibration performance of nurses and student nurses.
Design: A comparative confidence calibration analysis.
Settings: The study was conducted in a large university of Northern England.
Methods: Ninety-seven participants rated their confidence – using a scale that ranged from 0 (no confidence) to 100 (totally confident) on dichotomous clinical judgements of critical event risk. The judgements were in response to 25 paper-based and 25 higher fidelity scenarios using a computerised patient simulator and clinical equipment. Scenarios, and judgement criteria of 'correctness', were generated from real patient cases. Using a series of calibration measures (calibration, resolution and over/under-confidence), participants' confidence was calibrated against the proportion of correct judgements. The calibration measures generated by the paper-based and high fidelity clinical simulation conditions were compared.
Results: Participants made significantly less accurate clinical judgements of risk in the high fidelity clinical simulations compared to the paper simulations ($P = 0.0002$). They were significantly less confident in high fidelity clinical simulations than paper simulations ($P = 0.03$). However, there was no significant difference of over/under-confidence for participants between the two simulated settings ($P = 0.06$). Participants were no better calibrated in the high fidelity clinical simulations than paper simulations, $P = 0.85$. Likewise, participants had no better ability of discriminating correct judgements from incorrect judgements as measured by the resolution statistic in high fidelity clinical simulations than paper simulations, $P = 0.76$.
Conclusions: Improving the realism of simulated judgement tasks led to reduced confidence and judgement accuracy in participants but did not alter confidence calibration. These findings suggest that judgemental miscalibration of confidence in nurses may be a systematic cognitive bias and that simply making scenarios more realistic may not be a sufficient condition for correction.

[*] Corresponding author. Tel.: +44 01904 321064; fax: +44 01904 321041.
E-mail address: huiqin.yang@york.ac.uk (H. Yang).

0020-7489/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijnurstu.2012.08.005

What is already known about the topic?

- In clinical judgement, the ability of clinicians to match confidence to their judgemental abilities is a crucial link in the delivery of quality healthcare. Nurses' overconfidence and underconfidence in their clinical judgements have been demonstrated in some empirical studies.
- Almost all the clinical confidence calibration studies noted thus far used paper cases to represent clinical judgement tasks. However, their use as the cornerstone of much work in examining clinical judgement can be questioned, due to a lack of case fidelity (the degree of similarity between real situations and simulated situations).
- Therefore, apparent overconfidence and underconfidence in nurses making clinical judgements could be a feature of evaluative research designs that fail to accurately represent clinical environments.

What this paper adds

- This study investigates whether improving the realism of clinical scenarios impacts on nurses' over/underconfidence in their risk assessment judgements, and tests the hypothesis that clinicians' confidence miscalibration could be attributable to an experimental design that does not fully capture the 'realism' of real environment.
- The findings showed that improving the realism of simulated judgement tasks led to reduced confidence and judgement accuracy in participants but did not alter confidence calibration.
- These findings suggest that nurses' judgemental miscalibration of confidence may be a systematic cognitive bias and that simply making scenarios more realistic may not be a sufficient condition for correction.

1. Introduction

In clinical judgement, the ability of clinicians to match confidence to their judgemental abilities is a crucial link in the delivery of quality healthcare; it is often lacking (Baumann et al., 1991). The relationship between confidence and judgement success is known as calibration of confidence. In most calibration studies, confidence is often measured using a continuous scale of between 0 and 100 assigned by the judge on a particular judgement. Overconfidence is characterised by a positive score when the judge's mean confidence exceeds the mean judgement accuracy, whilst underconfidence is charactertised by a negative score when the judge's mean judgement accuracy exceeds the mean confidence. Miscalibration such as overconfidence or underconfidence when making judgements and decisions is an important form of bias in reasoning (Lichtenstein and Fischhoff, 1980, 1982; Petrusic and Baranski, 1997). Overconfidence in one's clinical judgement performance can increase the risk of iatrogenic harm (Croskerry and Norman, 2008). In the medical area of critical event risk assessment, overconfidence can result in delayed action (or worse, doing nothing) in the face of clinical data that merit an immediate response of intervention.

Much of previous medical research has demonstrated doctors' overconfidence phenomenon in their clinical judgements. For example, the study by Hausman et al. (1990) evaluated the over/underconfidence of a group of paediatric residents on their results of multiple choice examinations. This study showed a significant association between an increased level of overconfidence and lower examination scores, with a tendency for residents to be overconfident in the "correctness" of their judgements. Similar findings were revealed from a further study by Friedman et al. (2005), which assessed physicians' over/underconfidence in their diagnostic correctness and showed the tendency towards overconfidence in physicians' diagnostic judgements of 36–41% cases. Evidence has shown physicians' overconfidence as a contributing cause of diagnostic errors (Berner and Graber, 2008).

A number of nursing studies have equally paid attention to nurses' confidence levels in their clinical judgements (Baumann et al., 1991; McMurray, 1992; Thompson, 2003). A recent study (Yang and Thompson, 2010) using paper-based scenarios showed that nurses' confidence was systematically miscalibrated in their critical event risk assessment judgements. However, in using paper-based scenarios Yang and Thompson (2010) left themselves open to the quite reasonable criticism that the miscalibration may be due to a lack of realism in the judgement task used to measure performance.

The hypothesis that confidence miscalibration in decision makers can be attributed to experimental designs that fail to capture the 'realism' of real environments has been raised by other earlier researchers (Bjorkman et al., 1995; Gigerenzer et al., 1991; Juslin, 1993, 1994). Conversely, other studies suggest that decision makers' confidence levels for their judgements are more appropriate when judgement tasks adequately reflect natural judgement environments (Bjorkman et al., 1995; Juslin, 1993, 1994). However, to our knowledge the hypothesis that confidence miscalibration is associated with task realism (which was proposed about two decades ago) has not been quantitatively investigated in clinical calibration studies.

Almost all the clinical confidence calibration studies noted thus far used written case simulations to represent clinical judgement tasks. However, their use as the cornerstone of much work in examining clinical judgement can be questioned. Wigton et al. (1986) state that clinicians may not afford the same quality of attention to a written case simulation as they would to a real clinical situation. In contrast, computerised human patient simulators offer opportunities for capturing reality and enhancing the fidelity of clinical scenarios (Bond et al., 2001; Devitt et al., 2001; Waytt et al., 2007). In the context of clinical simulation, fidelity refers to the degree of similarity between real situation and simulated situation. In fact, fidelity is a 'proxy' of the notion of realism of simulated situations to real situations. Simulations using computerised human patient simulators substantially increase case fidelity, thereby improving the realism of clinical scenarios. With this in mind, this study investigates whether improving the realism of clinical scenarios impacts on nurses' over/underconfidence in their risk assessment judgements, and tests the hypothesis that clinicians' confidence miscalibration could be attributable to an experimental design that does not fully capture the 'realism' of real environment.

In this study we aimed to examine the effect of improved realism in clinical judgement tasks on nurses' confidence calibration in the context of critical event risk assessment judgements. Critical event risk judgements are an important clinical judgement and implicated in key indicators of healthcare quality such as failure to rescue (Goldhill and McNarry, 2004; McGloin et al., 1999).

2. Methods

2.1. Study design

This study was a prospective confidence calibration (Lichtenstein and Fischhoff, 1977) in two phases. In phase one, paper-based simulations were used to examine participants' confidence calibration. In phase two, high fidelity simulations with a computerised patient simulator (Laerdal SimMan^{TM}) were used to elicit participants' judgements and their confidence ratings on risk assessments.

2.2. Participants

To maximise the sample size of the paired comparisons between paper-based and high fidelity simulations we sampled both experienced nurses and student nurses, with student nurses being over sampled given the lower costs associated with their recruitment. We sampled 34 experienced registered nurses from the population of ward and critical care nurses in North Yorkshire, United Kingdom (UK) and 63 nursing students from the undergraduate student population in a large university in Northern England.

2.3. Constructing scenarios

Clinical scenarios were randomly derived from a real patient data set by Subbe et al. (2001). A booklet of clinical vignettes (paper-based simulations) was constructed to present 25 clinical scenarios for a fictitious emergency patient (see Appendix A). Each scenario included five cues with variable values: systolic blood pressure, respiratory rate, heart rate, temperature and consciousness levels. Cues were presented in natural units used in clinical practice (e.g. respiratory rate in breaths per minute). Patient consciousness was represented as four levels: alert (A), reacting to voice (V), reacting to pain (P) and unresponsive (U). A critical care specialist nurse with more than 10 years' specialist experience reviewed and approved the clinical vignette scenarios.

Whether a judgement was 'correct' or not (i.e. the judgement criterion) was also derived from the dataset prospectively collected by Subbe et al. (2001). Primary endpoints were death, intensive care unit (ICU) and high dependency unit (HDU) admission or cardiopulmonary resuscitation. If the patient had any one of these endpoints then they were classed as having experienced a critical event. Likewise, if one of these endpoints was not reached then the patient was classed as not 'at risk' of acute deterioration.

In phase two a computerised patient simulator (Laerdal SimMan^{TM}) was used to present higher fidelity simulations of the same cases using a clinical simulation lab. High fidelity clinical scenarios simulated the same information and units of the following four cues as in paper simulations: systolic blood pressure, heart rate, respiratory rate and temperature. These four cues were simulated on the bedside monitor in the simulation lab. In order to operationalise "level of consciousness", we converted different levels into different vocal sounds into the SimMan system. For example, the consciousness level of alert was represented by a vocal sound: "*Oh, I feel really ill.*" The consciousness level of reacting to voice was represented by a vocal sound: "*Ouch! Where is my wife? Get off me! Get off Me!*" In this high fidelity setting, participants were asked to observe clinical information cues on the monitor and listen the simulated vocal sound, enabling them to make a judgement for each scenario. All participants were not required to measure clinical information cues such as blood pressure in the high fidelity setting.

2.4. Data collection

ll the participants made their judgements in response to paper-based scenarios first and then underwent the simulated scenarios in high fidelity settings, In both settings participants were asked to make and record dichotomous judgements of either "yes" (at risk of a critical event) or "no" (not at risk of a critical event) on each scenario in a data collection sheet. Simultaneously, they were also asked how confident they were in the judgement and to record this as a confidence rating. Each participant's confidence was immediately assigned once a particular judgement was made for each scenario, so the confidences ratings well reflected their judgements at the time when the judgement was being made. Confidence ratings were assigned on a scale of 0–100, anchored at the ends by 0 (completely unconfident) and 100 (totally confident). Participants were given the same amount of time for both paper-based scenario and high fidelity scenario. To minimise any learning and maturation effect (Cook and Campbell, 1979) during the time gap between phase one and phase two of the study, the experiment of high fidelity simulations was conducted shortly (5–7 days) after participants completed the phase one of paper simulations.

2.5. Confidence calibration as methodology

2.5.1. Confidence calibration statistics

Three calibration statistics characterise the relationship between confidence and judgement correctness: (i) calibration score, (ii) over/underconfidence measure, and (iii) resolution score. The calibration score is the sum of squared deviations away from the 45° line in a scatter plot, weighted by the number of responses in each confidence category and divided by the total number of responses (Soll, 1996). It is represented by the following formula (Petrusic and Baranski, 1997):

$$\frac{1}{n}\sum_{j=1}^{J} n_j(\bar{p}_j - \bar{e}_j)^2$$

where n is the total number of responses; J is the total number of confidence categories; n_j is the number of responses in confidence category j; $\bar{p}_j$ is the mean confidence level associated with category j; $\bar{e}_j$ is the mean proportion correct associated with category j.

The calibration score provides a weighted squared deviation between the mean proportion correct and the mean confidence rating associated with each confidence category (Lichtenstein and Fischhoff, 1977, 1982). The calibration score ranges between 0 (perfectly calibrated) and 1 (worst calibrated); the lower the score the better the calibration. The worst possible score, 1.0, can be obtained by a judge who always assigns a confidence rating ($p = 100$) when the judgement was wrong and a confidence rating ($p = 0$) when the judgement was actually correct. In contrast, a calibration score of zero represents the perfect calibration: the proportion correct tallies perfectly with the confidence probability level of participants.

Over/underconfidence measures the deviation between confidence and proportion correct: $\bar{p} - \bar{e}$, where $\bar{p}$ denotes the mean confidence and $\bar{e}$ denotes the mean proportion correct. A negative score (the mean proportion correct exceeds the mean confidence) is indicative of underconfidence. In contrast, a positive score (the mean confidence exceeds the mean proportion correct) indicates overconfidence.

The final measure of the confidence/accuracy relationship is resolution. The resolution score measures how well participants use their confidence ratings to differentiate correct from incorrect responses of judgements (Petrusic and Baranski, 1997). The resolution score provides a weighted squared deviation between the mean proportion correct ($\bar{e}_j$) for each confidence category and the overall proportion correct ($\bar{e}$) for whole group level (Petrusic and Baranski, 1997):

$$\frac{1}{n}\sum_{j=1}^{J} n_j(\bar{e}_j - \bar{e})^2$$

The resolution score ranges from zero to the knowledge index $\bar{e}(1 - \bar{e})$, thus the resolution score is dependent on the mean proportion correct. Therefore this resolution score does not lead to a meaningful comparison of the discrimination skills between two judges. A normalised resolution score (NRS) overcomes this issue by adjusting for the knowledge index (Yaniv et al., 1991):

$$\mathrm{NRS} = \frac{(1/n)\sum_{j=1}^{J} n_j(\bar{e}_j - \bar{e})^2}{\bar{e}(1 - \bar{e})}$$

The normalised resolution score, which is not conditional on the mean proportion correct, offers a more robust statistic when comparing discrimination abilities. Normalised resolution scores vary between 0 (no resolution) and 1 (perfect discrimination). A higher score indicates the judge's greater ability to differentiate levels of achievement in judgement correctness. The resolution measures a participant's ability to separate out different levels of achievements in judgement correctness without necessarily knowing what the levels are, e.g. if a participant always expresses total confidence (a score of 100) on wrong judgements and has no (0) confidence on correct judgements, the calibration would be poor but the resolution could be high. We report only the normalised resolution scores in this study.

2.6. Hypothesis testing for calibration measures

Individual calibration statistics were calculated to enable hypothesis testing of differences in these measures between paper-based and high fidelity clinical simulation conditions. Where the assumptions necessary for parametric testing were met, paired sample t-tests were used to test the hypothesis that the mean difference in calibration indices between the two simulated settings was equal to zero. Where the data were not suitable for parametric testing, the Wilcoxon matched-pairs signed-ranks test was used to test the null hypothesis that the calibration statistics in one simulated setting did not tend to be greater or smaller than those in the other. All analyses were performed using Stata version 9 (http://www.stata.com/).

3. Results

3.1. The participants

Ninety-seven (34 experienced nurses and 63 student nurses) participated in both paper-based and high fidelity arms of the experiment. The mean age of experienced nurses was 36.6 years (standard deviation (SD) 10.0) and the mean age of students was 27.8 years (SD 8.22). Experienced nurses had an average of just over 12 years of clinical experience; 81% of experienced nurses were educated to Diploma or Degree level. The overwhelming majority of students were registered for nursing degrees: 93.7% of the students were 2nd and 3rd year Degree nurses, with 6.3% of the students being 1st year post graduate diploma students.

3.2. Proportion of correct judgements and confidence ratings

Participants were significantly less accurate in the higher fidelity (mean 73.65%; SD 7.68%) than in the paper simulations (mean 77.11%; SD 7.04%), $t(96) = 3.93$, $P = 0.0002$. Participants were significantly less confident in the higher fidelity simulations (mean 75.26; SD 11.18) than in the paper simulations (mean 76.63; SD 11.26), $t(96) = 2.19$, $P = 0.03$.

3.3. Over/underconfidence

No significant difference was observed in over/underconfidence for participants between paper based (mean −0.48, SD 14.16) and higher fidelity computerised patient clinical simulations (mean 1.61, SD 14.63), $t(96) = -1.91$, $P = 0.06$.

3.4. Calibration and resolution

Participants were not significantly better calibrated in the higher fidelity clinical simulations (Mdn 0.048) than in the paper simulations (Mdn 0.050), $z = 0.19$, $P = 0.85$. Equally, participants did not exhibit significantly better resolution in the higher fidelity clinical simulations (Mdn 0.196) than in the paper simulations (Mdn 0.213), $z = -0.31$, $P = 0.76$. Subgroup analysis showed that students had no better calibration in high fidelity simulations (Mdn 0.048) than paper simulations (Mdn 0.053), $z = 0.87$, $P = 0.38$. Experienced nurses were also not more calibrated in high fidelity simulations (Mdn 0.048) than paper simulations (Mdn 0.046), $z = -0.88$, $P = 0.38$. In term of resolution, subgroup analysis showed that students had no significantly better normalised resolution in high fidelity simulations (Mdn 0.198) than paper simulations (Mdn 0.227), $z = 0.29$, $P = 0.77$. Likewise, experienced nurses had no significantly better normalised resolution in high fidelity simulations (Mdn 0.192) than paper simulations (Mdn 0.199), $z = 0.35$, $P = 0.73$. The results of subgroup analyses were consistent with the overall results of this study.

4. Discussion

The main findings from this study were that improving the ‘realism’ of judgement tasks significantly decreased participants’ confidence and judgement accuracy but did not alter their confidence calibration or resolution. This study showed that, as the task realism was enhanced via clinical simulations, a significant decrease in participants’ judgement accuracy was accompanied by a parallel reduction in their confidence levels in the high fidelity simulation conditions. This pattern was well captured by the measures of calibration and resolution, as no significant difference in calibration or resolution was observed between the two conditions. The participants were no better calibrated nor able to demonstrate better resolution as a result of having their performance measured using higher fidelity simulations, compared with the performance measured using paper simulations. Because we knew the task structure, number of cues and relative uncertainty of the clinical task environment (derived from the Subbe dataset (2001)), we were able to present information cues in ways that mirrored the clinical “ecology”. By using the computerised patient simulator (SimMan^{TM}), monitoring equipment and vocal information we were able to increase the realism of the scenarios which allowed participants to make judgements in settings more similar to their real environments. Of course, this “realism” was still some ways short of actual clinical environments; with their attendant “noise”, distractions and multiple task presentation.

Improving realism of judgement tasks by using technology to simulate visual and perceptual cues, did not influence calibration and resolution amongst participants in this study. The information cues presented in the high fidelity setting were identical to those presented in the paper simulations. In the high fidelity setting, participants were asked to make a judgement and assign their confidence level for each scenario by observing clinical information cues on the bedside monitor and listening to a vocal sound being simulated. It should be noted that information acquisition on visual and other sensory information is a crucial component of the clinical judgement process amongst nurses. In our study, simulations using a computerised patient simulator (Laerdal SimMan^{TM}) enabled us to recreate actual judgement situations in which such judgements were usually made on the basis of information acquisition on visual and other perceptual information cues, thereby considerably enhancing the realism of judgement tasks. Despite this improved realism, there was no better calibration of participants in the high fidelity simulation setting when compared with paper-based scenarios. Whilst the two phases differed in terms of the available information presentation formats on which to base judgement there appears to be a common processing mechanism behind the evaluation of probabilistic cues as a means of obtaining a level of feeling of relative certainty in participants’ judgement.

This explanation, however, refutes the (previously highlighted) proposition that confidence miscalibration can be attributed to ecological non-realism in experimental design (Bjorkman et al., 1995; Gigerenzer et al., 1991; Juslin, 1993, 1994). In contrast, Soll (1996) predicts that non-realism in judgement tasks is a sufficient rather than necessary condition for overconfidence to occur. Soll (1996) concludes that true systematic biases towards over/underconfidence exist amongst individual participants and implies that these are somewhat resistant to changes in evaluative design. Baranski and Petrusic (1994) and Petrusic and Baranski (1997) also find that confidence calibration (e.g. overconfidence) is remarkably similar in judgements built around perceptual and non-perceptual information. On balance, our findings suggest that improving the realism of clinical judgement tasks alone does not change systematic cognitive biases such as over/underconfidence on nurses’ judgements.

Despite the fact that improving realism of judgement tasks significantly decreased participants’ confidence and judgement accuracy, similar patterns of confidence miscalibration were observed in both paper based and high fidelity clinical simulations. In general participants had a good calibration of confidence on their judgements in the two settings (Mdn 0.050 in paper simulations and Mdn 0.048 in high fidelity simulations). The similar pattern between paper-based and high-fidelity clinical simulations indicates that the relationship between participants’ confidence and accuracy has not been significantly influenced by improving the realism of judgement tasks. This ties in with Keren’s (1991) assertion that calibration is part of individual cognition rather than determined by judgement events/tasks. The finding suggests that miscalibration patterns seen in these participants may be transferred to real clinical situations. Further research is required on the evaluation of confidence calibration patterns of clinical judgements made in real clinical settings.

4.1. Strengths and limitations

The strengths of the study included the use of real patient data in constructing clinical scenarios and the use of the same patient case records to derive valid judgement criteria of 'correctness'. As opposed to conventional paper case simulations, we used a novel simulation technique to recreate high fidelity clinical scenarios to studying participants' risk assessment judgements, with a presentation of vital clinical information.

In real situations, however, there is a large amount of clinical information available to nurses for making risk assessments, including other perceptual clinical information. In reality, nurses may also use the additional information in assessing whether the patient is at risk of deterioration. Furthermore, in order to study the effect of improved realism on participants' judgements rather than their ability to correctly measure cues, participants were not required to perform active information seeking such as measuring blood pressure in the high fidelity simulation setting. It should be noted, however, in real practice situations nurses often actively seek out a variety of a patient's clinical information to enable them to make a clinical judgement. In addition, there were some limitations of the simulation for consciousness level. We used simulated vocal sounds to represent various levels of consciousness. In practice, nurse may assess the level of a patient's consciousness using methods such as talking to the patient in order to elicit a response.

All the participants made their judgements in response to paper-based scenarios first and then underwent the realistic scenarios in clinical simulations, which may introduce the familiarity effect of the first task on the second task (Hamm, 2008). Although randomisation of the task order could have minimised such familiarity effect, we believe that this effect was minimal in our results as we did not give any feedback of correctness to participants when they completed the first task and, notably, participants performed less well in judgement accuracy in the high fidelity simulation setting. Additionally, in this study we did not test the reliability of calibration score, which may be further investigated in future clinical calibration studies.

4.2. Implication for research

The findings from our study indicate that the prevalence of confidence miscalibration is independent of the improved realism of judgement tasks. Miscalibration does not disappear in simulations that are built around more "natural" judgement tasks in higher fidelity settings. It suggests that nurses' confidence miscalibration forms part of their cognitive make up. Thus, identifying effective interventions to minimise nurses' cognitive biases may be a fruitful avenue for future research. Such approaches can be effective; cognitive feedback built around calibration scores and visualisation of judgement trends have yielded positive results (Lichtenstein and Fischhoff, 1980). The effectiveness of providing cognitive feedback as a means of improving physicians' calibration performance has been demonstrated in medical research (Poses et al., 1992). Further studies are required to ascertain whether providing cognitive feedback could improve nurses' confidence calibration.

5. Conclusion

This study suggests that better calibration is not necessarily correlated with improved realism in judgement tasks. Misplaced confidence (miscalibration), with a similar pattern, was seen in both low and higher fidelity clinical simulations. The notion that miscalibration exists independently of the tasks used to measure it in artificial and controlled settings is given extra weight by the finding that overconfidence has been demonstrated amongst physicians' processing probabilistic information when treating possible pneumonia patients in real clinical settings (Christensen-Szalanski and Bushyhead, 1981). We therefore conclude that nurses' confidence miscalibration is not simply due to non-realism of judgemental tasks, but appears to be an intrinsic part of nurses' judgement process. As such, the true challenge lies in finding ways of correcting this miscalibration and improving the accuracy and therefore quality of clinical judgements. Calibration methodology has something to offer the medical educator and evaluator who seek to rise to this challenge.

Conflict of interest: None declared.

Funding: The study was funded by the Department of Health Sciences of University of York, United Kingdom (UK).

Ethical approval: Ethical approval was obtained from the Research Governance and Ethics Committee of the University of York, UK. All participants were given the same information about the study and gave written consent to take part. All participants completed a written informed consent document.

Acknowledgments

We thank Dr C. P. Subbe for consenting to us using his dataset in constructing our clinical scenarios. We also thank Alison Foster for clinical advice on high fidelity clinical simulations.

Appendix A. Clinical information

Mr. Robert Wright, 63 years old and 76 kg weight, was presented to the emergency room in your hospital, accompanied by his wife. He was generally feeling unwell, with a tender abdomen and vomited after each meal for past 2 days. He was born in England and he has been married for 38 years. He is a senior engineer in an automotive company. He has no food or medical allergies. There was no report of use of medications. He has no significant past medical history or history of mental illness. The details of family history are unclear. The following sets of information are available to you when you assess Mr. Wright on admission. Please make your clinical judgements for each scenario.

An example paper-based patient scenario

Systolic blood pressure 114 mmHg	Risk (circle)	
Heart rate 118 beats per minute	YES	NO
Respiratory rate 40 breaths per minute		
Temperature 38.0 °C	Confidence (0–100)	
Conscious level reacting to pain		

An example response sheet for higher fidelity clinical simulations

	Risk (circle)	
Frame (Scenario) 6	YES	NO
	Confidence (0–100)	

References

Baranski, J.V., Petrusic, W.M., 1994. The calibration and resolution of confidence in perceptual judgments. Perception and Psychophysics 55, 412–428.

Baumann, A.O., Deber, R.B., Thompson, G.G., 1991. Overconfidence among physicians and nurses: the "micro-certainty, macro-uncertainty" phenomenon. Social Science & Medicine 32 (2), 167–174.

Berner, E.S., Graber, M.L., 2008. Overconfidence as a cause of diagnostic error in medicine. American Journal of Medicine 121 (5 Suppl.), S2–S23.

Bjorkman, M., Juslin, P., Winman, A., 1995. Realism of confidence in sensory discrimination: the underconfidence phenomenon. Perception and Psychophysics 54, 255–259.

Bond, W.F., Kostenbader, M., McCarthy, J.F., 2001. Prehospital and hospital-based health care providers' experience with a human patient simulator. Prehospital Emergency Care 5 (3), 284–287.

Christensen-Szalanski, J.J., Bushyhead, J.B., 1981. Physicians' use of probabilistic information in a real clinical setting. Journal of Experimental Psychology: Human Perception and Performance 7 (4), 928–935.

Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: Design and Analysis Issues for Field Settings. Houghton Miffin, Boston.

Croskerry, P., Norman, G., 2008. Overconfidence in clinical decision making. The American Journal of Medicine 121 (5A), S24–S29.

Devitt, J.H., Kurrek, M.M., Cohen, M.M., Cleave-Hogg, D., 2001. The validity of performance assessments using simulation. Anesthesiology 95 (1), 36–42.

Friedman, C.P., Gatti, G.G., Franz, T.M., Murphy, G.C., Wolf, F.M., Heckerling, P.S., Fine, P.L., Miller, T.M., Elstein, A.S., 2005. Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. Journal of General Internal Medicine 20, 334–339.

Gigerenzer, G., Hoffrage, U., Kelenbolting, H., 1991. Probabilistic mental models: a Brunswikian theory of confidence. Psychological Review 98 (4), 506–528.

Goldhill, D.R., McNarry, A.F., 2004. Physiological abnormalities in early warning scores are related to mortality in adult inpatients. British Journal of Anaesthesia 882–884.

Hamm, R., 2008. Cue by hypothesis interactions in descriptive modeling of unconscious use of multiple intuitive judgment strategies. In: Plessner, H., Betsch, C., Betsch, T. (Eds.), Intuition in Judgment and Decision Making. Lawrence Erlbaum, Mahwah, NJ, pp. 55–70.

Hausman, C.L., Weiss, J.C., Lawrence, J.S., Zeleznik, C., 1990. Confidence weighted answer technique in a group of pediatric residents. Medical Teacher 12 (2), 163–168.

Juslin, P., 1993. An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. European Journal of Cognitive Psychology 5, 55–71.

Juslin, P., 1994. The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. Organizational Behavior and Human Decision Processes 57, 226–246.

Keren, G., 1991. Calibration and probability judgments: conceptual and methodological issues. Acta Psychologica 77, 217–273.

Lichtenstein, S., Fischhoff, B., 1977. Do those who know more also know more about how much they know? Organizational Behavior and Human Decision Processes 20, 159–183.

Lichtenstein, S., Fischhoff, B., 1980. Training for calibration. Organizational Behavior and Human Performance 26, 149–171.

Lichtenstein, S., Fischhoff, B., 1982. Calibration of probabilities: the state of the art to 1980. In: Kahneman, D., Slovic, P., Tverksy, A. (Eds.), Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, pp. 306–334.

McGloin, H., Adam, S.K., Singer, M., 1999. Unexpected deaths and referrals to intensive care of patients on general wards. Are some cases potentially avoidable? Journal of the Royal College of Physicians of London 33 (3), 255–259.

McMurray, A., 1992. Expertise in community health nursing. Journal of Community Health Nursing 9 (2), 65–75.

Petrusic, W.M., Baranski, J.V., 1997. Context, feedback, and the calibration and resolution of confidence in perceptual judgments. The American Journal of Psychology 110 (4), 543–575.

Poses, R.M., Cebul, R.D., Wigton, R.S., Centor, R.M., Collins, M., Fleischli, G., 1992. Controlled trial using computerized feedback to improve physicians' diagnostic judgments. Academic Medicine 67 (5), 345–357.

Soll, J.B., 1996. Determinants of overconfidence and miscalibration: the roles of random error and ecological structure. Organizational Behavior and Human Decision Processes 65 (2), 117–137.

Subbe, C.P., Kruger, M., Rutherford, P., Gemmel, L., 2001. Validation of a modified Early Warning Score in medical admissions. Quarterly Journal of Medicine 94 (10), 521–526.

Thompson, C., 2003. Clinical experience as evidence in evidence-based practice. Journal of Advanced Nursing 43 (3), 230–237.

Waytt, A., Archer, F., Fallow, B., 2007. Use of simulators in teaching and learning: paramedics' evaluation of a patient simulator? Journal of Emergency Primary Health Care 5 (2), 1–16.

Wigton, R.S., Hoellerich, V.L., Patil, K.D., 1986. How physicians use clinical information in diagnosing pulmonary embolism: an application of conjoint analysis. Medical Decision Making 6 (1), 2–11.

Yang, H., Thompson, C., 2010. Nurses' risk assessment judgements: a confidence calibration study. Journal of Advanced Nursing 66 (12), 2751–2760.

Yaniv, I., Yates, J.F., Smith, J.E.K., 1991. Measures of discrimination skill in probabilistic judgement. Psychological Bulletin 110, 611–617.