Design and validation of a scoring tool for performance measurement in the management of anaphylaxis under general anaesthesia

Dr David Wright[1], Dr Rohit Garkoti[2], Dr Makani Purva[1]

1 Hull Institute of Learning and Simulation, Clinical Skills, Hull Royal Infirmary, Anlaby Road, Hull, HU3 2JZ Makani.Purva@hey.nhs.uk

2 Consultant Anaesthetics & Medical Simulation Lead, QE Gateshead Health NHS Foundation Trust

Abstract

Introduction

Competent management of anaphylaxis under general anaesthesia is a core anaesthetic skill and reliable assessment of performance is essential. We aimed to create a scoring tool which was reliable and valid.

Methods

The modified Delphi technique was used to produce a checklist of weighted tasks. 2 groups (8 junior and 8 senior anaesthetic trainees) undertook an anaphylaxis simulation and a panel of raters scored their performance using the tool.

Results

The Delphi process reached concordance after 2 rounds (Kendall W 0.75, p<0.001), producing a checklist of 22 weighted tasks. Internal consistency was excellent (Cronbach's alpha 0.91-0.96). Total inter-rater reliability was high (ICC=0.98, 95% CI 0.96-0.99) and the majority of tasks independently showed good reliability. Principal component analysis indicated that the tool could be used reliably with a single rater. Mean total scores for junior vs senior, showed no significant difference (p>0.05).

Conclusions

The Delphi technique was effective and efficient. The checklist was highly reliable, suggesting it could be used for both formative and summative assessment. An unexplained variance of <5% with a single rater, allows deployment with limited resources. High reliability may also represent comprehensive sampling of the construct, and supported by Cronbach's alpha values, indicates good content validity. Failure to demonstrate a performance difference, between juniors and seniors, limits interpretation of construct validity. This could be resolved in future by increasing the experience gap. Given the resource implications of validating assessment tools for myriad specific scenarios, we instead suggest the development of a validated and standardised Delphi toolkit.

Introduction

Anaphylaxis in the perioperative period has an incidence of 1 in 10 000 to 1 in 20 000 [1,2]. It is a severe life-threatening hypersensitivity reaction, with a mortality of up to 10%, where rapid recognition and treatment improves outcomes [3]. In the context of current safety culture and ongoing changes to postgraduate medical training [4], it is recognised that competence should first be demonstrated in a safe environment [5]. Simulation training is well suited to such infrequently encountered, high-risk scenarios. Reliable and valid assessment of performance is also required to ensure that this training is effective and that a standard has been met. The reproducibility of simulation is ideal for the development and testing of such assessments. We used the modified Delphi technique to create an objective scoring tool, for the management of anaphylaxis during general anaesthesia. We then assessed its reliability and validity using the known-groups technique.

Methods

This study was approved by the local research ethics committee of the National Research Ethics Service.

Delphi technique

The Delphi technique is a method of producing a consensus expert opinion. It was developed in post-second world war North America, to improve the accuracy and objectivity of military planning [6] and has since been adapted for the development of performance metrics[7]. It involves a series of rounds where the opinion of several experts is sought, collated and redistributed, by an independent facilitator. This is repeated until consensus is reached. This method addresses the limitations of less formal processes, specifically the bias of face-to-face discussion and lack of a defined outcome for confirming consensus.

After obtaining informed consent, we recruited 6 anaesthetists with 5 or more years

consultant experience, forming the expert panel. They remained blind to other members. All undertook a Delphi training package (tutorial and written material). In round 1, a list of 22 management tasks, derived from the Association of Anaesthetists of Great Britain and Ireland (AAGBI) anaphylaxis guideline [8], was provided to all experts. They were asked to weight the importance of each task on a 5 point Likert scale (1: not important, 5: extremely important) and were able to recommend changes to tasks, exclude tasks or add new ones. The facilitator (one of the study authors) calculated median and ranges for each task. These were distributed to the panel, along with inclusion/exclusion recommendations and raw scores. In round 2, the panel had the opportunity to modify their scores and change recommendations. This process was repeated until adequate agreement was reached. The final weighted task list, formed the assessment tool.

Simulation

8 junior (core training year one and two) and 8 senior (specialty training year five or above) anaesthesia trainees were recruited by invitation. Participants remained blind to other's identity.

Simulation was conducted in a simulated operating theatre at the Hull Institute of Learning and Simulation, using a SimMan 3G® (Laerdal, Norway). Performance was recorded using a smots™ audio-visual system (Scotia, UK). Faces were pixelated prior to video review, to assist blinding of assessors to candidate identity and group.

The scenario

Participants received a standardised orientation to SimMan 3G® and the simulation environment. All participants undertook the same scenario and received an introduction, as a written case handover:

"You are the anaesthetic trainee on call. This is a 21 year old male who I have just anaesthetised for a laparoscopic appendicectomy. He has no past medical or anaesthetic history, no allergies and no regular medications. I performed a rapid sequence induction with thiopentone, suxamethonium, fentanyl and atracurium. He is on the table and the surgeon is about to start."

Following antibiotic administration, a programmed set of physiological changes consistent with anaphylaxis (tachycardia, hypotension, desaturation and wheeze), were introduced at fixed time intervals. Rash was reported if the participant enquired. The participant was able to manage the patient as they deemed appropriate and physiological responses to key interventions were standardised within the programme. Participants were debriefed afterwards.

Reliability and Validity

5 assessors, with 5 or more years consultant experience were recruited (none were involved in the Delphi process). Each undertook training in the use of the assessment tool (presentation and written material) and were blind to others identity. Assessors used the tool to score the performance of all 16 candidates using video review.

Administration of adrenaline is recognised as essential in the management of life-threatening anaphylaxis [9] and was identified by the investigators as a single, high priority task. Time to administration was recorded by one of the investigators, remaining blind to candidate group.

Statistics

Statistical analysis was performed using SPSS® version 20 (IBM corp) and Prism® version 5 (Graphpad Software Inc). Agreement during the modified Delphi process was assessed after each round using Kendall's W, with a concordance coefficient of ≥ 0.75 required to accept the tool. Inter-rater reliability amongst assessors was measured using intra-class correlation coefficients (2-way mixed effects), defining excellent reliability as ICC≥ 0.75 [10]. Internal consistency of the tool was measured for each assessor, using Cronbach's alpha. Total score and time to adrenaline administration for each group, were assessed using an unpaired 2-tailed t test. Principal component analysis was used to determine the number of assessors required to explain more than 95% of variation from the mean score.

Results

Delphi technique

Adequate concordance was achieved after 2 rounds (Kendall W 0.75, $p<0.001$). No additional tasks were added and non were excluded. This produced a checklist of 22 weighted tasks (figure 1).

Reliability and validity

17 of the 22 tasks had an ICC ≥ 0.75, (figure 2). Total score ICC was 0.98 (95% CI 0.96-0.99). Internal consistency of the assessment tool was high for all assessors (Cronbach's alpha 0.91-0.96). There was no significant difference in total score awarded to juniors and seniors. Mean (SD) total score for juniors was 55.3 (12.4) and seniors

Figure 1: Final checklist

Candidates actions	Round 1 Median (Range)	Round 2 Median (Range)	Final weighted score
Uses the ABC approach	5.0 (4-5)	5.0 (4.5-5)	5.0
Removes all positive causative agents	5.0 (3-5)	5.0 (4-5)	5.0
Calls for help	4.5 (3-5)	4.5 (4-5)	4.5
Administers 100% oxygen	5.0 (4-5)	5.0 (4-5)	5.0
Elevates the patients legs if hypotension	3.0 (2-4)	3.0 (2-3)	3.0
Starts CPR if necessary	5.0 (5)	5.0 (5)	5.0
Delivers adrenaline in timely manner	5.0 (4-5)	5.0 (4-5)	5.0
Adrenaline given in appropriate dose	5.0 (4-5)	5.0 (4-5)	5.0
Adrenaline given by appropriate route (IV)	4.0 (4-5)	4.0 (4-5)	4.0
Considers starting an IV infusion of adrenaline if several doses needed	3.5 (3-5)	3.5 (3-4)	3.5
Considers high rate of fluid therapy	4.0 (2-5)	4.0 (2-5)	4.0
Plans transfer to appropriate area	4.0 (2-5)	4.0 (3-4)	4.0
Gives antihistamine	2.5 (1-4)	2.5 (1-3)	2.5
Gives Steroids	3.0 (2-4)	3.0 (2-4)	3.0
Considers alternative vasopressor if needed	3.5 (2-5)	3.5 (3-5)	3.5
Treats persistent bronchospasm with IV salbutamol	3.0 (2-5)	3.0 (2-5)	3.0
Takes blood sample for mast cell tryptase	4.0 (3-4)	4.0 (3.5-4)	4.0
Takes/arranges for a second blood sample to be taken 1-2 hrs later	4.0 (3-4)	4.0 (3.5-4)	4.0
Liaises with hospital lab for analysis of sample	3.0 (2-3)	3.0 (3)	3.0
Arranges appropriate investigation of the patient (immunology)	4.0 (3-4)	4.0 (3-4)	4.0
Notifies as appropriate (GP, AAGBI anaphylaxis database)	3.5 (2-4)	3.5 (3-4)	3.5
Gives an explanation when the patient wakes up	5.0 (3-5)	5.0 (3-5)	5.0

was 55.8 (10.7), (95% confidence interval -4.6 – 5.7, p=0.83). There was also no difference in time to adrenaline administration. Mean (SD) time(s) was 95.5 (40.4) for juniors and 109.3 (39.5) for seniors (95% confidence internal -29.1 – 56.6, p=0.50).

Principal component analysis showed that an acceptable level of unexplained variance from mean total score (<5%), could be achieved using only a single assessor (figure 3).

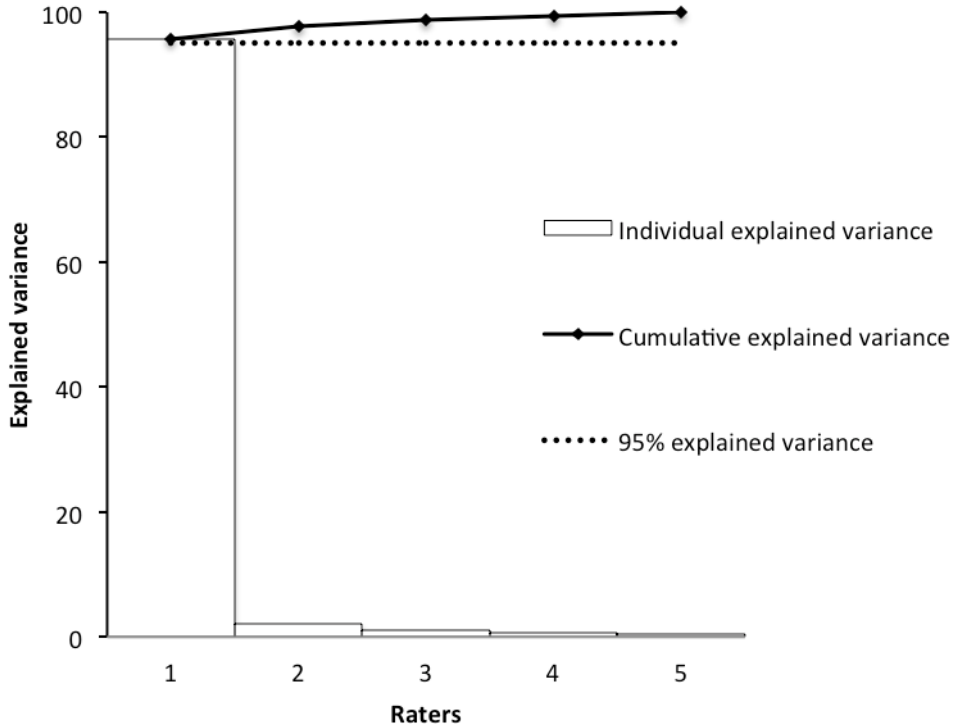
Discussion

The Delphi technique was effective and efficient, producing a task list in only 2 rounds, using an acceptable number of experts. At face value, the tasks appear appropriate to the construct (performance in managing anaphylaxis). High values for Cronbach's alpha (representing high internal consistency), suggest the tool is consistently measuring the construct. It is likely that use of unambiguous task descriptors contributed to this outcome. Choosing descriptors that are sufficiently objective, whilst retaining the ability to capture variation in clinical approaches, can

Figure 2: Intra-class correlation coefficients for the final checklist

Candidates actions	ICC	95% CI
Uses the ABC approach	0.68	0.46 - 0.83
Removes all positive causative agents	0.30	-0.77 - 0.59
Calls for help	0.95	0.92 - 0.97
Administers 100% oxygen	0.94	0.89 - 0.97
Elevates the patients legs if hypotension	0.96	0.94 - 0.98
Starts CPR if necessary	0.74	0.57 - 0.86
Delivers adrenaline in timely manner	0.95	0.91 - 0.97
Adrenaline given in appropriate dose	0.90	0.83 - 0.95
Adrenaline given by appropriate route (IV)	0.98	0.96 - 0.99
Considers starting an IV infusion of adrenaline if several doses needed	0.97	0.94 - 0.98
Considers high rate of fluid therapy	0.96	0.93 - 0.98
Plans transfer to appropriate area	0.92	0.87 - 0.96
Gives antihistamine	0.94	0.90 - 0.97
Gives Steroids	0.93	0.89 - 0.96
Considers alternative vasopressor if needed	0.93	0.89 - 0.96
Treats persistent bronchospasm with IV salbutamol	0.82	0.70 - 0.90
Takes blood sample for mast cell tryptase	0.93	0.88 - 0.96
Takes/arranges for a second blood sample to be taken 1-2 hrs later	0.74	0.53 - 0.86
Liaises with hospital lab for analysis of sample	0.61	0.36 - 0.79
Arranges appropriate investigation of the patient (immunology)	0.78	0.64 - 0.88
Notifies the appropriate (GP, AAGBI anaphylaxis database)	0.80	0.66 - 0.89
Gives an explanation when the patient wakes up	0.88	0.80 - 0.94
Total score	**0.98**	**0.96 – 0.99**

Figure 3: Explained variance from mean score determined by principal component analysis



be challenging. Simulation style also influences this – in this scenario all tasks must be directly observable, as the candidate was not questioned.

High reliabilities indicate the tool could be used by different assessors, retaining a consistent outcome. Very high reliabilities could indicate excessive measurement of the same variable. This may simply represent inefficient over-sampling of the entire construct or, more damagingly, repetitive sampling of a single element. We suggest the high reliabilities seen here be interpreted positively, as broad sampling, for a complex life-threatening situation requiring multiple actions, is desirable.

Several forms of validity have been described[11]. Content validity is the extent to which a test adequately samples the domain of interest. Construct validity indicates whether the test measures the intended construct (higher scores correlate with better true performance). Criterion validity describes how well results align with another test of the same construct and may be concurrent (correlates with a previously validated test) or predictive (used to predict a later test). Content validity here appears high, as tasks were derived from an accepted consensus guideline and refined through a panel of experts. Evidence for construct validity is more difficult to interpret. Similar scores for juniors and seniors may represent the inability of the tool to identify a true performance difference. It may also indicate a weakness in the use of the known groups technique [12] where our assumption that seniors perform better than juniors, is incorrect. We suggest the latter is most likely for several reasons. Firstly, high internal consistency indicates there is consistent measurement of the construct. Secondly, the absence of any difference in time taken to administer adrenaline, independently suggests there is no true performance difference (a form of criterion validity). Finally, we believe that junior trainees are conditioned for and receive frequent training in, emergency drill situations early in their careers. If this method is employed to develop further tools, it may be prudent to increase the experience gap, ensuring any true difference is maximised.

Taking all measures of validity into consideration, we suggest the tool has an acceptable validity profile, but that the experience gap betweenthe two groups did not guarantee a performance difference. To ensure an assessment delivers maximum utility, it's context of use should be suited to its characteristics[13]. As simulation is already a recommended modality for RCoA workplace based assessments[14], this tool may be betst employedas an adjunct to these formative assessmnents. Principal component analysisindicates that an acceptable level of score variance can be achieved with only one assessor, whicch would facilitate efficient resource allocation, whilst assuring reliabilty.

Conclusion

Development of specific, validated assessment tools is clearly an improvement on existing subjective methods. However, it remains unclear if it is practical or cost-effective to develop and validate tools for each individual clinical scenario. In future, it may be more realistic to agree a best practice tool development framework, with the Delphi process at its core.

Practice Points

- The Delphi process is an effective and efficient method of creating checklists
- Careful selection of unambiguous tasks is required for it to succeed
- Caution is required when using the known groups technique: seek independent confirmation of any assumed differences
- Remember not to forgo usability in the search for high reliability and validity
- Adopting a best practice approach to assessment development may be more practical than validating individual tools

References

1. Levy JH, Castells MC. Perioperative anaphylaxis and the United States perspective. *Anesth Analg.* 2011 Nov;113(5):979–81.
2. Mertes PM, Laxenaire M-C, Alla F. Anaphylactic and anaphylactoid reactions occurring during anesthesia in France in 1999-2000. *Anesthesiology.* 2003 Sep;99(3):536–45.
3. Johansson S, Hourihane J. A revised nomenclature for allergy: an EAACI position statement from the EAACI nomenclature task force. *Allergy.* 2001;56(9):813–24.
4. Greenaway D. Shape of Training: securing the future of excellent patient care. Final report of the independent review led by Professor David Greenaway. 2013.
5. Rodriguez-Paz JM, Kennedy M, Salas E, Wu a W, Sexton JB, Hunt E a, et al. Beyond "see one, do one, teach one": toward a different training paradigm. Postgrad Med J. 2009;85:244–9.
6. Dalkley N, Helmer O. An Experimental Application of the Delphi Method to the Use of Experts. Manage Sci. 1963;9(3):458–67.
7. Clayton MJ. Delphi: a technique to harness expert opinion for critical decision making tasks in education. Educ Psychol. 1997 Dec;17(4):373–86.
8. Harper NJN, Dixon T, Dugué P, Edgar DM, Fay a, Gooi HC, et al. Suspected anaphylactic reactions associated with anaesthesia. Anaesthesia. 2009 Feb;64(2):199–211.
9. McLean-Tooke APC, Bethune C a, Fay AC, Spickett GP. Adrenaline in the treatment of anaphylaxis: what is the evidence? BMJ Br Med J. 2003;327(December):1332–5.
10. Fleiss JL. *The Design and Analysis of Clinical Experiements.* 1999. 7 p.
11. Schurwith L, Van der Vleuten C. *How to design an useful test: the principles of assessment. In: Swanwick T, editor. Understanding Medical Education: Evidence, Theory and Practice.* 1st ed. Oxford: Wiley-Blackwell; 2010. p. 195–207.
12. Cronbach LEEJ, Meehl PE. Construct validity of psychological tests. Psychol Bull. 1955;52(4):281–302.
13. Van Der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. Adv Heal Sci Educ. Kluwer Academic Publishers; 1996;1(1):41–67.
14. 14. RCoA. *Curriculum for a CCT in Anaesthetics*. 2010;(August).