Inter-observer reliability of ten tests used for predicting difficult tracheal intubation

Keyvan Karkouti MD FRCPC,* D. Keith Rose MD FRCPC,* Lorraine E. Ferris PhD,† Daniel F. Wigglesworth BSc,* Tina Meisami-Fard BSc,* Henderson Lee MD*

Purpose: *To determine inter-observer reliability of ten preoperative airway assessment tests used for predicting difficult tracheal intubation.*

Method: *We prospectively assessed 59 patients undergoing elective surgery requiring tracheal intubation at a large metropolitan teaching hospital. Two experienced observers independently conducted the airway assessment tests on the same group of patients. Inter-observer reliability was examined using Kappa (K) and intraclass correlation coefficient (ICC).*

Results: *Two tests – mouth opening (ICC = 0.93) and chin protrusion (ICC = 0.89) – had excellent inter-observer reliability. Seven tests – thyromental distance (ICC =0.74), subluxation (K = 0.66), atlanto-occipital extension distance (ICC = 0.67) and angle (K = 0.66), profile classification (K = 0.58), ramus length (ICC = 0.53), oropharyngeal best view (K = 0.49) – were moderately reliable. One test – Mallampati technique of assessing oropharyngeal view (K = 0.31) – had poor reliability.*

Conclusion: *Many of the preoperative airway tests have only moderate inter-observer reliability. This may provide some insight into why previous research has failed to show that the tests accurately predict difficult tracheal intubation.*

Objectif: *Comparer la fiabilité de dix tests préopératoires de dépistage de l'intubation difficile.*

Méthode: *Cinquante-neuf patients soumis à une chirurgie non urgente avec intubation ont été évalués de façon prospective dans un important hôpital métropolitan. Deux observateurs d'expérience ont examiné séparément les voies aériennes des patients d'un même groupe. La fiabilité de l'examen a été comparée avec les coefficients Kappa et de corrélation interclasse (ICC).*

Résultats: *Deux tests – l'ouverture de la bouche (ICC = 0,93) et la protusion du menton (ICC = 0,89) avaient une excellente fiabilité comparée. Sept tests – la distance thyromentionnière (ICC = 0,74), la subluxation (K = 0,66), l'extension (ICC = 0,67) et (K = 0,66) l'angle atlanto-occipitaux, la classification du profil (K = 0,58), la longueur de la branche montante (ICC = 0,53), la meilleure visibilité oropharyngée (K = 0,43) – étaient modérément fiables. Un test – la technique d'évaluation oropharyngée de Mallampati – était peu fiable (K = 0,31).*

Conclusion: *Plusieurs des tests préopératoires ne sont que modérément fiables. Ceci peut expliquer pourquoi les études réalisées antérieurement n'ont pas réussi à démontrer que ces test pouvaient dépister l'intubation difficile.*

Key words
AIRWAY: assessment;
INTUBATION: tracheal.

From the Department of Anaesthesia*, St. Michael's Hospital, Toronto, Ontario; and the Department of Behavioural Science†, Faculty of Medicine, University of Toronto; Clinical Epidemiology Unit, Sunnybrook Health Science Centre.

Presented in part in the Canadian Anaesthetists' Society Resident's Competition, Ottawa, Ontario, June 1995.

Address correspondence to: Dr. Keyvan Karkouti, Department of Anaesthesia, St. Michael's Hospital, 30 Bond Street, Toronto, Ontario, Canada M5B 1W8.

Accepted for publication 16th February, 1996.

Unanticipated difficult laryngoscopic tracheal intubation remains a primary concern of anaesthetists. The failure of the anaesthetist to maintain a patent airway following the induction of general anaesthesia is one of the most frequent causes of anaesthesia related morbidity and mortality.[1] An American Society of Anesthesiologists Closed Claims analysis of adverse outcomes associated with anaesthesia showed that the most common cause of serious injury was due to inadequate ventilation, oesophageal intubation, and difficult tracheal intubation.[2] In 85% of these cases, the outcome was death or brain damage.[2] It is estimated that in the developed world, up to 600 people die every year as a result of complications occurring at the time of tracheal intubation.[3,4] Clearly, preoperative identification of patients whose tracheas will be difficult to intubate would decrease the rate of anaesthesia related adverse respiratory events.

To aid the anaesthetist in identifying these patients, several preoperative airway assessment tests have been

proposed.[3–10] Although the use of these tests, singly or in various combinations, has been advocated by several authors,[4,6,8,10–12] studies show that these tests have a low capability of predicting the difficult airway.[13–19] The inter-observer reliability (repeatability of results) of many of these tests has been questioned.[14,20,21] Since tests with low reliability estimates are not valid they will have little predictive value. Tests that do not have acceptable reliability properties should be replaced with more reliable tests or, factors contributing to the low repeatability of results should be addressed.

Measures of test reliability estimate the proportion of the total variance of test results that is attributable to true differences in the phenomenon under consideration. The higher the reliability, the lower the error variance (variance as a result of irrelevant conditions). While there are different types of reliability, all of which differ in the sources of error variance identified, the most commonly addressed type in tests that predict the difficult airway has been inter-observer reliability. Inter-observer reliability is based on the amount of agreement between independent observers when they conduct the same test on the same patient.[22]

Our review of the literature showed that some tests used to predict difficult tracheal intubation have been examined for inter-observer reliability. However, it was often difficult to determine whether the tests were described adequately to the rater and whether the raters had been trained in the technique with an expert to ensure appropriate technique. It was also difficult to compare the reliability estimates of the tests since we did not know if the reliability studies were conducted under the same conditions and with the same level of rater competence. To address this concern, we examined the inter-observer reliability of ten tests used to predict difficult tracheal intubation and ranked these tests in terms of their repeatability from best to worst.

Methods

Selection of patients

Following institutional approval, 59 patients were selected and assessed using an airway assessment. Fifty-five of the patients were selected from the operative schedule and were assessed preoperatively. Since reliability estimates are more accurate when based on a wide range of measurements, and since the prevalence rate of having a difficult tracheal intubation is low, we increased the number of difficult cases in our study population by including in our sample four patients who had recently required multiple laryngoscopies for tracheal intubation.

Patients were considered eligible if they were to receive, or had recently received, general anaesthesia for elective surgery requiring tracheal intubation. The exclusion criteria included an inability to give consent, an unstable C-spine, gross anatomical abnormalities of head or neck, or any recent surgery involving the head or neck.

Measures

We conducted a MEDLINE computer search from 1980 to 1995 for tests that could be useful in identifying patients with difficult tracheal intubation.[3–10] We selected 10 based on those that were the most common and could be completed at the bedside. For each test, based on the available data, the most valid method of examination was used.[23,24] Two of the selected tests assessed mandibular movement, four assessed mandibular space, two assessed atlanto-occipital extension, and two assessed the visibility of oropharyngeal cavity structures.

Mandibular movement was tested using "mouth opening (inter-incisal gap)"[25] and "subluxation" of the mandible.[7,23] For mouth opening, each patient maximally opened his/her mouth and the distance between the upper and lower incisors in the mid-line was measured. In the edentulous patient, the distance between the upper and lower gingiva was measured. For "subluxation" of the mandible, the patient protruded the lower incisors as far forward as possible. This assessment was ranked depending upon the amount of anterior mandibular movement: grade 1 if the lower incisors were anterior to the upper incisors; grade 2 if the lower incisors were equal to the upper incisors; and grade 3 if the lower incisors failed to reach the upper incisors and remained posterior.

Four of the tests measured the mandibular space. The first was "thyromental distance".[9,24] For this test, each patient extended his/her head and neck as far as possible with mouth closed. The straight distance from the inside of the mentum to the thyroid notch was measured. The second test was the "length of mandibular ramus".[10] The ramus length was measured from the temporal mandibular joint to the angle of the mandible. The remaining two tests measured the severity of mandibular retrognathia. One was "profile classification".[7] For this test, the raters used a straight edge to draw an imaginary line joining the most prominent part of the brow and the maxilla. If the most anterior part of the chin was behind this line, the patient was classified as having a retrognathic mandible. If the mandible was equal or anterior to this line, they were classified as having a neutral or protruding mandible respectively. The other test was "chin protrusion".[3] For this test, the patient assumed the sniffing position (complete flexion of the lower cervical spine

and extension of the atlanto-occipital joint, see below) and opened his/her mouth as wide as possible. An imaginary line was drawn from the tip of the upper incisors to the most anterior part of the thyroid cartilage. The perpendicular distance to the genial tubercle in front of this line was measured.

Two tests were used to assess atlanto-occipital extension.[3,23,26,27] For the first test, a goniometer was used to measure the angle transversed by the occlusal surfaces of the maxillary teeth as the atlanto-occipital joint was extended from complete flexion to the sniffing position. To obtain the sniffing position, the patient maximally flexed his/her cervical spine and then tilted the head up as far as possible without moving the neck. The rater's hand was placed on the neck to ensure immobilization of the lower cervical spine.[23] The patients were classified based on the extent of atlanto-occipital extension: greater than 66% extension, between 33% and 66% extension, and less than 33% extension. The second test was performed in two steps. The patient maximally flexed his/her head and neck and the distance between the genial tubercle and the sternal notch was measured. Then the patient assumed the sniffing position (described above) and the distance between the genial tubercle and the sternal notch was re-measured. The first measurement was subtracted from the second for a total atlanto-occipital extension score.[27]

The final two tests assessed the oropharyngeal view. For the "classic Mallampati technique," each patient sat with his/her head in neutral position, mouth open, and tongue protruded maximally without phonating.[5] Raters noted which oropharyngeal structures were visible. For the "best possible view" technique, all attempts were made to obtain the best possible view of the oropharyngeal structures by having the patient either remain quiet or phonate, extend or flex the head and neck, and retract or protrude the tongue.[21,24,28] For both methods, patients were graded as described by Mallampati and modified by Samsoon and Young: grade 1, good visualization of the soft palate, fauces, uvula, and tonsillar pillars; grade 2, pillars obscured by the base of the tongue, but posterior pharyngeal wall clearly visible below the soft palate; grade 3, soft palate and base of the uvula visible; and grade 4, soft palate not visible.[6]

Procedure

Before the start of the study, the two raters (specially trained senior anaesthesia residents) were instructed using patient volunteers. Exact measurements were made for each variable using an accurate measuring device (C-THRU Inch/Metric Protractor Ruler model B-75).

Once the rater training was complete, patients were selected and recruited by one of the raters and were then assessed by both raters. The raters assessed each patient independently and recorded the information on a preprinted form that also contained clear instructions and diagrams for each test. Most patients were recruited by rater #1, hence most patients were first assessed by rater #1 and then by rater #2. To avoid patient fatigue, the tests were carried out in an order that minimized patient exertion and discomfort, and patients were allowed to rest between the two assessments. To assure consistent performance for both raters, each rater had the patients repeat the required manoeuvres until performed correctly.

Analysis

Results from tests that used a discrete or categorical scale were analyzed using the Kappa statistic[29] whereas for those with a continuous scale the intraclass correlation coefficient (ICC) was used.[22] Scores for both these statistical measurements range from "0" to "1" where the former shows no reliability and the latter perfect reliability. Although there are no universally accepted standards, convention suggests that scores greater than or equal to 0.75 are excellent, between 0.4 and 0.75 are moderate and less than or equal to 0.4 are poor.[22,29]

The relative rater bias for tests with continuous variables was assessed by first calculating the mean of the differences between the measurements of the two raters for each patient. The mean of the difference for the test was then calculated by adding the mean of the differences for the entire patient population and then dividing this by the number of patients. Values significantly different from zero ($P < 0.01$) show that there is a systematic bias between the two raters for that test.[30] To demonstrate the clinical importance of this value, the mean of the difference for the test was divided by the mean of the measurement itself and expressed as a percentage.

Results

Table I shows the overall reliability estimates for the ten clinical tests. Kappas ranged from 0.31 to 0.66 and intraclass correlation coefficients from 0.53 to 0.93.

Tables II, III and IV depict details about the interobserver agreement for some of the tests where patients were graded according to a set criteria. Table II shows that the raters had a disagreement rate of 12% in classifying patients as Grade 1 or Grade 2 for subluxation (no patients were grade 3). Tables III and IV show that the raters had a disagreement rate of 32% in assessing oropharyngeal view using the classic Mallampati technique dropping to 3.4% using the best view technique.

The mean of the differences (rater bias) for tests using

TABLE I Reliability (reproducibility) of airway tests

Variable	*Test*	*K*[a]	*ICC*[b]	*Reliability*
Mandibular movement	Mouth opening	–	0.93	Excellent
	Subluxation	0.66	–	Moderate
Mandibular space	Chin protrusion	–	0.89	Excellent
	Thyromental distance	–	0.74	Moderate
	Profile classification	0.58	–	Moderate
	Ramus length	–	0.53	Moderate
Atlanto-occipital extension	Distance	–	0.67	Moderate
	Angle	0.66	–	Moderate
Oropharyngeal view	Best view	0.49	–	Moderate
	Classic mallampati	0.31	–	Poor

[a]K = Kappa.
[b]ICC = Intraclass Correlation Coefficient.
K and ICC interpreted as follows: excellent reliability if greater than or equal to 0.75, moderate reliability if between 0.4 and 0.75, poor reliability if less than or equal to 0.4.[22,29]

TABLE II Contingency table for subluxation

		Rater 1		
		Grade I	*Grade II*	*Grade III*
Rater 2	Grade I	41	4	0
	Grade II	3	11	0
	Grade III	0	0	0

In the 59 patients assessed, raters agreed with each other's classification in 52 patients while they disagreed in 7 patients (disagreement rate 7/59 or 12%, Kappa = 0.66).

TABLE III Contingency table for classic Mallampati technique of assessing oropharyngeal view

		Rater 1	
		Class I & II	*Class III & IV*
Rater 2	Class I & II	30	16
	Class III & IV	3	10

Raters agreed with each other's classification in 40 patients while they disagreed in 19 patients (disagreement rate 19/59 or 32%, Kappa = 0.31).

TABLE IV Contingency Table for best view technique of assessing oropharyngeal view

		Rater 1	
		Class I & II	*Class III & IV*
Rater 2	Class I & II	56	2
	Class III & IV	0	1

Raters agreed with each other's classification in 57 patients and disagreed in 2 patients (disagreement rate 2/59 or 3.4%, Kappa = 0.49).

a continuous scale was different from zero for all the tests ($P < 0.01$). These differences were less than 7% of the actual mean of all measurements (Table V).

Discussion

We found that under ideal conditions, the inter-observer reliability estimates for the ten tests varied. Only two of 10 clinical tests commonly used for predicting difficult tracheal intubation (mouth opening and chin protrusion) have excellent inter-observer reliability. Seven (subluxation, thyromental distance, profile classification, ramus length, atlanto-occipital extension distance and angle, and oropharyngeal best view) have moderate reliability. One of the tests (classic Mallampati technique of assessing oropharyngeal view) has poor reliability.

In clinical practice, there are several factors that may contribute to lower reliability estimates. For example, if patients do not follow instructions appropriately or consistently, or find it difficult to assume a position, reliability estimates will be lowered. To increase the reliability of the tests, patients need to have the required manoeuvres clearly described and, where necessary, to have them demonstrated. Asking patients to repeat the manoeuvres until performed correctly will also help.

Rater factors may also decrease inter-observer reliability estimates. This occurs if a rater consistently classifies or measures differently from another rater (rater bias). This rater error may be due to the ambiguity of the definitions. For example, Mallampati's description of oropharyngeal view is different from Samsoon and Young's classification.[5,6] Furthermore, Samsoon and Young's class III includes the visualization of the base

TABLE V The mean of differences (measure of rater bias) for tests with continuous variables

Test	Mean of differences (cm)[a]	Mean of all measurements (cm)[b]	a/b
Mouth opening	−0.12	4.27	2.8%
Chin protrusion	+0.10	2.90	3.4%
Thyromental distance	−0.14	7.75	1.8%
Ramus length	−0.37	6.15	6.0%
Atlanto-occipital extension distance	−0.81	11.50	7.0%

[a]Mean of the differences between the measurements of the two raters for each patient.
[b]Mean of all measurements (two for each of 59 patients) for each test.
[a/b]Expresses the mean of the differences (*a*) (converted to positive values) as a percentage of the mean of the measurements (*b*).
For all the tests, the mean of differences is significantly different from zero ($P < 0.01$) indicating that some systematic rater bias existed for all tests. Since this difference is small relative to the actual mean of all measurements (less than 7%), this amount of bias may not be clinically relevant.

of the uvula, and this feature may be prone to classification errors.[6] Moreover, different researchers have used different land marks and head positions to measure the thyromental distance.[8,9,12,18,24] It is important that anaesthetists are provided with clear and detailed definitions and diagrams for the tests.

Lastly, the measurement technique may affect the reliability of the tests. Factors such as digit preference (rounding off to the nearest whole number) and measurement error will decrease the reliability of the test.[22] This is especially true for many of the tests in this study since the measured values are relatively small. To reduce this source of variability, an accurate measuring device should be used, rounding off avoided, and results recorded immediately on a pre-printed form.

This inter-observer reliability study was performed under optimal conditions. The required manoeuvres were clearly described to the patients and were demonstrated to them when necessary. Raters were provided with clear and detailed definitions with accompanying diagrams, and were trained until they reached proficiency. Moreover, we used an accurate measuring device, rounding off was avoided and results were immediately recorded. However, even under these optimal conditions, the majority of the tests were only moderately reproducible, establishing that the sources of variability were not entirely eliminated. Some of the remaining variability may be related to the patient; for example, they may find it difficult to assume a position consistently. Another source may be due to rater factors. Examining the classification of oropharyngeal view in Table III, we can conclude that rater bias was not eliminated for this test since rater #1 classified significantly more patients into classes III and IV than rater #2. The persisting rater bias for the other tests was too small to be clinically important (Table V). The variability due to rater factors may have been higher, and the reliability of the tests lower, if more than two raters were included in the study. Since the goal of this study was to obtain the maximum inter-observer reliability of the tests, we only used two raters. The reliability of these tests may also be improved if devices such as positioning aids, photographs, or radiographs are used. However, a recent study that assessed the reliability of the Mallampati test using photographs of the oropharynx found that the test was still only moderately reliable.[31] Since these tests need to be easy and quick to perform at the bedside by the anaesthetist, we did not use any extra accessories.

To maximize further the reliability of the tests by increasing the between-patient variance, the sample population was enriched with patients who had abnormal airways. However, since this study was designed only to assess the reliability of the tests and not their ability to predict difficult intubations (i.e., valid answers), we did not make any attempts to assess the predictive capability of the tests. Future large scale studies are required to assess the validity of the tests.

In conclusion, many of the current tests used for predicting difficult tracheal intubation have only moderate inter-observer reliability and this is under optimal conditions. If optimal conditions are not provided, the reliability of the tests will be even lower, and they will have low predictive values.[22] Our findings may have an impact on clinical practice since they suggest that to use these tests for predicting difficult laryngoscopic tracheal intubations, they must be performed in a precise manner. Moreover, our results suggest one possible reason why research has failed to show that the tests accurately predict difficult tracheal intubations. Further reliability and validity studies are needed to address this issue.

References

1 *Marks JD, Bogetz MS.* New concepts in the management of the difficult airway. *In:* Barash PG, Cullen BF, Stoelting RK (Eds.). Clinical Anaesthesia Update, 2nd ed. Philadelphia: J.B. Lippincott Company, 1994: 1–11.

2 *Caplan RA, Posner KL, Ward RJ, Cheney FW.* Adverse respiratory events in anesthesia: a closed claim analysis. Anesthesiology 1990; 72: 828–33.

3 *Bellhouse CP, Doré C.* Criteria of estimating likelihood of difficulty of endotracheal intubation with the Macintosh laryngoscope. Anaesth Intensive Care 1988; 16: 329–37.

4 *King TA, Adams AP.* Failed tracheal intubation. Br J Anaesth 1990; 65: 400–14.

5 *Mallampati SR, Gatt SP, Gugino LD, et al.* A clinical sign

to predict difficult tracheal intubation: a prospective study. Can Anaesth Soc J 1985; 32: 429–34.
6 *Samsoon GLT, Young JRB.* Difficult tracheal intubation: a retrospective study. Anaesthesia 1987; 42: 487–90.
7 *Wilson ME, Spiegelhalter D, Robertson JA, Lesser P.* Predicting difficult intubation. Br J Anaesth 1988; 61: 211–6.
8 *Mathew M, Hanna LS, Aldrete JA.* Pre-operative indices to anticipate difficult tracheal intubation. Anesth Analg 1989; 68: S187.
9 *Frerk CM.* Predicting difficult intubation. Anaesthesia 1991; 46: 1005–8.
10 *Chou H-C, Wu T-L.* Mandibulohyoid distance in difficult laryngoscopy. Br J Anaesth 1993; 71: 335–9.
11 *Benumof JL.* Management of the difficult adult airway. With special emphasis on awake tracheal intubation. Anesthesiology 1991; 75: 1087–110.
12 *Rocke DA, Murray WB, Rout CC, Gouws E.* Relative risk analysis of factors associated with difficult intubation in obstetric anesthesia. Anesthesiology 1992; 77: 67–73.
13 *Deller A, Schreiber MN, Gramer J, Ahnefeld FW.* Difficult intubation: incidence and predictability. A prospective study of 8,284 adult patients. Anesthesiology 1990; 73: A1054.
14 *Oates JDL, MacLeod AD, Oates PD, Pearsall FJ, Howie JC, Murray GD.* Comparison of two methods for predicting difficult intubation. Br J Anaesth 1991; 66: 305–9.
15 *McDonald JS, Gupta B, Cook RI.* Proposed methods for predicting difficult intubation: prospective evaluation of 1501 patients. Anesthesiology 1992; 77: A1125.
16 *Bellhouse CP.* Predicting difficult intubation (Letter). Anaesthesia 1992; 47: 440–1.
17 *Bellhouse CP, Dore C.* Using clinical assessment to predict difficult direct laryngoscopy. Can J Anaesth 1992; 39: A117.
18 *Savva D.* Prediction of difficult tracheal intubation. Br J Anaesth 1994; 73: 149–53.
19 *Rose DK, Cohen MM.* The airway: problems and predictions in 18,500 patients. Can J Anaesth 1994; 41: 372–83.
20 *Wilson ME, John R.* Problems with the Mallampati sign (Letter). Anaesthesia 1990; 45: 486–7.
21 *Tham EJ, Gildersleve CD, Sanders LD, Mapleson WW, Vaughan RS.* Effects of posture, phonation and observer on Mallampati classification. Br J Anaesth 1992; 68: 32–8.
22 *Fleiss JL.* Reliability of measurement. *In:* The Design and Analysis of Clinical Experiments. New York: John Wiley & Sons. 1986: 1–32.
23 *Calder I.* Predicting difficult intubation (Letter). Anaesthesia 1992; 47: 528–9.
24 *Lewis M, Keramati S, Benumof JL, Berry CC.* What is the best way to determine oropharyngeal classification and mandibular space length to predict difficult laryngoscopy? Anesthesiology 1994; 81: 69–75.
25 *Aiello G, Metcalf I.* Anaesthetic implications of temporomandibular joint disease. Can J Anaesth 1992; 39: 610–6.
26 *Youdas JW, Carey JR, Garrett TR.* Reliability of measurements of cervical spine range of motion – comparison of three methods. Phys Ther 1991; 71: 98–106.
27 *Chow FL, Duncan PG, Code WE, Yip RW.* Can bedside neck extension predict difficult intubation? Can J Anaesth 1993; 40: A4.
28 *Oates JDL, Oates PD, Pearsall FJ, McLeod AD, Howie JC.* Phonation affects Mallampati class (Letter). Anaesthesia 1990; 45: 984.
29 *Rosner B.* Hypothesis testing: categorical data. *In:* Fundamentals of Biostatistics, 4th ed. Belmont, CA: Wadsworth Publishing Company. 1995: 423–6.
30 *Altman DG, Bland JM.* Measurement in medicine: the analysis of method comparison studies. The Statistician 1983; 32: 307–17.
31 *Pilkington S, Carli F, Dakin MJ, et al.* Increase in Mallampati score during pregnancy. Br J Anaesth 1995; 74: 638–42.