Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub. 2011 Sep; 155(3):275–282. DOI 10.5507/bp.2011.041 275

INTER-OBSERVER REPRODUCIBILITY OF 15 TESTS USED FOR PREDICTING DIFFICULT INTUBATION

Milan Adamus[a]*, Ondrej Jor[b], Tereza Vavreckova[b], Lumir Hrabalek[c], Jana Zapletalova[d,e], Tomas Gabrhelik[a], Hana Tomaskova[f], Vladimir Janout[g]

[a] *Department of Anesthesiology and Intensive Care Medicine, University Hospital Olomouc and Faculty of Medicine and Dentistry, Palacky University Olomouc, Czech Republic*
[b] *Faculty of Medicine and Dentistry, Palacky University Olomouc*
[c] *Department of Neurosurgery, University Hospital Olomouc and Faculty of Medicine and Dentistry, Palacky University Olomouc*
[d] *Department of Medical Biophysics, Faculty of Medicine and Dentistry, Palacky University Olomouc*
[e] *Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc*
[f] *Department of Epidemiology and Public Health, Faculty of Medicine, Ostrava University*
[g] *Department of Preventive Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc*
E-mail: milan.adamus@seznam.cz

Received: May 17, 2011; Accepted: July 12, 2011

Key words: Reproducibility/Prediction/Difficult intubation/Inter-Observer Variability

Aim. To determine the inter-observer reproducibility of 15 tests used for predicting difficult tracheal intubation (DI).

Material and methods. Following local ethics committee approval and informed consent, 101 volunteers were examined by two assessors using 15 tests for predicting DI. The two assessors who were blinded to the results of the other, examined each volunteer independently. Cohen's kappa (κ) or first-order agreement coefficient (AC1) were used to measure agreement between assessor ratings on a qualitative scale. Agreement between two quantitative outcomes was described using the intraclass correlation coefficient (ICC) and Pearson's (PCC) or Spearman's (SCC) correlation coefficients. The following interpretation of the coefficients was used: poor (< 0.20), fair (0.21–0.40), satisfactory (0.41–0.60), good (0.61–0.80), and excellent (0.81–1.00).

Results. Respective coefficients of inter-rater agreement and correlation coefficients were determined for the following parameters: pathologies associated with DI (κ=0.662, AC1=0.990), clinical impression (κ=-0.013, AC1=0.969), modified Mallampati test (κ=0.503, AC1=0.861), upper lip bite test (κ=0.370, AC1=0.897), temporo-mandibular joint movement (κ=0.088, AC1=0.797), max. anteroflexion of C-spine (ICC=0.136, SCC=0.391), max. retroflexion of C-spine (ICC=0.020, SCC=0.284), mandibular length (ICC=0.301, SCC=0.553), neck circumference (ICC=0.832, SCC=0.928), hyo-mental distance (ICC=0.378, SCC=0.472), thyro-mental distance (ICC=–0.002, PCC=0.265), sterno-mental distance (ICC=0.674, PCC=0.815), and finally, inter-incisor gap (ICC=0.695, PCC=0.785). Two tests (positive history of DI and retrogenia), were excluded from calculation because no positive cases were found.

Conclusion. Best inter-rater agreement was found for the assessment of neck circumference while the highest discrepancies between raters were in goniometrically-measured mobility of the C-spine.

Many of the pre-operative airway tests had only fair inter-observer reproducibility. This may be one reason why models for predicting difficult intubation are not universally reliable.

INTRODUCTION

Tracheal intubation is a mainstay of airway management during general anesthesia and usually performed uneventfully. However, if the intubation appears to be difficult or impossible after induction of anesthesia, critical oxygen desaturation may occur. Unanticipated difficult intubation can be more dangerous than a predicted one when a potential airway problem is detected before anesthesia.

There are many clinical tests for predicting difficult intubation (DI). Unfortunately, single tests, such as Mallampati classification, are not reliable enough to prospectively detect all cases of DI (ref.[1-4]). Combining several tests may be more effective and may improve the accuracy of the assessment. However, studies have demonstrated conflicting results in the predictive value of models consisting of different test combinations[5-7]. A correct model must be both reliable in classifying patients' airway and give the same results when performed by different assessors (reproducibility).

The aim of this study was to determine the inter-rater agreement between two assessors (medical students) using fifteen parameters for predicting difficult intubation.

MATERIALS AND METHODS

Following local ethics committee approval and informed consent, 101 volunteers (medical students) were

Table 1. Parameters for predicting difficult intubation used in the study.

Parameter	Methods	Ranges and Units
Positive history of DI	Does the patient have a history of difficult intubation?	yes – no
Pathologies associated with DI	Does the patient have a condition commonly associated with difficult intubation?	specify (ankylosing spondylitis, acromegaly, etc.)
Clinical impression	Hard to define – a feeling of a potential airways problem	yes – no
Mallampati classification as modified by Samsoon and Young (MMT, Modified Mallampati test) (ref. [16])	Head in neutral position, full mouth opening, protrusion of the tongue: pharyngeal view	1 = the soft palate, fauces, uvula and pillars visible 2 = the soft palate, fauces and base of uvula visible 3 = the soft palate visible 4 = the hard palate only visible
Upper lip bite test (ULBT) (ref.[17-19])	Biting the upper lip with the lower incisors	1 = the incisors in front of the lip 2 = the lip partly visible 3 = the lip visible
Retrogenia (receding mandible)	A line drawn from the upper eye lid to the maxilla	yes – the chin behind the line no – the chin in front of the line
Hyo-mental distance (HMD)	Distance: the body of the hyoid bone – the mentum	mm
TM joint movement	Full mouth opening (IIG) + slux	1 = IIG > 50 mm + slux > 0 2 = IIG < 50 mm + slux > 0 3 = IIG < 50 mm + slux < 0
Maximal anteroflexion of the C-spine	• The goniometer head to the ear canal • First arm in the long axis of the neck above the ear • Second arm to the nasal wing	degrees
Maximal retroflexion of the C-spine		degrees
Mandibular length	Distance: outer angle – the middle of the chin (follow the shape of the mandible)	cm
Neck circumference	Measured at the level of the cricoid, perpendicular to the long axis of the neck	cm
Thyro-mental distance (TMD)	Distance: superior thyroid notch – the lower edge of the middle of the chin	mm
Sterno-mental distance (SMD)	Distance: jugulum – the lower edge of the middle of the chin	mm
Inter-incisor gap (IIG)	Full mouth opening, distance between the incisors (gums)	mm

DI = difficult intubation, TM = temporo-mandibular, slux = subluxation (maximal forward protrusion of the lower incisors beyond the upper incisors)

examined (2–5/day) with 15 tests for predicting difficult intubation. In random order, each volunteer was examined in one session independently by two co-authors (O. J., T. V.) who were thoroughly instructed in carrying out the tests. Examinations performed by assessor O. J. created the O group while the group T included the corresponding examinations done by T. V. The measurements were done under standardized conditions and each assessor was blinded to the results of the other. The airway assessment consisted of fifteen parameters and measurements (see Table 1).

The data were recorded into an Excel spreadsheet application (Microsoft Office 2007 SP2, Microsoft Corporation), and statistically analyzed (SPSS v. 15.0 statistical software, SPSS Inc., Chicago, USA).

Descriptive statistics was used to summarize the demographic data of the volunteers and comparison of genders was done with a Mann-Whitney U test. A p-value less than 0.05 was considered significant. Agreement between assessors (percentage), Cohen's kappa (κ), or first-order agreement coefficient (AC1) were used for comparison of qualitative parameters. The inter-rater agreement in measurement of quantitative parameters was analyzed with the intraclass correlation coefficient (ICC) with 95% confidence intervals and Pearson's or Spearman's correlation coefficients. We used the following interpretation[8] of the inter-rater agreement in the kappa values and correlation coefficients: poor (< 0.20), fair (0.21–0.40), satisfactory (0.41–0.60), good (0.61–0.80), and finally excellent (0.81–1.00). The distribution of inter-observer differences was tested with the Kolmogorov-Smirnov test. Data with normal distributions were compared using the paired Student's t-test; the Wilcoxon signed-rank test was used for data that did not pass the normality test. Scatter plot and Bland-Altman plot[9] were used to demonstrate the systematic bias for measurements between assessors.

RESULTS

A total of 101 volunteers were enrolled and they all successfully finished the study with no drop-outs. Thirty (29.7%) were males (median age 23 years, range 20–26 years; median height 184 cm, range 173–190 cm; median weight 82 kg, range 66–100 kg; median BMI 24.0 kg m^{-2}, range 21.1–31.7 kg m^{-2}) and 71 (70.3%) were females (median age 23 years, range 20–25 years; median height 168 cm, range 153–182 cm; median weight 61 kg, range 45–83 kg; median BMI 21.5 kg m^{-2}, range 16.2–30.5 kg m^{-2}). There was no significant difference in age for the two genders (p=0.847). Compared to females, males were significantly taller (p<0.0001), heavier (p<0.0001) and had higher BMI (p<0.0001).

Two tests (positive history of DI and retrogenia), were excluded from calculation because no positive cases were found. The coefficients of the inter-rater agreements of the qualitative and quantitative tests are given in Tables 2 and 3, respectively.

There was a systematic bias between assessors for measurements of all quantitative parameters (see Table 4, Fig. 1 and Fig. 2).

DISCUSSION

We demonstrated variable inter-observer reproducibility of tests for predicting DI. The best agreement between assessors was found for determining neck circumference; the worst results were obtained for the goniometric measurements (anteroflexion and retroflexion of cervical spine). Inter-observer reproducibility of a test depends upon factors related to both rater and person examined[10]. Rater components of errors include incorrect/inconsistent measurement technique that may be due to insufficient instructions and/or inaccurate methodology.

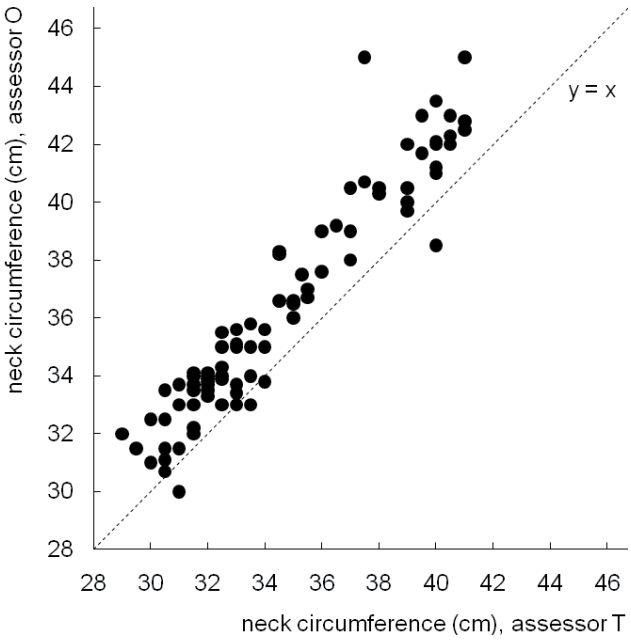

Fig. 1. Scatter plot of assessor's T measurements of neck circumference against the paired measurements of assessor O. Bias is present, the measurements of assessor T were systematically lower than the measurements of assessor O.

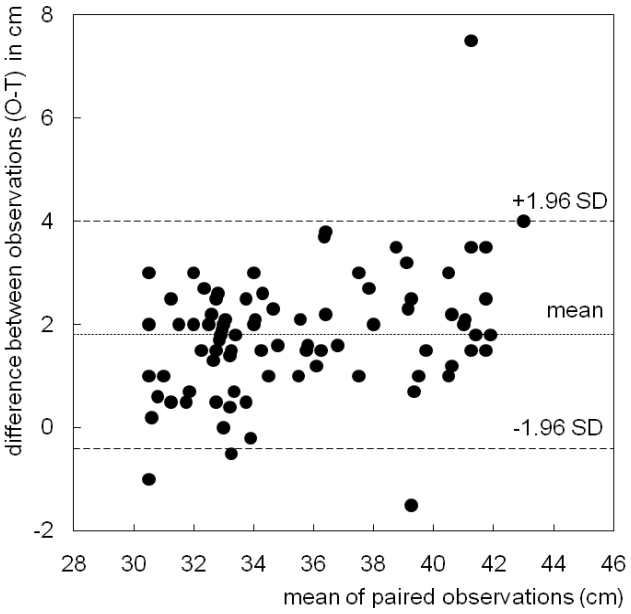

Fig. 2. Bland-Altman plot of differences between assessors in paired measurements of neck circumference against mean neck circumference, in 101 volunteers.

Each test must be described as simply as possible but the accuracy of the measurement procedure has to be maintained. To reduce this potential source of error, we used both written description of all measurements and practical training of the assessors before the start of the study. However, it is doubtful whether these steps were

Table 2. Inter-observer agreement – qualitative parameters.

Parameter	Inter-observer agreement (%)	Cohen's kappa ± SE	Strength of agreement (based on kappa)	AC1	Strength of agreement (based on AC1)
Pathologies associated with DI	99	0.662 ± 0.316	good	0.990	excellent
Clinical impression	97	-0.013 ± 0.009	poor	0.969	excellent
MMT (I, II, III, IV)	68	0.452 ± 0.073	satisfactory	n/A	n/A
MMT (I+II, III+IV)	89	0.503 ± 0.127	satisfactory	0.861	excellent
ULBT (I, II, III)	68	0.443 ± 0.079	satisfactory	n/A	n/A
ULBT (I+II, III)	91	0.370 ± 0.156	fair	0.897	excellent
TM joint movement (I, II, III)	64	0.390 ± 0.067	fair	n/A	n/A
TM joint movement (I+II, III)	83	0.088 ± 0.083	poor	0.797	good

SE = standard error, AC1 = first-order agreement coefficient[14,15], n/A = not applicable
DI = difficult intubation, MMT = Modified Mallampati test[16], ULBT = Upper lip bite test[17–19], TM = temporo-mandibular
Inter-observer agreement[8]: poor (< 0.20), fair (0.21–0.40), satisfactory (0.41–0.60), good (0.61–0.80), excellent (0.81–1.00)

Table 3. Inter-observer agreement – quantitative parameters.

Parameter	ICC	95%CI	Strength of agreement (based on ICC)	Correl. coef.	Strength of agreement (based on correl. coef.)
Maximal anteroflexion of the C-spine	0.136	-0.060–0.322	poor	0.391[a]	fair
Maximal retroflexion of the C-spine	0.020	-0.175–0.213	poor	0.284[a]	fair
Max. anteroflexion + retroflexion of the C-spine	-0.109	-0.297–0.087	poor	0.313[a]	fair
Mandibular length	0.301	0.113–0.468	fair	0.553[a]	satisfactory
Neck circumference	0.832	0.761–0.884	excellent	0.928[a]	excellent
HMD	0.378	0.198–0.533	fair	0.472[a]	satisfactory
TMD	-0.002	-0.196–0.192	poor	0.265[b]	fair
SMD	0.674	0.553–0.768	good	0.815[b]	excellent
IIG	0.695	0.579–0.784	good	0.785[b]	good

ICC = Intraclass correlation coefficient, 95%CI = Confidence interval 95%, [a] Spearman's correlation coefficient, [b] Pearson's correlation coefficient
HMD = Hyo-mental distance, TMD = Thyro-mental distance, SMD = Sterno-mental distance, IIG = Inter-incisor gap
Inter-observer agreement[8]: poor (< 0.20), fair (0.21–0.40), satisfactory (0.41–0.60), good (0.61–0.80), excellent (0.81–1.00).

Table 4. Systematic deviation (bias) in the measurements of quantitative parameters. Minimum, maximum, mean, and percentiles of differences between assessors.

Parameter [difference between assessors O and T]	mean	range	percentiles		
			25 (lower quartile)	50 (median)	75 (upper quartile)
dif Max. anteroflexion of the C-spine (degrees) [O-T]	-8.2	-34.0–35.0	-15.0	-8.0	-1.0
dif Max. retroflexion of the C-spine (degrees) [O-T]	-10.1	-35.0–14.0	-17.0	-9.0	-2.5
dif Max. anteroflexion+Max. retroflexion (degrees) [O-T]	-18.2	-51.0–35.0	-27.5	-19.0	-7.5
dif Mandibular length (cm) [O-T]	0.7	-1.3–2.5	0.1	0.5	1.1
dif Neck circumference (cm) [O-T]	1.8	-1.5–7.5	1.3	2.0	2.3
dif HMD (mm) [O-T]	-4.3	-25.0–21.0	-10.0	-3.0	0
dif TMD (mm) [O-T]	-13.6	-55.0–38.0	-22.0	-13.0	-7.0
dif SMD (mm) [O-T]	-10.6	-77.0–30.0	-15.0	-10.0	-5.0
dif IIG (mm) [O-T]	-3.5	-19.0–7.0	-5.0	-3.0	-0.5

O = assessor O, T = assessor T

HMD = Hyo-mental distance, TMD = Thyro-mental distance, SMD = Sterno-mental distance, IIG = Inter-incisor gap

sufficient because the inter-rater bias in the measurements of quantitative parameters was substantial. Factors related to the examined volunteer may be based on misunderstanding or not following the instructions appropriately. When necessary, the required maneuvers were clearly described several times and demonstrated repeatedly[11].

This study is relevant not only to pre-anesthetic airway assessment and predicting DI. The results present a **statistical challenge**, too.

The Cohen's kappa coefficient (κ) is a statistical measure of inter-rater agreement for qualitative (categorical) items[12,13]. Kappa-values range from -1.0 to 1.0. Negative values occur when agreement is weaker than expected by chance. When we get κ=0, the agreement is the same as would be expected by chance, κ=1.0 indicates perfect agreement above chance. However, in our study, the high percentage agreement between assessors for some parameters did not correspond to low κ-values (see Table 2). For these parameters, first-order agreement coefficient (AC1) was calculated as an alternative to the κ coefficient. Some authors believe that the AC1 value reflects the inter-rater agreement more realistically than the Cohen's kappa coefficient[14,15]. The limitation of the AC1 is that it can be used for contingency tables 2 × 2 only. When a qualitative parameter has more than two values, either another agreement coefficient (AC2) has to be used or some groups have to be merged for AC1 calculations. In our study, MMT (Modified Mallampati test) (ref.[16]), ULBT (Upper lip bite test)[17-19] and Temporo-mandibular (TM) joint movement were the relevant parameters suitable for merging groups.

In three volunteers, clinical impression of potential DI was positive (one in the O group, two in the T group). This parameter had poor inter-observer correlation when measured with Cohen's kappa coefficient (-0.013), but excellent if AC1 was used (0.969). Very low incidence of positive cases is a limitation of these results.

No volunteer declared a positive history of DI. This could be for two reasons. Either the volunteer had had no anesthesia or the intubation for his/her previous anesthesia was not difficult. As we were unable to distinguish between these two groups and the incidence of positive anamnesis of DI was zero, this parameter was excluded from statistical analysis. The same applied to retrogenia: there was no positive case in the groups.

Pathologies associated with DI were detected only in two volunteers. The agreement between the assessors was as high as 99%. Based on the Cohen's kappa coefficient, good agreement was detected (κ=0.662). When AC1 was used, the strength agreement was graded as excellent (AC1=0.99). These results show the advantage of AC1 over Cohen's kappa coefficient. As we intuitively feel, when the examinations of the assessors were identical in 99% cases, the degree of agreement should be described as excellent. On the other hand, one must take into consideration the low incidence of pathologies determined and probably not present in the study group.

Four Grades (I, II, III, IV) in the MMT (Modified Mallampati test) (ref.[1-4,16]) were merged into two – (I + II) and (III + IV). This is reasonable because Grades I and II are considered an unrestricted pharyngeal view while Grades III and IV indicate a limited view[3,16]. 89% volunteers were assigned the same MMT Grade by both assessors, the Cohen's kappa coefficient was 0.503 (satisfactory), and AC1 was 0.861 (excellent). In a previous study, we demonstrated that the MMT, when performed alone, was not a reliable predictor of DI. This limitation of MMT may partially be due to inconsistent assessments by the observers.

For the ULBT (ref.[17-19]), there was agreement between the assessors in 91% cases (κ=0.370, AC1=0.897). Similarly, in the parameter TM joint movement, the inter-observer agreement was in 83% cases (κ=0.088, AC1=0.797).

The inter-rater agreement for quantitative measurements was assessed with ICC and correlation coefficients (based on data distribution, Pearson's or Spearman's coefficient were used). Generally, strength of agreement based on ICC was more conservative (giving "worse" results) than the strength judged by Pearson's or Spearman's correlation coefficients. Best agreement between the assessors was in the measurement of neck circumference (both ICC and Spearman's correlation coefficient described it as excellent). One can only speculate that the best degree of agreement was due to a relative simplicity and well-defined and unambiguous measurement method for the neck circumference. In contrast, the lowest level of agreement was in the measurement of neck movements. Maximal anteroflexion and retroflexion were examined with a goniometer. Despite thorough pre-study instructions, at the end of the study, both assessors felt and expressed uncertainty about the accuracy of the method. The assessment of neck mobility is more complex and relies on correct identification of anatomical landmarks and patient co-operation[8]. The most important difficulty was in determination of neutral head position from which both maximal anteroflexion and retroflexion were measured. This is a serious limitation of the goniometric measurements used.

The range of the ICC coefficient and Pearson's or Spearman's correlation coefficients lies between -1 and +1. All coefficients describe the relationship between the paired measurements made by two assessors; the correlation coefficients provide no information about systematic deviation (bias). Bias means that one assessor **consistently** measures or classifies the given parameter differently from the second one. The Pearson's or Spearman's correlation coefficients can be very high even in the presence of a high inter-observer bias.

Wilcoxon signed-rank test and paired Student's t-test, respectively, showed a significant bias in the measurements of all quantitative parameters between the assessors ($p<0.001$). The medians of differences (O - T) and upper quartiles for anteroflexion, retroflexion, anteroflexion + retroflexion, HMD, TMD, SMD and IIG were negative (see Table 4). This indicates that the measurements of these parameters performed by the assessor T were systematically higher than the measurements done by the assessor O (measurements of T were positively biased). In contrast, the medians of differences (O - T) and lower quartiles were positive for the neck circumference and mandibular length (see Table 4), in other words, that the measurements of observer T were lower than those of assessor O (measurements of T were negatively biased).

Systematic deviation can be demonstrated graphically on a simple scatter or a Bland-Altman plot. Fig. 1 and Fig. 2 show an example of these plots for one of the quantitative parameters (neck circumference) when bias was present.

Fifteen years ago, Karkouti et al. (ref.[11]) studied the inter-observer reproducibility between two experienced residents in ten tests for predicting DI (IIG, subluxation of the mandible, TMD, length of mandibular ramus, retrogenia, chin protrusion, atlanto-occipital movement and score, Mallampati test and MMT). Two tests – mouth opening (ICC=0.93) and chin protrusion (ICC=0.89) – had excellent inter-observer reproducibility. Classical Mallampati test had poor reproducibility (κ=0.31) and the remaining seven tests were moderately reproducible. Due to the different tests used for assessing inter-observer agreement, our results cannot be directly compared to the Karkouti et al.[11] study. However, when the four Grades of MMT (I, II, III and IV) were merged into two (I + II and III + IV), the kappa was comparable in both studies (Karkouti et al.[11] κ=0.49, our results κ=0.503).

Thyro-mental distance (Patil-Aldreti test[20,21]) is credited as a valuable tool in predicting DI. Unfortunately, data describing its **reproducibility** are contradictory. Hilditch et al.[10] found excellent agreement between a nurse and anesthesiologist in the measurement of TMD (ICC=0.85). We found poor to fair agreement depending on the coefficient used (ICC= -0.002, Pearson's correlation coefficient = 0.265). This is comparable to Merino García et al.[22] (κ<0.21) and Rosenstock et al.[8] (Spearman's correlation coefficient for specialists = 0.41 and residents = 0.48, respectively). In the same study, the inter-rater agreement for TMD was even worse in DI cases (Spearman's correlation coefficient for specialists = 0.23 and for residents = 0.21, respectively).

Our study has limitations. Two medical students at the end of their undergraduate period performed all assessments. They were given written instructions and underwent practical training in how to carry out the tests. However, they had only a short previous practice in anesthesia and airway management in the operating room. Although previous studies have shown **generally** variable agreement and inconsistency between assessors irrespective of their qualification (two specialists[8], two residents[8,11], specialist *vs.* resident[22], anesthetist *vs.* nurse[10]), the assessor's level of expertise may also influence the accuracy[8].

The only goal of the study was to compare the agreement of paired measurements of two assessors and not to determine the ability of the tests to predict DI. Young and healthy volunteers were enrolled and their participation in the study ended with completion of the measurements. They did not have subsequent anesthesia and hence the

ease of tracheal intubation could not be determined. For this reason, in spite of inter-observer bias in most parameters, we were not able to distinguish which measurement, if any, reflected the reality.

To be useful in clinical settings, a model for predicting DI should be simple and feasible, with high accuracy, sensitivity and positive predictive value to identify **all** patients in whom intubation will be difficult[5]. These criteria can only be met when the input data are correct and consistent. If not, the construction of a model predicting DI may become rather a mathematic entertainment for the particular rater than a valuable clinical tool.

CONCLUSION

Although performed under standardized conditions, not all tests for predicting DI achieved acceptable inter-observer reproducibility in our study. Best agreement was demonstrated for the assessment of neck circumference while the highest discrepancies between raters were in goniometrically-measured mobility of the C-spine (max. anteroflexion and retroflexion). The high inter-observer variability of examinations may be one reason why the models for predicting DI are not reliable in all cases.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from the Czech Ministry of Health Internal Grant Agency – project No. NS 9618–4/2008.

The authors express their sincere thanks to Prof. Jaroslav Vesely and Ms. Vera Langerova (Department of Pathological Physiology, Faculty of Medicine and Dentistry, Palacky University Olomouc, Czech Republic) for valuable advice and critical reviewing the manuscript.

REFERENCES

1. Lee A, Fan LT, Gin T, Karmakar MK, Ngan Kee WD. A systematic review (meta-analysis) of the accuracy of the Mallampati tests to predict the difficult airway. Anesth Analg 2006;102:1867-78.
2. Bindra A, Prabhakar H, Singh GP, Ali Z, Singhal V. Is the modified Mallampati test performed in supine position a reliable predictor of difficult tracheal intubation? J Anesth 2010;24:482-5.
3. Adamus M. Comment on the article by Bindra A et al. Is the modified Mallampati test performed in supine position a reliable predictor of difficult tracheal intubation? J Anesth 2011;25:135.
4. Law JA. Relying on just a few predictors of easy airway management may bite back! Anesth Analg 2008;106:668.
5. Naguib M, Scamman FL, O'Sullivan C, Aker J, Ross AF, Kosmach S, Ensor JE. Predictive performance of three multivariate difficult tracheal intubation models: a double-blind, case-controlled study. Anesth Analg 2006;102:818-24.
6. Wilson ME, Spiegelhalter D, Robertson JA, Lesser P. Predicting difficult intubation. Br J Anaesth 1988;61:211-6.
7. Arné J, Descoins P, Fusciardi J, Ingrand P, Ferrier B, Boudigues D, Ariès J. Preoperative assessment for difficult intubation in general and ENT surgery: predictive value of a clinical multivariate risk index. Br J Anaesth 1998;80:140-6.
8. Rosenstock C, Gillesberg I, Gätke MR, Levin D, Kristensen MS, Rasmussen LS. Inter-observer agreement of tests used for prediction of difficult laryngoscopy/tracheal intubation. Acta Anaesthesiol Scand 2005;49:1057-62.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;Feb 8;1(8476):307-10.
10. Hilditch WG, Kopka A, Crawford JM, Asbury AJ. Interobserver reliability between a nurse and anaesthetist of tests used for predicting difficult tracheal intubation. Anaesthesia 2004;59:881-4.
11. Karkouti K, Rose DK, Ferris LE, Wigglesworth DF, Meisami-Fard T, Lee H. Inter-observer reliability of ten tests used for predicting difficult tracheal intubation. Can J Anaesth 1996;43:554-9.
12. Uebersax JS. Kappa Coefficients. Cited [2011, April 11]. Available from: http://www.john-uebersax.com/stat/kappa.htm
13. Ben-David A. Comparison of classification accuracy using Cohen's Weighted Kappa. Expert Systems with Applications 2008;34:825-32.
14. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61:29-48.
15. Blood E, Spratt KF. Disagreement on Agreement: Two Alternative Agreement Coefficients. Statistics and Data Analysis. SAS Global Forum 2007. Paper 186–2007. Cited [2011, April 11, 2011]. Available from: http://www2.sas.com/proceedings/forum2007/186-2007.pdf
16. Samsoon GL, Young JR. Difficult tracheal intubation: a retrospective study. Anaesthesia 1987;42:487-90.
17. Khan ZH, Kashfi A, Ebrahimkhani E. A comparison of the upper lip bite test (a simple new technique) with modified Mallampati classification in predicting difficulty in endotracheal intubation: a prospective blinded study. Anesth Analg 2003;96:595-9.
18. Hester CE, Dietrich SA, White SW. A comparison of preoperative airway assessment techniques: the modified Mallampati and the upper lip bite test. AANA J 2007;75:177-82.
19. Myneni N, O'Leary AM, Sandison M, Roberts K. Evaluation of the upper lip bite test in predicting difficult laryngoscopy. J Clin Anesth 2010;22:174-8.
20. Orozco-Díaz E, Alvarez-Ríos JJ, Arceo-Díaz JL, Ornelas-Aguirre JM. Predictive factors of difficult airway with known assessment scales. Cir Cir 2010;78:393-9.
21. Torres K, Patel AA, Styliński K, Błoński M, Torres A, Staśkiewicz G, Maciejewski R, Wojtaszek M. The body constitution of patients and intubation scales as predictors of difficult intubation considered in relation to the experience of the intubator. Folia Morphol (Warsz). 2008;67:171-4.
22. Merino García M, Marcos Vidal JM, García Pelaz R, Díez Burón F, España Fuente L, Bermejo González JC. Evaluation of a protocol for predicting difficult airway in routine practice: interobserver agreement. Rev Esp Anestesiol Reanim 2010;57:473-8.