Accurate Classification of Difficult Intubation by Computerized Facial Analysis

Christopher W. Connor, MD, PhD,* and Scott Segal, MD, MHCM†

BACKGROUND: Bedside airway evaluation is conduced before anesthesia, but all current methods perform modestly, with low sensitivity and positive predictive value. We hypothesized that subjective features of patients' anatomies improve anesthesiologists' ability to predict difficult intubation, and derived a computer model to do so, based on analysis of photographs of patients' faces.

METHODS: Eighty male patients were divided into 2 equal cohorts for model derivation and validation. Each cohort consisted of 20 easy and 20 challenging intubations, defined as >1 attempt by an operator with at least 12 months of anesthesia experience, grade 3 or 4 laryngoscopic view, need for a second operator, or nonelective use of an alternative airway device. Photographs of each subject's face were analyzed by software that resolves each face into 61 facial proportions derived from an algorithm that models the face as a single point in a 50-dimensional eigenspace. Each parameter was tested for discriminatory ability by logistic regression, and combinations of 11 variables with $P \leq 0.1$, plus Mallampati class and thyromental distance, were tested exhaustively by all possible binomial quadratic logistic regression models. Candidate models were cross-validated by maximizing the product of the area under the receiver operating characteristic curves obtained in the derivation and validation cohorts.

RESULTS: The best model included 3 facial parameters and thyromental distance. It correctly classified 70 of 80 subjects ($P < 10^{-8}$). In contrast, the best combination of Mallampati class and thyromental distance correctly classified 47 of 80 ($P = 0.073$). Sensitivity, specificity, and area under the curve for the computer model were 90%, 85%, and 0.899, respectively.

CONCLUSIONS: Computerized analysis of facial structure and thyromental distance can classify easy versus difficult intubation with accuracy significantly outperforming popular clinical predictive tests. (Anesth Analg 2011;112:84–93)

All patients undergoing preoperative evaluation are assessed for anatomic features that might predict difficulty in performing endotracheal intubation under general anesthesia. Typically, at least 2 examinations are used: the Mallampati (MP) test[1,2] is performed and the thyromental distance (TMD)[3] is measured. The MP test involves an examination of oropharyngeal structures that are visible when the seated patient maximally opens the mouth and extends the tongue without phonation. The TMD is a measure of the space between the superior tip of the thyroid cartilage and the inside of the tip of the mandible. Both tests perform only modestly, with sensitivity of 30% to 60%, specificity of 60% to 80%, and positive predictive value of only 5% to 20%.[4] Even so, the combination of MP test and TMD performed better than any other bedside screening test in a meta-analysis of 35 trials studying >50,000 subjects.[4] In practice, anesthesiologists likely consider other subjective factors in anticipating a difficult airway, including habitus, facial appearance, and perhaps other poorly understood hunches. It is our belief that this gestalt may outperform conventional airway examinations.

In this study, we attempted to derive a computer model that similarly classifies the ease or difficulty of endotracheal intubation, from analysis of facial structure based on 3 photographs. The computer model was derived and validated against cohorts of patients with known airway anatomy, identified at surgery to be either easy or difficult to intubate. We hypothesized that this may allow an improved airway examination tool to be derived.

METHODS

Recruitment of subjects at Brigham and Women's Hospital was performed in accordance with a protocol approved by Partners Healthcare Human Research Committee. The protocol was noninvasive, requiring only a customary airway examination, review of the anesthetic record, and photography of the head and neck of the patient. Because the protocol contained no risk of harm to the patient, approval for recruitment by solely verbal consent was obtained from the IRB. To limit any potential confounding effects of gender and racial group in this initial study, only male Caucasians were recruited.

Patients were defined as easy to intubate if their anesthetic record described a single attempt with a Macintosh 3 blade resulting in a grade 1 laryngoscopic view (full exposure of the vocal cords).[1,5] Difficult intubation was defined by at least 1 of the following: >1 attempt by an operator with at least 1 year of anesthesia experience, grade 3 or 4 laryngoscopic view on a 4-point scale,[5] need for a second operator, or nonelective use of an alternative airway

From the Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Boston, Massachusetts.

Accepted for publication September 24, 2010.

Supported by departmental funds.

Current affiliations: From the *Department of Anesthesiology, Boston Medical Center; and †Department of Anesthesiology, Tufts University School of Medicine, Boston, MA.

The authors declare no conflicts of interest.

Address correspondence and reprint requests to Christopher W. Connor, MD, PhD, Department of Anesthesiology, Boston Medical Center, Newton Pavilion, 88 East Newton St., Room 2817, Boston, MA 02118. Address e-mail to chris.connor@bmc.org.

DOI: 10.1213/ANE.0b013e31820098d6

device such as a bougie, fiberoptic bronchoscope, or intubating laryngeal mask airway.[6,7] We also calculated a modified intubation difficulty scale (IDS),[8] with the assumption that subjective force applied during intubation was increased when >1 attempt at direct laryngoscopy was performed.

Patients meeting our entry criteria were identified by examination of their anesthesia records in the postanesthesia care unit. Suitable patients were recruited postoperatively when adequately recovered from the effects of anesthesia. Photographs were obtained either in the postanesthesia care unit or on the ward on the first postoperative day. Patients who had undergone head or neck surgery were excluded. Patients in whom central venous catheters or other interventions prevented full view of the features of the face in frontal and profile views were excluded. Patients who were neither easy nor difficult to intubate by our criteria were not recruited. All patients meeting entry criteria were recruited until the recruitment goals were reached. Patients were informed of their right to not participate in the study, but none refused.

Data Acquisition

Eighty patients were recruited by cohorts. Forty patients were used as the model derivation set and the other 40 patients were used as the model validation set. Each set was composed of 20 easy to intubate and 20 difficult to intubate subjects. Digital photographs of the head and neck of each patient in frontal view and in left and right profiles were obtained. Patient demographics (height, weight, age, gender, type of surgery), MP class, TMD (in fingerbreadths), and the details regarding ease of intubation were obtained from the anesthetic record. Any data found to be absent from the record were collected by the authors at the time of patient enrollment. Preoperative assessment of the patients was performed in a Preoperative Anesthesia Testing Clinic staffed by a small cadre of specially trained preanesthetic nurse practitioners using a structured electronic medical record and under the direct supervision of an experienced attending anesthesiologist. Because the clinical assessments of MP test and TMD were performed before surgery to determine the ease or difficulty of intubation, the clinic staff was blinded to the ultimate cohort assignation at the time of assessment. TMD was measured in fingerbreadths with the head in a neutral position, as is the usual clinical practice at our institution and elsewhere.[9]

The photographs were analyzed by facial structure analysis software (FaceGen Modeller v3.3; Singular Inversions, Toronto, Canada). This software uses an algorithm to generate a mathematical model of the face based on a weighted contribution of predetermined "eigenfaces." An example of a completed model is shown in Figure 1 (image of the first author). The eigenface method allows the structure of any particular individual face to be expressed in an elegant and compact form.[10] Each of the weighting values for the eigenfaces can be considered to be a coordinate value, allowing the whole physiognomy of an individual face to be represented solely as a point in a 50-dimensional space.[11] The facial analysis software implements a further improvement to the eigenface method, such that the weightings of the eigenfaces can be specified in terms of descriptive (but not directly measurable or observable) facial proportions[12,13] expressed as standard deviations from an androgynous normal (Fig. 2) derived from a reference population of 300 individuals.[14] Table 1 shows the 61 descriptive facial proportions used. Some of these 61 indices are interdependent but it can be demonstrated that this higher-dimensional model is related directly to the underlying 50 eigenfaces by a straightforward linear transformation.[14]

Statistical Analysis

Model Derivation

The task of recognizing the difficult intubation can be conceived as the task of deriving an algorithm capable of separating the points representing the cohort of easy patients from those representing the cohort of difficult patients within a defined variable space. Patients who were easy to intubate were assigned a classification value of zero, difficult to intubate patients a value of 1. All 61 variables and the physical properties of MP test and TMD were subjected to variable reduction by univariate analysis at $P \leq 0.1$ (χ^2 distribution, G^2 goodness-of-fit statistic).[15] Those variables that individually showed a statistical trend in discrimination between easy and difficult airways are shown with an asterisk in Table 1 and were used as a subset to derive the predictive model. Binomial logistic regression to segregate the easy to intubate and difficult to intubate cohorts[16] was performed exhaustively on all possible variable combinations of this reduced subset of variables using a quadratic logit.[17]

The mathematical theory underlying the choice of the quadratic logit function is described in Appendix A. Briefly, the quadratic form of the logit uses both the value of an input variable and its square. We chose this model because we hypothesized that factors influencing difficulty of intubation may not behave linearly, but instead be either easier or harder on both sides of a central value. As an example, one could hypothesize that if both the small jaw length of micrognathia and the large jaw length of acromegaly suggest difficulty, then ease of intubation may be optimum at some middle value and fall away on either side. Furthermore, use of the quadratic logit does not exclude that a linear relationship may be found, because the fitting of the quadratic logit may produce a curve that is locally straight over the region of interest. Hence, using only data taken from the 40 patients in the model derivation cohorts, each variable and its square were used as inputs to a logit function in all possible combinations of inclusion or noninclusion. The coefficients of these logits were optimized to produce candidate models. The area under the curve (AUC) of the receiver operating characteristic (ROC) of each candidate model was calculated[18] and stored as $\mathrm{AUC}_{\mathrm{derivation}}$ for that candidate model. We selected models optimized for AUC rather than raw accuracy because this technique is more robust with regard to future data.[19]

Model Validation

Model validation was then performed to avoid choosing a candidate model that is overfitted to the derivation dataset.

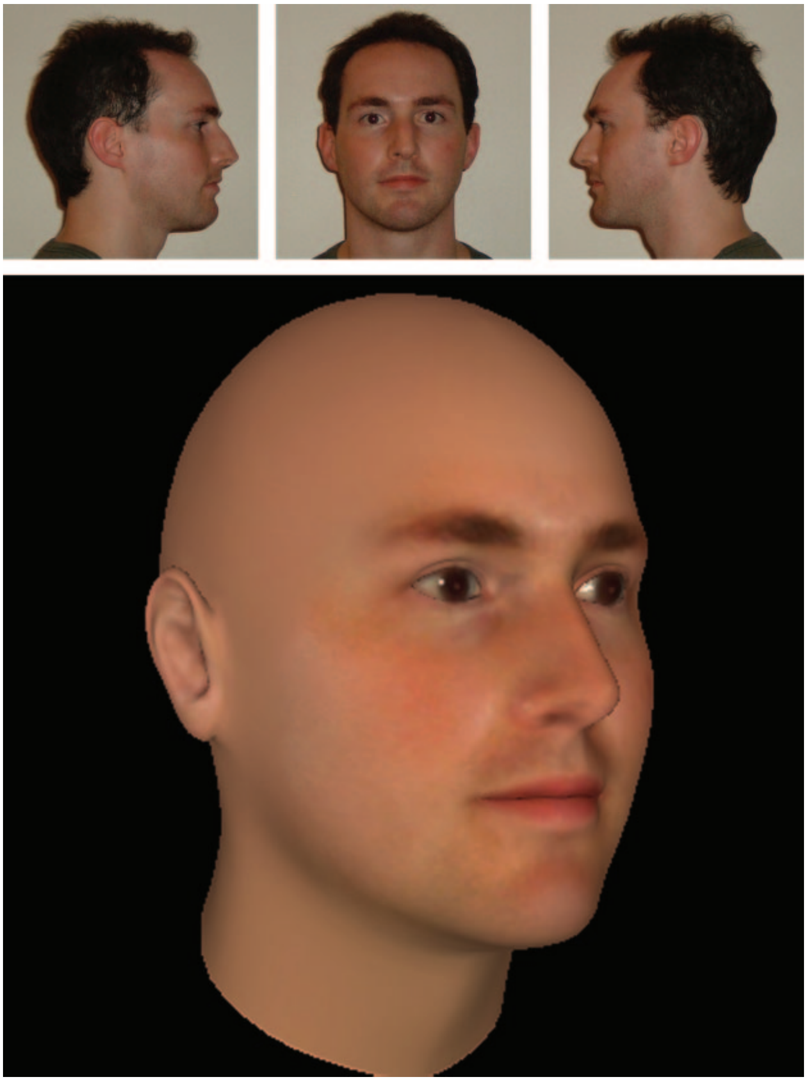

Figure 1. Computer reconstruction of the head from profile and face-on photographs (image of first author).

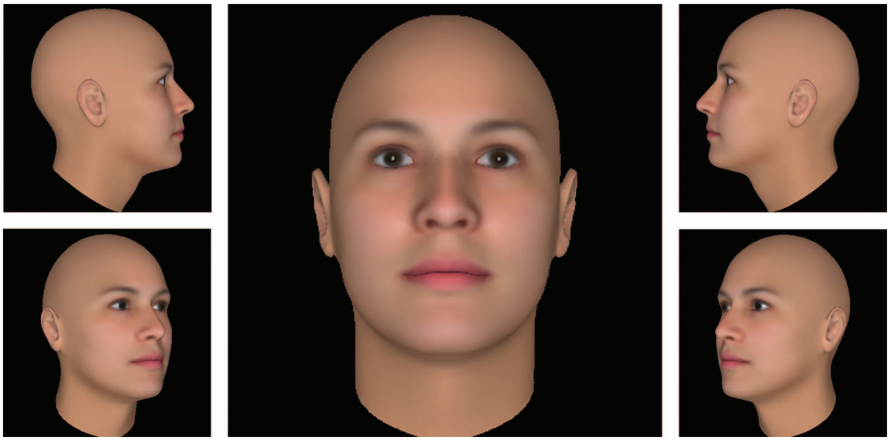

Figure 2. Appearance of the average head of the reference population.

The candidate models produced in the model derivation stage were applied without further adjustment to the data taken from the 40 patients in the model validation cohorts, producing an $\text{AUC}_{\text{validation}}$ for each candidate model.

Final model selection was performed by choosing the candidate model with the maximum value of $\text{AUC}_{\text{derivation}} \times \text{AUC}_{\text{validation}}$. This method excludes models that show evidence of overfitting, a problem with large variable spaces. The mathematics underlying the rationale of AUC product maximization are described in Appendix B. Briefly, using the observation that any overfitting to gain performance in one subset will lead to a comparable loss of performance in a

Table 1. The 61 Variables Defining Photographic Reconstruction of the Head

Descriptive facial proportions

Brow ridge: high/low
Brow ridge inner: down/up
Brow ridge outer: up/down*
Cheekbones: low/high
Cheekbones: shallow/pronounced
Cheekbones: thin/wide
Cheeks: concave/convex
Cheeks: round/gaunt
Chin: forward/backward
Chin: pronounced/recessed
Chin: retracted/jutting
Chin: shallow/deep
Chin: small/large
Chin: tall/short
Chin: wide/thin*
Eyes: down/up*
Eyes: small/large
Eyes: tilt inward/outward*
Eyes: apart/together
Face: brow-nose-chin ratio*[a]
Face: forehead-sellion-nose ratio
Face: heavy/light
Face: round/gaunt
Face: tall/short
Face: up/down
Face: wide/thin
Forehead: small/large
Forehead: tall/short
Forehead: tilt forward/back
Head: thin/wide

Jaw: retracted/jutting
Jaw: wide/thin
Jaw: neck slope high/low*[a]
Jawline: concave/convex
Mouth: drawn/pursed
Mouth: happy/sad
Mouth: lips deflated/inflated
Mouth: lips large/small
Mouth: lips puckered/retracted
Mouth: lips thin/thick
Mouth: protruding/retracted
Mouth: tilt up/down
Mouth: underbite/overbite
Mouth: up/down*
Mouth: wide/thin
Mouth: chin distance, short/long*
Nose: bridge shallow/deep
Nose: bridge short/long
Nose: down/up
Nose: flat/pointed
Nose: nostril tilt down/up*
Nose: nostrils small/large
Nose: nostrils wide/thin
Nose: region concave/convex
Nose: sellion down/up
Nose: sellion shallow/deep (1)
Nose: sellion shallow/deep (2)
Nose: sellion thin/wide
Nose: short/long
Nose: tilt down/up*[a]
Temples: thin/wide

Thyromental distance*[a]
Mallampati class

The thyromental distance and Mallampati class are included in the table as 2 further variables that were used for modeling.

* Demonstrated at least a statistical trend ($P \leq 0.1$) with identified difficult intubation.

[a] Appear in the final model.

subsequent subset, product maximization creates a measure in which any apparent overfitting is penalized.

The sensitivity and specificity of the selected model were calculated. The *P* value of the model was determined by its classification accuracy (exact binomial distribution) and the Bonferroni correction for multiple tests was applied (MATLAB; MathWorks, Natick, MA).

Facial Structure Test-Retest Validation

The facial structure analysis software requires some user interaction to place certain fiducial markers on the images to guide reconstruction. Ten patients were selected at random from the 80 study patients and their photographs were rerendered into 3-dimensional models to test the reproducibility of the reconstruction process.

RESULTS

Of the 61 descriptive facial proportions and the physical properties of MP test and TMD, 11 showed a univariate statistical trend in discriminating between easy and difficult intubations. These 11 variables, shown with an asterisk in Table 1, were evaluated as possible inputs to the model, producing a total of $2^{11} - 1 = 2047$ candidate models. The final model was chosen by AUC product maximization and found to depend on only 3 facial proportions plus TMD, as marked with a superscript "a" in Table 1. The parameters of the model are stated numerically in Appendix C. Relative to the population normal shown in Figure 2, the variations in facial appearance described by the 3 facial proportions used in the airway algorithm are shown in Figure 3.

Clinical Interpretation of the Model

Figure 4 shows the classification and statistical behavior of the algorithm when applied to the model derivation dataset, the validation dataset, and the 2 datasets combined. The algorithm successfully clusters the easy and difficult airways toward opposite ends of the logit curve (Fig. 4, A, C, and E). ROC curves were constructed for each test

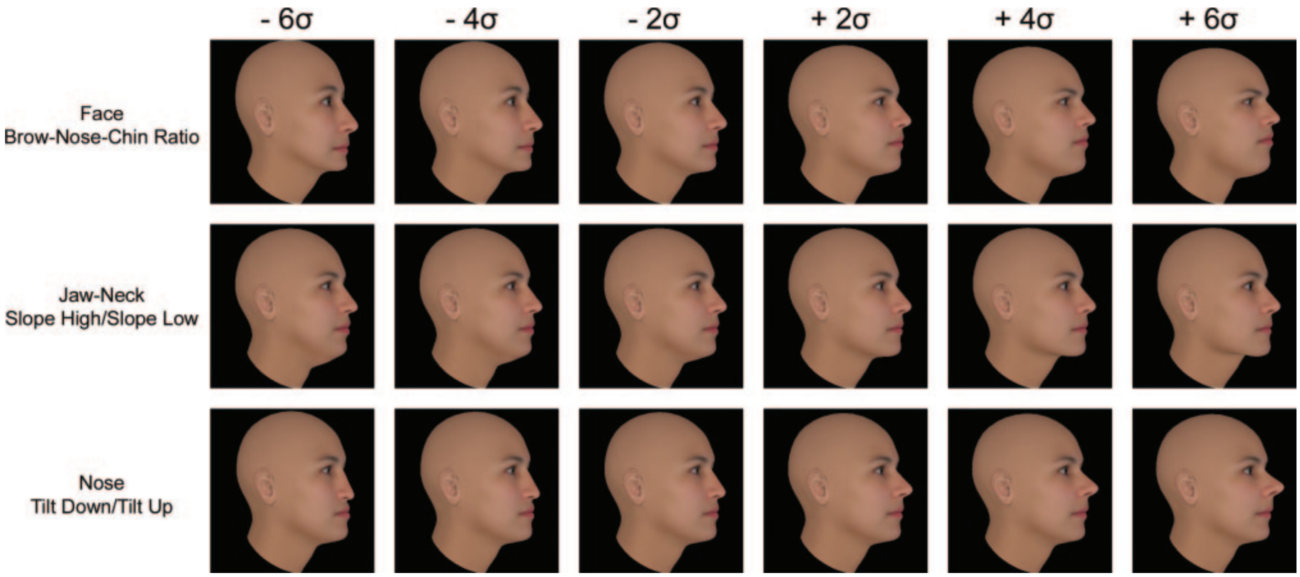

Figure 3. Variations in facial appearance from the average head shown in Figure 2 by standard deviations of the descriptive facial proportions used in the airway algorithm. σ is the standard deviation from the normal head derived from 300 individuals.

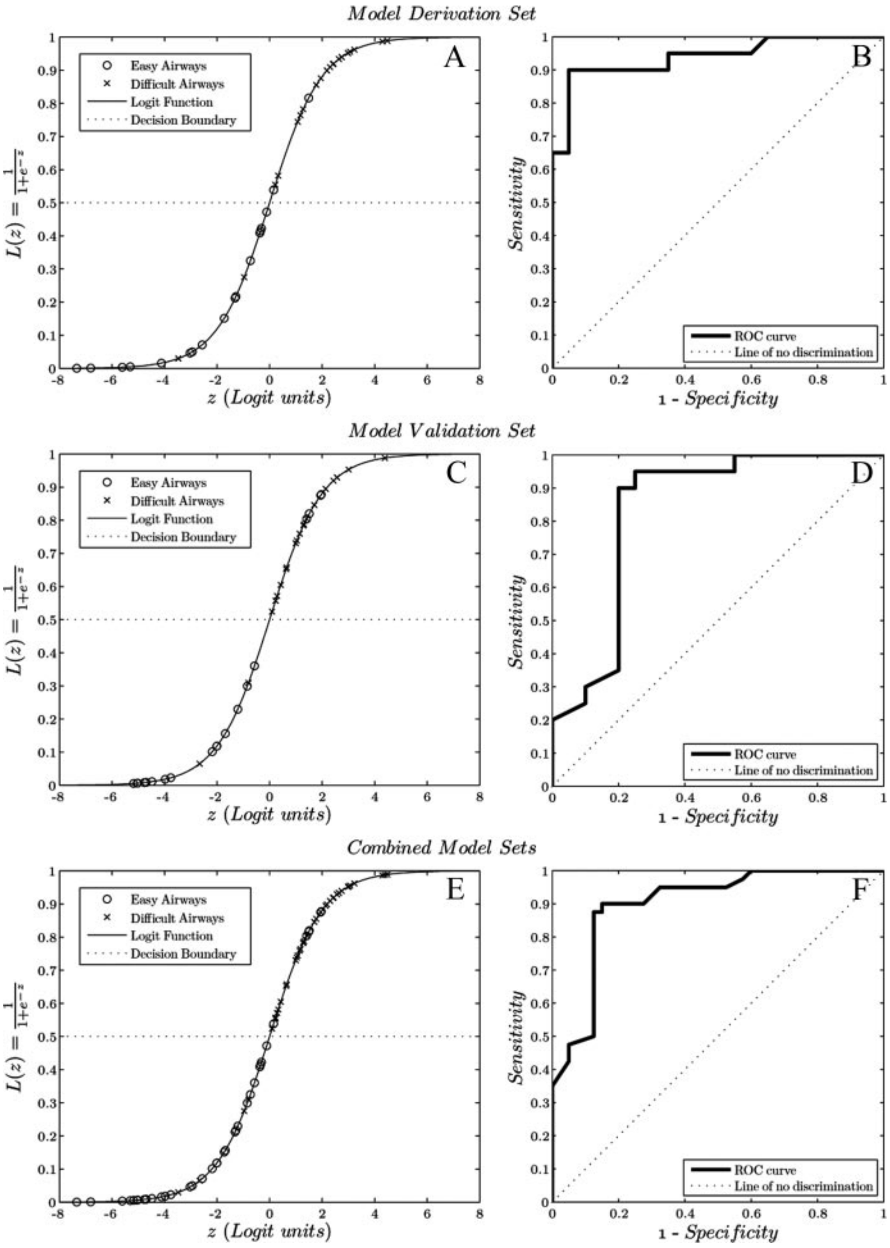

Figure 4. Classification and statistical behavior of the airway algorithm. Panels A, C, and E show the logit function and calculated values [L(z)] for the studied faces in the derivation, validation, and combined cohorts, respectively. Panels B, D, and F show the associated receiver operating characteristic (ROC) curves for each cohort.

population and are shown adjacent to the classification behavior (Fig. 4, B, D, and F). Table 2 contains the numerical representations of these statistical properties. The performance of the algorithm is stated in terms of its accuracy as a binomial classifier, allowing calculation of *P* values according to the binomial distribution. In selecting a model based on the data, the problem of multiple comparisons must be addressed. Because 11 variables participated in generation of the model, and all possible combinations of models were exhaustively evaluated, each variable participated in 2^{10} models. Application of the Bonferroni correction for these multiple comparisons still yielded highly significant *P* values (Table 2). When applied to the combined dataset, the performance of the airway classification algorithm showed a sensitivity of 90% and a specificity of 85%. The area under the ROC curve (AUC) was 89.9%. When applied to the combination of all easy patients and only those difficult patients with an IDS score[8] >5, the model correctly classified 89%, with a sensitivity of 96%, specificity 85%, and AUC 90.2%.

The MP test did not show a statistical trend with ease or difficulty of intubation and so did not form part of the described process of model selection. Even when the MP test was explicitly forced into the reduced set of variables, it did not affect the final model selection and so inclusion of MP test did not add further predictive information.

To allow comparison of the model to classical airway assessment tools, the ability of the MP test and TMD to classify difficult intubation was tested against the study population. Table 3A shows the statistical performance of the MP test and TMD both alone and together when used as variables for the quadratic logit model. This analysis ascertained the maximum performance of these tests when their thresholds are allowed to be optimized against the

Table 2. Statistical Performance Measures for the Airway Algorithm

Statistical properties	Derivation set	Validation set	Combined[a]	Combined easy and IDS score >5[b]
Sensitivity	0.9	0.9	0.9	0.96
Specificity	0.9	0.8	0.85	0.85
True positives	18	18	36	24
True negatives	18	16	34	34
False positives	2	4	6	6
False negatives	2	2	4	1
Accuracy (correct/total)	36/40	34/40	70/80	58/65
Exact binomial probability test	$P = 9.29 \times 10^{-7}$	$P = 4.18 \times 10^{-6}$	$P = 1.58 \times 10^{-12}$	$P = 2.14 \times 10^{-11}$
Bonferroni correction	2^{10}	2^{10}	2^{10}	2^{10}
Corrected probability	$P = 9.51 \times 10^{-4}$	$P = 4.28 \times 10^{-3}$	$P = 1.62 \times 10^{-9}$	$P = 2.19 \times 10^{-8}$

IDS = intubation difficulty scale.

[a] The combination of both the derivation set and validation set.

[b] The set of all easy patients with all difficult patients for whom an IDS score >5 was estimated ($n = 65$).

Table 3A. Statistical Performance Measures of Classical Airway Assessment Tools When Optimized with Respect to the Study Model Derivation Population

	Mallampati test		Thyromental distance		Bivariate model (MP and TMD)	
Statistical properties	Derivation	Validation	Derivation	Validation	Derivation	Validation
Sensitivity	0.4	0.1	0.85	0.8	0.8	0.65
Specificity	0.85	0.9	0.5	0.5	0.7	0.7
True positives	8	2	17	16	16	13
True negatives	17	18	10	10	14	14
False positives	3	2	10	10	6	6
False negatives	12	18	3	4	4	7
Accuracy	25/40	20/40	27/40	26/40	30/40	27/40

Table 3B. Performance of Classical Airway Assessment Tools Alone and in Combination When Used with Their Frequently Ascribed Thresholds

Statistical properties	MP score ≥3	TMD <3	MP score ≥3 and TMD <3	MP score ≥3 or TMD <3
Sensitivity	0.25	0.125	0.05	0.325
Specificity	0.875	0.925	0.95	0.85
True positives	10	5	2	13
True negatives	35	37	38	34
False positives	5	3	2	6
False negatives	30	35	38	27
Accuracy	45/80	42/80	40/80	47/80

MP = Mallampati; TMD = thyromental distance.

model derivation data in the same manner used in the derivation of the new airway model. The classical tools nevertheless demonstrate substantially weaker performance than the model (57 of 80 correctly classified, Fisher exact test, $P = 0.018$ for difference from computer model). Table 3B shows the performance of the MP test and TMD when used in the usual clinical manner with their classically ascribed thresholds, without the inclusion of a squared term. The performance here is again inferior and the greatest achieved accuracy of 47 of 80 did not rise to the level of statistical significance when compared with chance ($P = 0.073$, exact binomial distribution) and was inferior to the computer model ($P < 0.0001$, Fisher exact test).

Because this new airway model describes appearance, it is possible to generate pictures of faces that would appear to have certain degrees of ease or difficulty of intubation. Figure 5A illustrates the head that is theoretically most difficult to intubate according to the model. Figure 5B represents a head that the model would classify as easy to intubate. The parameter values for this head are set such that the value produced by the model is of the same magnitude but opposite to Figure 5A. Figure 5B might therefore be considered to represent a patient as easy to intubate as the patient in Figure 5A would be difficult.

Mathematical Interpretation of the Model

A mathematical interpretation of the model is given in Appendix C, which gives a more precise interpretation of the meaning of the various parameters and the resulting logit calculated by the model.

Facial Structure Test-Retest Validation

A correlation coefficient of $r = 0.80$ was established across the 610 variables, indicating a high degree of reproducibility. The classifications (predicted easy versus difficult) of these 10 patients by the algorithm described in Appendix C were unchanged by rerendering.

DISCUSSION

In our study, computerized facial structure analysis combined with a widely used bedside airway evaluation method yielded a model that significantly outperformed

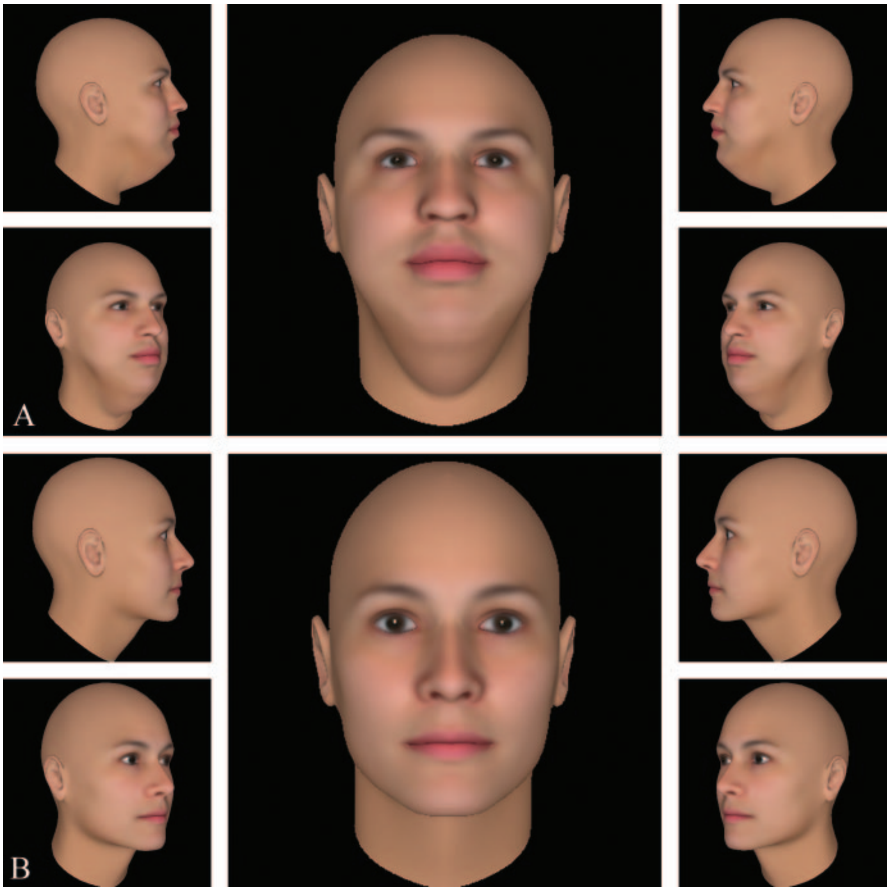

Figure 5. A, Appearance of the face rated most difficult to intubate by the algorithm, defined by x = [2.995, −13.683, 0.557, 2.032], where the values of the vector refer, respectively, to face (brow-nose-chin ratio), jaw-neck (slope high/slope low), nose (tilt down/tilt up), and thyromental distance (fingerbreadths). B, Appearance of a face rated easy to intubate. The ease is comparable in magnitude to the difficulty associated with Figure 4A, defined by x = [−1.06, 4.85, −0.20, 4], where the values of the vector refer, respectively, to face (brow-nose-chin ratio), jaw-neck (slope high/slope low), nose (tilt down/tilt up), and thyromental distance (fingerbreadths).

popular clinical predictive tests. Our model accurately classified 70 of 80 airways compared with 47 of 80 for MP test plus TMD using classical thresholds.[1,3,5]

Use of a bedside examination to predict difficult intubation is considered the standard of care in modern anesthesiology practice. It has been incorporated into the difficult airway algorithm of not only the American Society of Anesthesiologists[7] and those of several other countries,[20] but also most recently into the World Health Organization Surgical Safety Checklist,[21] the use of which is being encouraged in every operating room in the world. Unfortunately, all easily performed examination systems in clinical practice perform only modestly, with sensitivities of 20% to 62%, specificities of 82% to 97%, and very low positive predictive values, generally <30%, unless very liberal definitions of difficulty are used.[22] There are likely a number of reasons for this poor performance, including the relative rarity of difficult intubation,[22] the multifactorial etiology and varying definition of difficult intubation, interobserver variability in test results,[23,24] failure to validate potential systems in patients independent of those used to derive the test,[22] and the inadequacy of the tests themselves. Conversely, experienced anesthesiologists almost certainly use cues other than those derived from formal bedside tests to formulate their clinical impression of the ease of intubating any given patient. There may be a large number of anatomic factors that enter into such a judgment.[25] However, bedside scores based on such factors have not proven to be accurate.[4] Indeed, getting anesthesiologists to pay attention to the airway may be the principal benefit of routinely performing airway examinations before induction.[22]

Our study differs from previous work using facial imaging to evaluate the airway. Suzuki et al.[26] used digital photographs of subjects' faces to calculate 5 ratios and angles from measurements derived from placement of anatomic markers on the photographs. They found one, the "submandibular angle," to be correlated with difficult tracheal intubation. They also used morphing software to construct "average" easy and difficult to intubate faces, which we believe bear some subjective resemblance to our Figure 5. Similarly, Naguib et al.[27] measured 22 indices from plain radiographs and 8 from 3-dimensional computed tomographic scans of the head in patients who were easy or difficult to intubate. They constructed a model containing 3 bedside tests (MP test, TMD, and thyrosternal distance) and 2 radiographic features that accurately separated the easy and difficult cohorts with an AUC of the ROC curve of 0.97. Both of these previous investigations, however, used a priori assumptions of which anatomic features might relate to difficult laryngoscopy and intubation. Both also required actual measurement of anatomic features. In contrast, our method modeled the entire physiognomy of the face with no such assumptions and no direct measurements. Moreover, the method does not require time-consuming and potentially dangerous radiographs. If implemented on high-speed computers, perhaps accessed by end users transmitting images of patients over a network, our model could be used for rapid bedside or field assessment of the airway, even by inexperienced practitioners.

Our study has several limitations. First, it is likely that there are causes of difficult intubation not included in our study cohorts. For example, some patients with limited neck mobility but otherwise normal airways are difficult to intubate.[28] Further refinement of the model could include subjective or measured indices of neck extension, or indeed other predictors of difficult intubation such as body mass index. Second, we measured TMD in the neutral position rather than full neck extension, and we used fingerbreadths rather than measured distance. This is the method in routine use in our institution but some evidence suggests measuring TMD in extension is more predictive of intubation difficulty[29] and, although popular, fingerbreadth measurements are inferior to ruler measurements.[9] Conversely, if the model included this potentially inferior measurement, refining it by using full-extension TMD could only improve performance of the model. Third, we sought to eliminate potential racial or gender-based confounding by confining our sample to Caucasian males. Finally, because photographs were obtained postoperatively, we cannot entirely exclude the possibility of changes in facial appearance caused by anesthesia and surgery. Only a large, prospective study in a diverse patient population would be able to verify the performance of our model in general clinical use. It is encouraging that the model predicts ease in the computerized normal face, and that it performs better than bedside tests within the study cohorts. It is also possible that deriving and validating it within a larger fraction of the difficult airway "space" could further refine the model.

Another potential limitation is the method used to categorize the subjects as easy or difficult to intubate. We used a liberal definition of difficult intubation, which causes the positive predictive value to increase. In real-world clinical use, anesthesiologists are likely more interested in very difficult intubation, and the positive predictive value will be lower, as it is for all difficult airway predictive methods. Conversely, use of a liberal definition makes the statistical task of separating the easy and difficult cohorts more difficult, not less so,[22] because the 2 groups of patients are more similar. This makes the strong performance of our model notable. Moreover, the model performed even better in the subset of patients with an IDS score >5, who had comparatively more difficult intubations. However, it is decidedly problematic to infer the comparative difficulty of an intubation from the after-the-fact description of the technical maneuvers required to manage that airway.[6] For example, would an intubation achieved over a bougie on the third attempt be considered more or less difficult than one in which the anesthesiologist decided to use a video technique after the first unsuccessful attempt? Even presuming equally experienced laryngoscopists, the comparison is confounded by differences in comfort with, and availability of, other adjunct techniques. This ambiguity also complicates the use of research tools such as IDS score, which, for example, could be low (and thus descriptive of an easy intubation) in a patient in whom a single direct laryngoscopy produced a poor view and who was then intubated fiberoptically.

The clinical utility of our methodology and model remains an important research question. First, technical issues would need to be solved. The software currently requires approximately 15 minutes to model each face from digital photographs. It relies presently on a relatively inefficient iterative algorithm to do so, and exerts considerable computing power on modeling the coloration and texture of the skin. Certainly, a more efficient one could be written, particularly if only a few parameters need to be derived to predict difficult intubation. Indeed, we have a prototype algorithm that can analyze a face, derive the relevant parameters, and calculate the intubation prediction in less than 1 minute (data not shown). If proven practical for widespread clinical use, this would represent a significant advance over previously published methods involving offline measurements taken from radiographs or photographs. Second, the computing power required is modest but exceeds that of current handheld devices. We envision that clinical use of our model would be most efficiently deployed using high-speed computers accessible to clinicians over a network, perhaps using handheld computers or smartphones incorporating digital cameras as input devices. Third, the requirement for manual placement of fiducial markers to guide reconstruction is a potential source of user error. However, it is encouraging that our test-retest results revealed no cases in which the overall judgment of ease or difficulty of intubation varied. Finally, the performance of the model should be compared with that of experienced clinicians given similar data. The model would be particularly useful if it could predict difficult intubation when an experienced clinician had not suspected it, a more dangerous clinical situation than the converse error of judgment. If the model outperforms human experts, then its applicability would potentially be quite broad and would include even seasoned anesthesiologists. Conversely, if human operators can match the model's performance, then the software may be of greater utility to nonairway experts. This assessment is an area of active research by our group.

In summary, the model presented herein significantly outperformed the current standard of the combination of MP and TMD examinations, and is based on quantification of facial anatomy performed by an unbiased computer algorithm. Additional work should define the ability of experienced clinicians presented with similar photographs and bedside airway examination results, and the ability of the computer model to prospectively predict difficult intubation in a large and diverse patient population. If the superiority of the method can be confirmed, the model could represent an important advance in the assessment of the airway.

Appendix A: A Real-World Representation of the Quadratic Logit

The quadratic logit function admits both the value of a parameter and its square:

$$\mathrm{L}(z) = \frac{1}{1+e^{-z}} \text{ where } z = b_0 + b_1x_1 + b_2x_1^2 + b_3x_2 + b_4x_2^2 + \ldots \quad \text{(A1)}$$

Completing the squares, we can equivalently write:

$$z = \beta_0 - \beta_1\,(x_1 - \alpha_1)^2 - \beta_2(x_2 - \alpha_2)^2 - \ldots \quad \text{(A2)}$$

And hence:

$$z = \beta_0 - \zeta_1\frac{(x_1-\alpha_1)^2}{2\sigma_1^2} - \zeta_2\frac{(x_2-\alpha_2)^2}{2\sigma_2^2} - \ldots \quad \text{(A3)}$$

where ζ_i takes only the values ±1, without which σ_i would be imaginary if $\beta_i < 0$.

Note that in equation (A1), the term z becomes the exponent. It is useful to recall the probability distribution function of the normal distribution:

$$\phi_{\mu_i\sigma_i^2}(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{(A4)}$$

The similarity between the 2 equations suggests that an alternative, probabilistic, interpretation of the quadratic logit function is possible. It can be shown that the output of the logit function L(z) can be written as a weighting of normal probability distributions of the inputs x_i:

$$\mathrm{L}(z) = \frac{\prod\limits_{\varsigma_i=+1}\phi_{\mu_i,\sigma_i^2}(x_i)}{\prod\limits_{\varsigma_i=+1}\phi_{\mu_i\sigma_i^2}(x_i) + C\cdot\prod\limits_{\varsigma_i=-1}\phi_{\mu_i,\sigma_i^2}(x_i)} \quad \text{(A5)}$$

where C is a numerical constant determined by the model, defined as:

$$\mathrm{C} = \frac{\prod\limits_{\varsigma_i=-1}(\sigma_i\sqrt{2\pi})}{\prod\limits_{\varsigma_i=+1}(\sigma_i\sqrt{2\pi})}e^{-\beta_0} \quad \text{(A6)}$$

Equation (A5) is a curious result, because we believe it mathematically describes a mental process similar to that performed by anesthesiologists given the task of assessing an unknown airway. The combination of factors favorable to intubation are weighed against the countervailing unfavorable factors and, based on their relative preponderance, a decision is reached with some greater or lesser degree of confidence. The apparent artificiality of the quadratic logit model thus leads to a surprisingly natural result.

Appendix B: Cross-Validation by Product of Area Under the Curve

We measure the performance of a model by the area under the curve (AUC) on its receiver operating characteristic plot. A cross-validation and model selection technique is required that can relatively suppress models that show evidence of overfitting.

Let us suppose that some ideal, optimal model exists, and that this model has an AUC of $\mathrm{AUC}_{\mathrm{ideal}}$. It will likely be possible to produce some other model that seems to perform better on the model derivation set. However, because this model is by definition not the ideal model, its performance must have been artificially improved by overfitting. Let us define the AUC of this model as $\mathrm{AUC}_{\mathrm{ideal}} + \varepsilon_0$, where ε_0 represents the performance erroneously obtained through overfitting.

Now, consider the model validation dataset. We would expect that the ideal model would have an AUC of $\mathrm{AUC}_{\mathrm{ideal}}$ when tested against either dataset. If the datasets are sufficiently large, then we can expect that any improvement in performance that was erroneously obtained by overfitting in the derivation set will appear as an equal penalty in the validation set. Therefore, the model with an AUC of $\mathrm{AUC}_{\mathrm{ideal}} + \varepsilon_0$ in the derivation set should have an AUC of approximately $\mathrm{AUC}_{\mathrm{ideal}} - \varepsilon_0$ in the validation set. The effect of ε_0 must be symmetric in this way because if this were not so it would imply that some residual information is available that could be used to improve $\mathrm{AUC}_{\mathrm{ideal}}$, contradicting the initial statement that $\mathrm{AUC}_{\mathrm{ideal}}$ is the optimal model.

Although we cannot know the values of either $\mathrm{AUC}_{\mathrm{ideal}}$ or ε_0, we can use them as the basis for selecting the best candidate models by maximizing the product of the AUCs for the derivation and validation set, i.e.:

$$\mathrm{AUC}_{\mathrm{derivation}} = (\mathrm{AUC}_{\mathrm{ideal}} + \boldsymbol{\varepsilon}_0) \quad \text{(B1)}$$

$$\mathrm{AUC}_{\mathrm{validation}} = (\mathrm{AUC}_{\mathrm{ideal}} - \boldsymbol{\varepsilon}_0) \quad \text{(B2)}$$

$$\mathrm{AUC}_{\mathrm{derivation}}\cdot\mathrm{AUC}_{\mathrm{validation}} = (\mathrm{AUC}_{\mathrm{ideal}} + \boldsymbol{\varepsilon}_0)(\mathrm{AUC}_{\mathrm{ideal}} - \boldsymbol{\varepsilon}_0) = \mathrm{AUC}_{\mathrm{ideal}}^2 - \boldsymbol{\varepsilon}_0^2 \quad \text{(B3)}$$

The value of equation (B3) is maximized only when $\varepsilon_0 = 0$, describing no overfitting. The candidate model that generates the greatest AUC product as defined by equation (B3) is therefore likely to be the model that most closely approximates the theoretically ideal model and has the least overfitting.

Appendix C: Mathematical Interpretation of the Final Logistic Model

The parameters of the selected airway classification model are given in the following table:

Parameter x	α	σ	ς
Face (brow-nose-chin ratio)	2.995	2.417	+1
Jaw-neck (slope high/slope low)	−13.683	3.255	+1
Nose (tilt down/tilt up)	0.557	0.735	+1
Thyromental distance (fingerbreadths)	2.032	0.738	+1
Greatest modeled difficulty (logit units)		β_0 = 10.85	

The terms in the model are likewise defined as:

$$\mathrm{L}(z) = \frac{1}{1+e^{-z}}, \text{ in which } z = \beta_0 - \zeta_1\frac{(x_1-\alpha_1)^2}{2\sigma_1^2} - \zeta_2\frac{(x_2-\alpha_2)^2}{2\sigma_2^2} - \ldots \quad \text{(C1)}$$

The value of L(z) is always within the range of 0 to 1 and is the predicted likelihood of belonging to class 1. The value of 1 − L(z) is the predicted likelihood of belonging to class

0. Therefore, if L(z) is ≤ 0.5, then the patient is predicted as class 0 (easy to intubate) and if L(z) is >0.5, then the patient is predicted as class 1 (difficult to intubate). The meanings of the parameters of the model are defined fully in Appendix A, but can also be described simply. In the quadratic logit model, the α terms identify the apex of the quadratic curve, and the σ terms represent the steepness of the sides of the curve. The variable ζ defines whether ease of intubation improves (+1) or worsens (−1) as the value of the variable moves away from α. As $\zeta = +1$ for all terms, β_0 describes the value in logit units that would be produced by the head that is most difficult to intubate according to the model, as shown in Figure 5A.

When the derivation and validation data contain such a high prevalence of difficult intubations, one might be suspicious that an algorithm produced from that data might overcall the prevalence of difficult intubation in the general population. We can address this concern by calculating the predicted difficulty of the average head, to which the model had not previously been exposed. The average head (Fig. 2) is defined as the head for which all observable parameters have 0 deviance from the population normal,[16] and hence for which all the values of x for observable parameters in the model are 0. Assigning a thyromental distance of 4 fingerbreadths, we calculate $z = -2.60$, and therefore L(z) for the average face is 0.069, which suggests a likelihood of 93.1% that the average head will be easy to intubate.

Furthermore, the meaning of the value L(z) returned by the model is unclear beyond its definition as a binomial classifier above and below $L(z) = 0.5$. It is tempting but untested to conclude that the magnitude of L(z) predicts the degree of difficulty. This would be to impose a further level of structural meaning, to say that those points that lie to the upper right of the distribution in the logit plots of Figure 4 represent not just difficult intubations but instead represent intubations comparatively "more difficult" than those represented by points lying closer to the center. The present investigation cannot address this intriguing possibility, and the difficulty in testing it against agreed upon clinical definitions will complicate future attempts to do so.

REFERENCES

1. Mallampati SR, Gatt SP, Gugino LD, Desai SP, Waraksa B, Freiberger D, Liu PL. A clinical sign to predict difficult tracheal intubation: a prospective study. Can Anaesth Soc J 1985; 32:429–34
2. Samsoon GL, Young JR. Difficult tracheal intubation: a retrospective study. Anaesthesia 1987;42:487–90
3. Frerk CM. Predicting difficult intubation. Anaesthesia 1991; 46:1005–8
4. Shiga T, Wajima Z, Inoue T, Sakamoto A. Predicting difficult intubation in apparently normal patients: a meta-analysis of bedside screening test performance. Anesthesiology 2005;103:429–37
5. Cormack RS, Lehane J. Difficult tracheal intubation in obstetrics. Anaesthesia 1984;39:1105–11
6. Crosby ET, Cooper RM, Douglas MJ, Doyle DJ, Hung OR, Labrecque P, Muir H, Murphy MF, Preston RP, Rose DK, Roy L. The unanticipated difficult airway with recommendations for management. Can J Anaesth 1998;45:757–76
7. American Society of Anesthesiologists Task Force on Management of the Difficult Airway. Practice guidelines for management of the difficult airway: an updated report by the American Society of Anesthesiologists Task Force on Management of the Difficult Airway. Anesthesiology 2003;98:1269–77
8. Adnet F, Borron SW, Racine SX, Clemessy JL, Fournier JL, Plaisance P, Lapandry C. The intubation difficulty scale (IDS): proposal and evaluation of a new score characterizing the complexity of endotracheal intubation. Anesthesiology 1997;87: 1290–7
9. Baker PA, Depuydt A, Thompson JM. Thyromental distance measurement: fingers don't rule. Anaesthesia 2009;64:878–82
10. Turk M, Pentland A. Eigenfaces for recognition. J Cogn Neurosci 1991;3:71–86
11. Valentine T. A unified account of the effects of distinctiveness, inversion, and race in face recognition. Q J Exp Psychol A 1991;43:161–204
12. Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. SIGGRAPH'99. Proceedings of the 26th annual conference on computer graphics and interactive techniques. New York: ACM Press/Addison-Wesley, 1999:187–94
13. Blanz V, Vetter T. Face recognition based on fitting a 3D morphable model. IEEE Trans Patt Anal Machine Intell 2003;25:1063–74
14. Chen TG, Fels S. Exploring gradient-based face navigation interfaces. Graphics Interface 2004. ACM International Conference Proceedings Series. Ontario: Canadian Human-Computer Communications Society, 2004;62:65–72
15. Hosmer DW, Hosmer T, Le CS, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16:965–80
16. Weisberg S. Applied Linear Regression. 3rd ed. Hoboken, NJ: Wiley-Interscience, 2005
17. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. New York: Wiley, 2000
18. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36
19. Huang J, Lin CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17:299–310
20. Frova G, Sorbello M. Algorithms for difficult airway management: a review. Minerva Anestesiol 2009;75:201–9
21. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AH, Dellinger EP, Herbosa T, Joseph S, Kibatala PL, Lapitan MC, Merry AF, Moorthy K, Reznick RK, Taylor B, Gawande AA. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 2009;360:491–9
22. Yentis SM. Predicting difficult intubation: worthwhile exercise or pointless ritual? Anaesthesia 2002;57:105–9
23. Wilson ME, John R. Problems with the Mallampati sign. Anaesthesia 1990;45:486–7
24. 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
25. Wilson ME, Spiegelhalter D, Robertson JA, Lesser P. Predicting difficult intubation. Br J Anaesth 1988;61:211–6
26. Suzuki N, Isono S, Ishikawa T, Kitamura Y, Takai Y, Nishino T. Submandible angle in nonobese patients with difficult tracheal intubation. Anesthesiology 2007;106:916–23
27. Naguib M, Malabarey T, AlSatli RA, Al Damegh S, Samarkandi AH. Predictive models for difficult laryngoscopy and intubation: a clinical, radiologic and three-dimensional computer imaging study. Can J Anaesth 1999;46:748–59
28. Santoni BG, Hindman BJ, Puttlitz CM, Weeks JB, Johnson N, Maktabi MA, Todd MM. Manual in-line stabilization increases pressures applied by the laryngoscope blade during direct laryngoscopy and orotracheal intubation. Anesthesiology 2009;110:24–31
29. Rosenstock C, Gillesberg I, Gatke 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