Vocal cord movement – can it be accurately graded?

Catriona Douglas[1], Radhika Menon[2], Jenny Montgomery[1], Richard Townsley[3], Omar Hilmi[1], Malcolm Buchanan[1], Stuart M. Robertson[1], Lykourgos Petropoulakis[2], John soraghan[2], Heba Lakany[2], and Kenneth MacKenzie[1]

[1]Queen Elizabeth University Hospital
[2]University of Strathclyde
[3]University Hospital Crosshouse

September 16, 2020

Abstract

Background: Flexible nasendoscopy is the principle method of assessment of vocal cord movement. As this procedure is inherently subjective it may be that it is not possible for clinicians to grade degree of vocal cord movement reliably. Aim: The aim of this study was to assess the accuracy and consistency of grading of vocal cord movement viewed via flexible nasendoscopy. Design: Prospective video analysis study. Methods: Thirty flexible nasendoscopy videos, without sound or clinical information, were assessed by 6 consultant Head and Neck surgeons. They were asked to assess and grade right and left vocal cord movement independently, based on a 5-category scale. This process was repeated three times at separate time intervals. Agreement and reliability were assessed. Participants: 6 consultant head and neck surgeons. Results: The mean overall percentage of observed inter-rater agreement was 67.7% (SD 1.9) with the 5-categories scale, increasing to 91.4% (SD 1.9) when a 3-category scale was derived. The mean overall percentage of observed intra-rater agreement was 78.3% (SD 9.7) for 5 categories, increasing to 93.1% (SD 3.3) for 3 categories. Discriminating vocal cord motion using the 5-category scale is less reliable (k= 0.52) than with the 3-category scale (k = 0.68). Conclusion: This study demonstrates quantitatively that it is challenging to accurately and consistently grade subtle differences of vocal cord movement as proven by lesser agreement and reliability when using a 5 point scale instead of a 3 point scale. It highlights the need to have an objective measure to help in the assessment of vocal cord movement.

Introduction

Flexible nasendoscopy (FNE) is the most commonly performed examination used for assessment of vocal cord movement. It is routinely performed on patients with voice complaints and management is frequently based on the findings. FNE is the current gold standard method of evaluation of vocal cord motion, helping to distinguish between normal and reduced vocal cord movement. However, this subjective assessment can lead to inaccuracies and variability in diagnosis, especially in challenging cases. There is no reliable objective measure of categorising vocal cord movement from normal to complete paralysis. Furthermore, limited data exist on how consistent otolaryngologists are at rating vocal cord movement [1,2].

The reliability of clinicians differentiating between binary categories of mobile and immobile vocal cords[1,2] and the presence or absence of paresis have been reported[3]. The former studies reported a 95% inter-rater reliability and a 99% intra-rater agreement for binary vocal cord assessment and the latter an inter-rater reliability of 0.334 (Fleiss's kappa). A three category scale (paralysis, paresis, normal) was used in reliability studies in paediatric patients[4]. They reported an inter-rater reliability (Cohen's kappa) of 0.67 for diagnosis of normal vs impaired movement, and lesser reliability of 0.49 when identifying the degree of movement (normal, paresis, paralysis). The intra-rater reliability ranged from 0.48 to 1 (Fleiss's kappa). There is currently no

reliable grading system for categorising vocal cord movement from normal to complete paralysis, for example similar to the House Brackman scale used to routinely grade facial nerve paralysis.

The aim of this study was to determine if experienced consultant head and neck surgeons were accurate and consistent with their assessment and grading of vocal cord movement.

Materials and Methods

Thirty flexible fibreoptic nasendoscopy videos of laryngeal movement were captured in a laryngology clinic. These ranged from normal vocal cord movement to complete laryngeal paralysis (nine normal cases, four palsies, three nodules, two cases each of cysts, functional dysphonia, and inflammation, one case each of Reinke's oedema, presbyphonia, polyp, hypo pharynx lesion, supraglottic lesion, crescentic defect of vocal cord, weakness and slower right vocal cord movement). The videos were pre-processed to reduce the effect of the honeycomb artefact caused by the fibreoptic endoscopes[5]. Six consultant head and neck surgeons (JM, RT, OH, MB, SR, KM) were asked to subjectively assess vocal cord motion by visual inspection of the laryngeal videos and individually rate the movement of the left and right vocal cords independently on a scale of 0 to 4, see table 1. There was no clinical history or sound associated with the videos. This process was repeated with the same videos, in a different order, on three separate occasions with a minimum of two weeks between each rating session. Each consultant rated the videos three times giving a total of 180 individual ratings (2 [R & L cord] x 30 x 3 = 180) per consultant and 1080 (180 x 6 consultants) ratings in total. The consultants were blinded to their previous and other raters' scores. Ethical approval was not required for this study.

Statistical analysis

*Agreement*Agreement was computed using the 'percentage agreement' measure, which provides the percentage of cases in which two or more raters scored identically. To assess inter-rater agreement, two percentage agreement measures were computed, i.e., the overall agreement between raters for all categories combined (overall percentage agreement); and the agreement specific to a category (specific agreement). The purpose of 'specific agreement' is to objectively demonstrate if the clinicians are in better agreement while rating cases belonging to some categories more than others (such as fully mobile category as opposed to paresis). Intra-rater agreement (i.e., test-retest) was also computed for each consultant over the three sessions using overall percentage agreement.

*Reliability*Inter- and intra-rater reliability was calculated using the generalised Fleiss's kappa [4,6,7] to compare with comparable studies reported in the literature. The kappa statistic ranges from 0 to 1, where 0 depicts that raters are in agreement only by chance. Any value over 0 may be interpreted as representing: poor (below 0.40), fair to good (between 0.40-0.75) and excellent (above 0.75) agreement beyond chance. The rating scale was considered as an ordinal scale and an ordinal weighting scheme was used in the computation of Fleiss's kappa[4,6].

For the intra rater study, we had 3 sessions (i.e., replicates) per sample, which is appropriate[8,9] since moderately high (>0.60) reliability was expected based on the trend in the literature[1,3,4]. Since reliability was expected to be lower in the inter-rater study (as low as 0.33 [3]), 6 raters are appropriate[10]

Rating scales

The study was conducted using a five-category scale (Table 1), to determine if subtle differences in vocal cord motion can be visualised consistently between clinicians. It goes beyond the routine practice of describing motion as normal, paresis and paralysis, which is effectively a three-category scale. Hence a three-category scale was derived from the original five category scale to know the agreement/reliability using categories (normal/paresis/paralysis) that clinicians would normally use. This would allow comparison between the three and five category scales. The recategorization from five to three category was as follows; scores assigned to categories 0 and 1 were grouped together to form the 'immobile' category, the scores assigned to categories 3 and 4 were grouped together to form the 'fully mobile' category, and category 2 remained effectively a

'paresis' category, resulting in the derived clinically relevant 3-category scale. Inter and intra rater agreement and reliability measures were repeated using the derived scale.

Results -

All 6 Consultants completed all the video assessments, giving a total of 1080 individual vocal cord assessments. The results for the recorded five category and derived three category scales are reported.

Agreement measures

The exact agreement in scores between the consultants, averaged over the three sessions is provided in Table 2. The overall percentage of observed inter-rater agreement, as shown in Table 2, is consistent across sessions and has a mean value of 67.7% with the 5-categories scale, which increases to 91.4% when the 3-categories scale is used.

There is greater variability in performance of the consultants in the 5 category intra-rater study, with overall percentage agreement for a consultant between the three sessions ranging from 63.9% to 88.9%. The mean intra-rater agreement of the six consultants is 78.3% with a standard deviation of 9.7%. With the 3-category case, not only does the mean intra-rater agreement improve by 14.8% to attain 93.1% agreement, the variability in performance between consultants reduces as shown by the standard deviation of the mean agreement measure reducing 3-fold.

The specific agreement between consultants for each category, averaged over the three sessions, is provided in Table 3.

Reliability measures

The consistency of discriminating vocal cord motion between the consultants (inter-) and between sessions for a given consultant (intra-) are provided in Table 4. Kappa values are consistent across sessions and the reported inter-rater reliability is the mean reliability of all sessions. Discriminating vocal cord motion using the 5-category scale is less reliable ($\varkappa = 0.52$) than with the 3-category scale ($\varkappa = 0.68$).

The intra-rater or test-retest reliability is the mean reliability of each consultant over the three sessions. With the 5-category scale, intra-rater reliability ranged from 0.55 (fair) to 0.82 (excellent), with a mean of 0.69. The kappa values increased with the 3-category scale and ranged from 0.64 to 0.87, with a mean kappa of 0.75. Two out of six consultants had excellent reliability (0.78 and 0.82) with the 5-category scale and three consultants had excellent reliability with the 3-category scale (0.78, 0.87 and 0.87).

Discussion

The correct diagnosis of a vocal cord movement abnormality is vital to help guide management of the patient, with potential medicolegal implications if misdiagnosed. There are many causes of abnormal movement, with movement ranging from fully mobile, paresis to complete paralysis. Ideally clinical assessment would result in a reliable five category scale to allow use in the range of clinical situations such as reduction in movement in early invasive cancer or post thyroid surgery. Although the present gold standard for assessing the movement of a vocal cord is flexible nasendoscopy, there are few published studies assessing the consistency between different raters.

Comparison between raters – agreement measures

The six raters were asked to assess movement on a 5-point scale, ranging from no movement to fully mobile. Inter-rater specific agreement was less than 60% for 4 of the 5 categories; immobile, slightly reduced movement, minimal residual mobility and paresis. The only category to have high inter-rater specific agreement of 83.04%, was the fully mobile category. This may simply be because this is what clinicians see most commonly when performing flexible nasendoscopy - a fully mobile vocal cord - with the high agreement being a reflection of pattern recognition. Furthermore, since the dataset was formed from routine clinical cases, about 70% of them are of fully mobile vocal cords. Therefore, due to the high prevalence, the positive predictive value of the clinicians for this score category would be high[11]. Furthermore, when assessing each individual

rating in the five-point scale the combined agreement measure in each category varied considerably, ranging from only 16.6% for score 1 (minimal movement) to 83% for score 4 (fully mobile). This significant range in agreement highlights the difficulty in assessment of vocal cord mobility. When the options are limited to three categories, there was improved inter-rater specific agreement, with fully mobile agreement at 96.11%, and no mobility at 75.11%.

Analysis of specific agreement scores provides an insight into the categories the consultants were in greater agreement and the reason for the improvement of the scores with the 3-category scale. Clearly, much of the variability in scoring between clinicians is in the categories 1 (minimal residual mobility), 2 (paresis) and 3 (slightly reduced mobility) in the 5-category scale. The agreement in these categories for any session was less than 31%.

Comparison between raters – reliability measures

The consistency of discriminating vocal cord motion between the consultants was assessed. Discriminating vocal cord motion using the 5-category scale was less reliable ($\varkappa = 0.52$) compared to using the 3-category scale ($\varkappa = 0.68$), with both values falling in the fair to good grouping of reliability measures[11]. Liu et al, when assessing paediatric patients, reported a reliability of k=0.49 for 3 categories[4]. Assuming that nasendoscoping is more challenging in the paediatric population and that they too did not use audio, our results seem comparable. Madden et al reported higher inter-rater reliability of 95%, but they used a binary scale, i.e., purposeful vocal fold motion or no purposeful vocal fold motion, and their video data included audio. Nevertheless, Rosow et al who also included audio and employed a binary scale, reported the reliability of identifying the presence or absence of volitional adduction as only k=0.335[2]. However their assessment was based on stroboscopy making it difficult to draw any firm comparisons.

Repeatability of assessment

Consistency of re-examination affects clinical outcome and management decisions. When the 5-point scale is used, it is clear that the intra-rater consistency is lower, compared to the 3 point score.

The diagnosis of vocal cord paresis is felt to be more challenging than vocal cord paralysis[12]. This is highlighted in this study with the low inter rater specific agreement for the scores 1, 2 and 3 in the 5 point scale, and score 1 in the 3 point scale (Table 3), which demonstrates that clinicians disagree with what they are seeing when vocal cord paresis is present. Vocal cord movement is a continuum with paresis not as well recognised or studied as paralysis. Wu et al highlighted that in laryngology practice in North America, the most common diagnostic tool for diagnosing paresis was stroboscopy, not flexible nasendoscopy. Simpson et al reported that in a large series of 739 patients presenting to a tertiary laryngology service with a chief complaint of dysphonia, of the 26.4% with paresis or paralysis on stroboscopy, only 1.8% of the patients had LEMG confirmed vocal fold paresis. In stark comparison Satalof et al demonstrated that in his series of 689 patients with suspected paresis or paralysis, the LEMG confirmed this diagnosis in 95.9% of the patients. This significant variation between diagnosis and confirmation on LEMG highlights that we are still not able to consistently differentiate between these diagnoses. Although LEMG is the only way to confirm definitively that a patient has a paralysis or paresis, it is not routinely performed in clinical practice.

Limitations of the study

This study aimed at assessing the consistency of clinicians evaluating the movement of the vocal cord on a rating scale. "Worst case scenario" clinical situations were used, where the clinician had no history from the patient and was unable to hear the patient's voice when they assessed the video of vocal cord movement. There was no extra information asked on the numerous other clinical findings that are seen in patients with vocal fold paralysis such as arytenoid prolapse, posterior gap, height and length mismatch. The wide variation in interrater scores for the 5 point scale may be related to the fact that there was no accompanying clinical history or sound with the videos, making it an artificial situation. Madden et al, when assessing consistency of vocal fold motion, included sound with their videos and they demonstrated higher inter-rater reliability, suggesting that a "complete picture" is required when assessing vocal cord movement. All the

endoscopies performed were fibreoptic flexible nasendoscopy, which rendered poorer video quality than newer generation distal chip views, possibly making the more subtle movement of the vocal cords more difficult to judge and categorise. However, the videos very much reflected the reality of seeing patients in clinics and wards.

Conclusion

This study demonstrates quantitatively that it is challenging to accurately and consistently grade subtle differences of vocal cord movement as proven by lesser agreement and reliability when using a 5 point scale instead of a 3 point scale. Therefore, it highlights the need to have an objective measure to improve the accuracy of assessment of vocal cord movement. Image processing of endoscopy videos could be employed for measurement of vocal cord movement symmetry to quantify the degree of vocal cord motion, thus providing a reliable measure to assist in diagnosis and evaluate post treatment outcomes.

References

1. Madden LL, Rosen CA. Evaluation of Vocal Fold Motion Abnormalities: Are We All Seeing the Same Thing? *J Voice.* 2017;31(1):72-77.

2. Rosow DE, Sulica L. Laryngoscopy of vocal fold paralysis: evaluation of consistency of clinical findings. *Laryngoscope.*2010;120(7):1376-1382.

3. Estes C, Sadoughi B, Mauer E, Christos P, Sulica L. Laryngoscopic and stroboscopic signs in the diagnosis of vocal fold paresis.*Laryngoscope.* 2017;127(9):2100-2105.

4. Liu YC, McElwee T, Musso M, Rosenberg TL, Ongkasuwan J. The reliability of flexible nasolaryngoscopy in the identification of vocal fold movement impairment in young infants. *Int J Pediatr Otorhinolaryngol.* 2017;100:157-159.

5. Menon R PL, Soraghan JJ, Lakany H, MacKenzie K, Hilmi O, DiCaterina G. . Automatic quantification of vocal cord paralysis: An application of fibre-optic endoscopy video processing. . Paper presented at: 10th International Joint Conference on Biomedical Engineering Systems and Technologies.2017.

6. Gwent KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters (4th ed.). In: Gaithersburg M, ed.: Advanced Analytics; 2014.

7. Girard JM. Master Inter-Observer Reliability. 2019; http://mreliability.jmgirard.com.

8. Shoukri MM, Asyali, M.H. and Donner, A. Sample Size Requirements for the Design of Reliability Study: Review and Results. *Statistical Methods in Medical Research.* 2004;13:251-271.

9. Giraudeau B, Mary JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Stat Med.* 2001;20(21):3205-3214.

10. Donner A, Eliasziw M. Sample size requirements for reliability studies. *Stat Med.* 1987;6(4):441-448.

11. Fleiss JL. Measuring Nominal Scale Agreement among Many Raters. .*Psychological Bulletin.* 1971;76:378-382.

12. Wu AP, Sulica L. Diagnosis of vocal fold paresis: current opinion and practice. *Laryngoscope.* 2015;125(4):904-908.

Table 1. Rating scale used by the consultants

Score	Definition
0	No motion - The vocal cord is completely paralysed and shows no movement at all
1	Almost no motion - The vocal cord is not completely paralysed, but shows only very slight movement
2	Half the range of motion - The vocal cord moves about half the range of motion of that of a healthy vocal cord

Score	Definition
3	Almost full motion - The vocal cord moves with almost full range of motion, but not completely
4	Full range of motion - The vocal cord moves completely with full range of motion

Table 2. Overall percent agreement

Rating scale—	Inter-rater agreement in %	Inter-rater agreement in %	Intra-rater agreement in %	Intra-rater agreem
	Mean	**SD**	**Mean**	**SD**
5 categories	67.7	1.9	78.3	9.7
3 categories	91.4	1.9	93.1	3.3

Key: Mean inter-rater agreement is the agreement between consultants in a given session, averaged over the three sessions

Mean intra-rater agreement is the agreement in the scores of a consultant between the three sessions, averaged over all consultants

SD is the standard deviation above or below the means

Table 3. Inter-rater specific agreement

Rating scale—	Inter-rater agreement in % (±SD)	Inter-rater agreement in % (±SD)	Inter-rater agreement in % (±SD
5 categories	Immobile	Minimum residual mobility	Paresis
	58.6 (8.4)	16.7 (10.1)	23.9 (5.9)
3 categories	Immobile	Immobile	Paresis
	75.1 (4.3)	75.1 (4.3)	23.9 (5.9)

Table 4. Reliability measures

Rating scale—	Inter-rater reliability: Fleiss's kappa (±SD)	Intra-rater reliability: Fleiss's kappa (±SD)
5 categories	0.52 (0.03)	0.69 (0.11)
3 categories	0.68 (0.06)	0.75 (0.1)