Received: 26 May 2024 | Accepted: 9 June 2024

DOI: 10.1002/lio2.1298

ORIGINAL RESEARCH

Laryngoscope Investigative Otolaryngology

Inter-rater reliability and clinical relevance of subjective and objective interpretation of videofluoroscopy findings

Jonna Kuuskoski MD[1] | Jaakko Vanhatalo MD[2] | Jussi Hirvonen MD, PhD[2,3] | Jami Rekola MD, PhD[1] | Leena-Maija Aaltonen MD, PhD[4] | Pia Järvenpää MD, PhD[4]

[1]Department of Otorhinolaryngology—Head and Neck Surgery, Turku University Hospital and University of Turku, Turku, Finland

[2]Department of Radiology, Turku University Hospital and University of Turku, Turku, Finland

[3]Faculty of Medicine and Health Technology, Tampere University Hospital and Tampere University, Tampere, Finland

[4]Department of Otorhinolaryngology—Head and Neck Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

Correspondence
Jonna Kuuskoski, Department of Otorhinolaryngology—Head and Neck Surgery, Turku University Hospital and University of Turku, Turku, Finland. P.O. Box 52, FI-20521. Finland.
Email: jonna.kuuskoski@varha.fi

Abstract

Background: Dysphagia is commonly evaluated using videofluoroscopy (VFS). As its ratings are usually subjective normal-abnormal ratings, objective measurements have been developed. We compared the inter-rater reliability of the usual VFS ratings to the objective measurement VFS ratings and evaluated their clinical relevance.

Methods: Two blinded raters analyzed the subjective normal-abnormal ratings of 77 patients' VFS. Two other blinded raters analyzed the objective measurements of pharyngeal aerated area with bolus held in the oral cavity (PAhold), the pharyngeal area of residual bolus during swallowing (PAmax), the pharyngeal constriction ratio (PCR), the maximum pharyngoesophageal segment opening (PESmax), pharyngoesophageal segment opening duration (POD), airway closure duration (ACD), and total pharyngeal transit time (TPT). We evaluated the inter-rater agreement in the subjective ratings and the objective measurements. Clinical utility analysis compared the measurements with the VFS findings of pharyngeal phase abnormality, penetration/aspiration, and cricopharyngeal relaxation.

Results: In the pharyngeal findings, the subjective analysis inter-rater agreement was mainly moderate to strong. The strongest agreements were on the pharyngeal residues and penetration/aspiration findings. The objective measurements had fair to good inter-rater agreement. Clinical utility analysis found statistically significant connections between TPT and pharyngeal phase abnormality, normal PCR and lack of penetration/aspiration, and normal PESmax and normal cricopharyngeal relaxation.

Conclusions: The subjective analysis had moderate to strong inter-rater agreement in the pharyngeal VFS findings, especially concerning pharyngeal residues and penetration/aspiration detection, reflecting the efficacy and safety of swallowing. The objective measurements had fair to good inter-observer reproducibility and could thus improve the reliability of VFS diagnostics.

Level of evidence: 4.

KEYWORDS
deglutition, dysphagia, inter-rater reliability, videofluoroscopy

1 | INTRODUCTION

Difficulty swallowing can derive from numerous etiologies and may seriously impact an individual's health. Dysphagia can impair a person's nutritional status and their ability to take medications, lead to dehydration and malnutrition, considerably weaken quality of life, and even lead to death.[1] Several methods can be used to examine dysphagia, among which videofluoroscopy (VFS) and fiberoptic endoscopic examination of swallowing (FEES) are considered reference standards.[2] VFS effectively visualizes both the oropharyngeal and esophageal phases of swallowing. However, even though VFS is commonly used for evaluating dysphagia, the way in which its findings are interpreted and analyzed varies.

Several systems have been developed for rating VFS findings, such as the Videofluoroscopic Dysphagia Scale, the Modified Videofluoroscopic Dysphagia Scale, the Group for Learning Useful and Performant Swallowing (GLUPS) score, and the Modified Barium Swallow Impairment Profile (MBSImp).[3–8] Most of these rating systems consider the oropharyngeal phase of swallowing, but the GLUPS score also covers esophageal findings.[7]

The inter-rater reliability of most of these subjective methods has been reported as unsatisfactory.[3,4,6,9] Therefore, objective measurement-based methods have been developed to rate VFS findings. The MBSImp has been reported to have good inter- and intra-rater reproducibility.[8] However, this method requires licensed, relatively expensive training and is not widely accessible.

Kendall and Leonard et al. have published temporal and spatial parameters for normal and abnormal swallowing.[10,11] These objective measurements are time consuming, and they have not yet gained popularity in clinical use, but they might enable objectivity in the evaluation of swallowing.

The purpose of this study was to compare the inter-rater reliability and clinical utility of subjective normal-abnormal ratings based on the GLUPS score and the objective measurement-based rating method of Kendall and Leonard et al. for evaluating VFS findings in our tertiary care hospital.[7,10,11]

2 | METHODS

The study was conducted in the Turku University Hospital in Finland. The study cohort consisted of 77 dysphagia patients with multiple etiologies referred for VFS between December 2021 and June 2022. The VFS findings were subjectively rated and objectively measured.

2.1 | Videofluoroscopy

VFS was imaged using a multipurpose twin robotic X-ray system (Multitom Rax, Siemens Healthcare GmbH, Erlangen, Germany) with a tube voltage of 73 kV and Cu Filter of 0.2 mm. Patients were generally imaged using 15 pulses per second in accordance with local "as low as reasonably achievable" (ALARA) principles, but a 30 pulses per second continuous mode was used if clarification was needed. Analyses for subjective ratings and objective measurements were made with 10 and 20 mL boluses of liquid contrast Iodine contrast agent Iohexol (Omnipaque 300 mg/mL, GE Healthcare, Princeton, New Jersey) with IDDSI 0 (International Dysphagia Diet Standardization Initiative Framework[12,13]) consistency level. A 2 cm calibration ball was used in lateral view imaging. All the images were saved on the Turku University Imaging Server (Philips Vue PACS). Imaging schema has been described in detail in the previous publication.[14]

2.2 | Subjective ratings

The VFS findings were analyzed by an experienced radiologist (JV) and a clinician (ear, nose, and throat doctor and phoniatrician JK) who were blinded from each other's ratings by simultaneously filling a Videofluoroscopy Finding Form in real time as they examined the VFS and later checking the results through frame-by-frame analysis. The Videofluoroscopy Finding Form was modified on the basis of the GLUPS score.[7] The findings were categorized into oral findings (lip closure, tongue movement, possible premature pharyngeal spillage, possible oral residue), pharyngeal findings (swallow onset, velopharyngeal closure, epiglottal retroflexion, laryngeal elevation, possible vallecular residue, possible pyriform sinus residue, penetration/aspiration), and esophageal findings (upper esophageal sphincter opening, peristalsis of the esophagus, possible stasis of the esophagus, and lower esophageal sphincter function). Each of these 15 criteria was classified as normal or abnormal.

2.3 | Objective measurements

Objective analysis was performed 9–15 months later, and the raters were blinded from the original subjective ratings. All the VFS lateral view images of the liquid bolus consistencies were analyzed using a software program designed for quantitative VFS analysis (Swallowtail, Belldev Medical, Chicago, USA, version 3 and 4). The following parameters were measured from the lateral views of the most pathological image series with 10 or 20 mL bolus by two experienced radiologists (JV and JH): the pharyngeal aerated area when the iodine bolus was held in the oral cavity (PAhold), the pharyngeal area of the residual bolus during swallowing (PAmax), the pharyngeal constriction ratio (PCR), the maximum pharyngoesophageal segment opening (PESmax), the pharyngoesophageal segment opening duration (POD), the airway closure duration (ACD) and the total pharyngeal transit time (TPT), as described in previous publications.[10,11,15–18] Normative values of most of the measurements were obtained from the book by Leonard and Kendall,[19] and values were considered abnormal if they were not within two standard deviations (SD) of the normal mean for the given parameter. Normative PCR values were obtained from the publication of Leonard et al.[18]

2.4 | Ethical considerations

Before agreeing to take part in the study, the patients were given both oral and written information on the research and signed their written consent. The Ethics Committee of the Hospital District of Southwest Finland approved the study protocol, and permission for the research was granted by the Hospital District of Southwest Finland. The study was conducted in accordance with the Declaration of Helsinki (The World Medical Association 2013).

2.5 | Statistical analysis

We measured inter-rater reliability using Cohen's kappa statistics with 95% confidence intervals (CI) to determine the consistency of the subjective normal-abnormal ratings. Cohen's kappa values were interpreted as: 0.0–0.20, no agreement; 0.21–0.39, minimal agreement; 0.40–0.59, weak agreement; 0.60–0.79, moderate agreement; 0.80–0.90, strong agreement; >0.90, almost perfect agreement.[20] The inter-rater agreement between the objective measures was assessed using the intraclass correlation coefficient (ICC). The ICC with 95% CI was assessed using a two-way mixed-effect model that was based on single ratings and absolute agreements. The results were interpreted as: <0.50, poor; 0.50–0.75, fair; 0.75–0.90, good; >0.90, excellent.[21]

Fisher's exact test was used to determine whether the objective measurements and the subjective ratings of pharyngeal phase abnormality, penetration/aspiration, or cricopharyngeal relaxation were significantly associated. *p* values less than 0.05 were considered statistically significant.

All the statistical analyses were performed using IBM SPSS Statistics for Windows (version 26.0; IBM Corp., Armonk, NY, USA).

3 | RESULTS

3.1 | Inter-rater correlations of subjective ratings

The inter-rater correlations between the subjective normal-abnormal ratings of the radiologist and the clinician are shown in Tables 1–3. In the case of one patient who had severe aspiration in a thick consistency, the examination was ceased, and liquid consistency was not

TABLE 1 Inter-rater correlations (κ-values) and proportion of normal findings in oral phase of videofluoroscopy by radiologist and ear, nose, and throat (ENT) doctor.

Finding	Consistency	Proportion of findings classified as normal by radiologist *n*/%	Proportion of findings classified as normal by ENT *n*/%	*n* total	Cohen's kappa (95% CI)	*p* value	Evaluation of reliability
Lip closure normal	Liquid	76/100%	76/100%	76	-		-
	Moderately thick	76/98.7%	75/97.4%	77	−0.013 (−0.031–0.005)	0.908	-
	Solid	77/100%	77/100%	77	-		-
Tongue movement	Liquid	75/98.7%	75/98.7%	76	1.000	<0.001	Perfect agreement
	Moderately thick	76/98.7%	76/98.7%	77	−0.013 (−0.031–0.005)	0.908	-
	Solid	74/96.1%	74/96.1%	77	0.653 (0.195–1.111)	<0.001	Moderate agreement
Premature pharyngeal spillage	Liquid	71/93.4%	68/89.5%	76	0.414 (0.054–0.774)	<0.001	Weak agreement
	Moderately thick	72/93.5%	71/92.2%	77	0.706 (0.386–1.026)	<0.001	Moderate agreement
	Solid	71/92.2%	68/88.3%	77	0.191 (−0.127–0.509)	0.086	No agreement
Oral residue	Liquid	73/96.1%	69/90.8%	76	0.365 (−0.033–0.763)	<0.001	Minimal agreement
	Moderately thick	75/97.4%	70/90.9%	77	0.421 (0.013–0.829)	<0.001	Weak agreement
	Solid	74/96.1%	73/94.8%	77	0.551 (0.091–1.011)	<0.001	Weak agreement
Oral phase normal		63/85.1%	56/72.7%	77	0.525 (0.299–0.751)	<0.001	Weak agreement

Abbreviation: CI, Confidence interval.

Laryngoscope Investigative Otolaryngology

TABLE 2 Inter-rater correlations (κ-values) and proportion of normal findings in pharyngeal phase of videofluoroscopy by radiologist and ear, nose, and throat (ENT) doctor.

Finding	Consistency	Proportion of findings classified as normal by radiologist *n*/%	Proportion of findings classified as normal by ENT *n*/%	*n* total	Cohen's kappa (95% CI)	*p* value	Evaluation of reliability
Onset of swallow	Liquid	74/97.4%	69/90.8%	76	0.421 (0.013–0.829)	<0.001	Weak agreement
	Moderately thick	74/96.1%	71/92.2%	77	0.648 (0.276–1.020)	<0.001	Moderate agreement
	Solid	72/93.5%	72/93.5%	77	0.572 (0.188–0.956)	<0.001	Weak agreement
Velopharyngeal closure	Liquid	76/100%	76/100%	76	-		-
	Moderately thick	76/98.7%	76/98.7%	77	-		-
	Solid	77/100%	77/100%	77	-		-
Epiglottal retroflexion	Liquid	73/96.1%	75/98.7%	76	0.490 (−0.122–1.102)	<0.001	Weak agreement
	Moderately thick	74/96.1%	77/100%	77	-		-
	Solid	75/97.4%	77/100%	77	-		-
Laryngeal elevation	Liquid	76/100%	76/100%	76	-		-
	Moderately thick	77/100%	77/100%	77	-		-
	Solid	77/100%	77/100%	77	-		-
Vallecular residue	Liquid	50/65.8%	46/60.5%	76	0.774 (0.624–0.924)	<0.001	Moderate agreement
	Moderately thick	60/77.9%	58/75.3%	77	0.855 (0.715–0.995)	<0.001	Strong agreement
	Solid	53/68.8%	49/63.6%	77	0.653 (0.471–0.835)	<0.001	Moderate agreement
Pyriform sinus residue	Liquid	59/77.6%	58/76.3%	76	0.666 (0.460–0.872)	<0.001	Moderate agreement
	Moderately thick	65/84.4%	68/88.3%	77	0.725 (0.493–0.957)	<0.001	Moderate agreement
	Solid	63/81.8%	63/81.8%	77	0.825 (0.657–0.993)	<0.001	Strong agreement
Penetration or aspiration	Liquid	69/90.8%	72/94.7%	76	0.708 (0.392–1.024)	<0.001	Moderate agreement
	Moderately thick	73/94.8%	75/97.4%	77	0.655 (0.203–1.107)	<0.001	Moderate agreement
	Solid	75/97.4%	76/98.7%	77	0.661 (0.027–1.295)	<0.001	Moderate agreement
Pharyngeal phase normal		38/49.3%	29/37.7%	77	0.661 (0.495–0.827)	<0.001	Moderate agreement

Abbreviation: CI, Confidence interval.

TABLE 3 Inter-rater correlations (κ-values) and proportion of normal findings in esophageal phase of videofluoroscopy by radiologist and ear, nose, and throat (ENT) doctor.

Finding	Consistency	Proportion of findings classified as normal by radiologist *n*/%	Proportion of findings classified as normal by ENT *n*/%	*n* total	Cohen's kappa (95% CI)	*p* value	Evaluation of reliability
Cricopharyngeal relaxation	Liquid	58/76.3%	57/75.0%	76	0.679 (0.481–0.877)	<0.001	Moderate agreement
	Moderately thick	59/76.6%	56/72.7%	77	0.692 (0.502–0.882)	<0.001	Moderate agreement
	Solid	59/76.6%	57/74.0%	77	0.581 (0.365–0.797)	<0.001	Weak agreement
Esophageal peristalsis	Liquid	49/65.3%	42/56.0%	75	0.529 (0.333–0.725)	<0.001	Weak agreement
	Moderately thick	52/67.5%	40/51.9%	77	0.526 (0.340–0.712)	<0.001	Weak agreement
	Solid	48/63.2%	27/35.5%	76	0.389 (0.215–0.563)	<0.001	Minimal agreement
Esophageal obstruction/ stasis	Liquid	68/91.9%	73/98.6%	74	0.269 (−0.161–0.699	<0.001	Minimal agreement
	Moderately thick	71/93.4%	74/97.4%	76	0.555 (0.103–1.007)	<0.001	Weak agreement
	Solid	67/89.3%	73/97.3%	75	0.373 (−0.009–0.755)	<0.001	Minimal agreement
LES function	Liquid	51/68.9%	42/56.8%	74	0.459 (0.253–0.665)	<0.001	Weak agreement
	Moderately thick	54/71.1%	39/51.3%	76	0.548 (0.370–0.726)	<0.001	Weak agreement
	Solid	49/65.3%	30/40.0%	75	0.472 (0.300–0.644)	<0.001	Weak agreement
Esophageal phase normal		18/23.4%	12/15.6%	77	0.590 (0.362–0.818)	<0.001	Weak agreement

Abbreviations: CI, Confidence interval; LES, Lower esophageal sphincter.

examined. Three patients lacked esophageal ratings in some consistencies, so the total amount of patients analyzed varied from 74 to 77 in these ratings.

Overall, the agreement between the subjective ratings varied from weak to strong. The weakest agreement was in the oral and esophageal findings, whereas the agreement was mainly moderate or strong in the pharyngeal findings (Tables 1–3). The strongest agreements were in the pharyngeal residues with κ-values between 0.653 and 0.855 (95% CI 0.461–0.995, $p < 0.001$). Some categories had no abnormal findings, and in these cases, it was not possible to calculate the agreement.

3.2 | Inter-rater correlations of objective measurements

Table 4 shows the inter-rater correlation results of the objective measurements of the two radiologists. Seventy-six patients were measured because one patient's information on liquid consistency was missing due to aspiration. The ICC values varied from 0.648 (95% CI 0.497–0.761, $p < 0.001$) of PAmax to 0.864 (95% CI 0.795–0.912, $p < 0.001$) of PAhold, so the measurements had fair to good inter-observer reproducibility. PCR had a poor ICC value.

In comparison to the published normative values, the radiologists' normal-abnormal classifications varied from minimal to moderate agreement (Table 5).[18,19] The clinical utility of the objective measurements was assessed by comparing them with the normal-abnormal results of the subjective ratings of pharyngeal phase normality (Table 6), cricopharyngeal relaxation (Table 7), and penetration/aspiration (Table 8). We found only a few statistically significant connections: TPT and pharyngeal phase abnormality with a sensitivity of 71.8% and a specificity of 60.5%, $p = 0.006$; normal PESmax and cricopharyngeal relaxation normality with a sensitivity of 98.3% and a specificity of 16.7%, $p = 0.039$; and normal PCR and lack of penetration/aspiration with a sensitivity of 98.6% and a specificity of 28.6%, $p = 0.021$.

4 | DISCUSSION

VFS is the gold standard in dysphagia evaluation, but the examination is not usually standardized, and interpretations of the findings vary. In

TABLE 4 Inter-rater reliability (intraclass correlation coefficients) of two radiologists' analyses of objective measures in videofluoroscopy.

Measure	ICC (95% CI)	*p* value	Evaluation of reliability
ACD	0.783 (0.679–0.856)	<0.001	Good agreement
PAhold	0.864 (0.795–0.912)	<0.001	Good agreement
PAmax	0.648 (0.497–0.761)	<0.001	Fair agreement
PESmax	0.776 (0.669–0.852)	<0.001	Good agreement
POD	0.749 (0.631–0.833)	<0.001	Fair agreement
TPT	0.787 (0.684–0.859)	<0.001	Good agreement
PCR	0.001 (−0.222–0.224)	0.496	Poor agreement

Abbreviations: ICC, Intraclass correlation coefficient; CI, confidence interval; ACD, Airway closure duration; PAhold, Pharyngeal area when bolus held in oral cavity: PAmax, Pharyngeal area at the point of maximum pharyngeal constriction; PESmax, Maximum pharyngoesophageal segment opening; POD, Pharyngoesophageal segment opening duration; TPT, Total pharyngeal transit time, PCR, Pharyngeal construction ratio.

TABLE 5 Inter-rater correlations (κ-values) of two radiologists' normal-abnormal classification in comparison to normative values.

Measure	Cohen's kappa (95% CI)	*p* value	Evaluation of reliability
ACD	0.703 (0.461–0.944)	<0.001	Moderate agreement
PAhold	0.682 (0.445–0.919)	<0.001	Moderate agreement
PAmax	0.325 (−0.042–0.692)	<0.001	Minimal agreement
PESmax	0.309 (−0.185–0.802)	<0.001	Minimal agreement
POD	0.570 (0.319–0.821)	<0.001	Weak agreement
TPT	0.716 (0.563–0.869)	<0.001	Moderate agreement
PCR	0.551 (0.100–1.002)	<0.001	Weak agreement

Abbreviations: ACD, Airway closure duration; CI, Confidence interval; PAhold, Pharyngeal area when bolus held in oral cavity; PAmax, Pharyngeal area at the point of maximum pharyngeal constriction; PCR, Pharyngeal construction ratio; PESmax, Maximum pharyngoesophageal segment opening; POD, Pharyngoesophageal segment opening duration; TPT, Total pharyngeal transit time.

TABLE 6 Comparisons of objective measurements and subjective evaluation of pharyngeal phase ($n = 77$).

	Pharyngeal phase normal *n* /% within normal	Pharyngeal phase abnormal *n* /% within abnormal	*p* value
ACD normal	35/92.1%	35/89.7%	
ACD abnormal	3/7.9%	4/10.3%	1.000
PAhold normal	32/84.2%	35/89.7%	
PAhold abnormal	6/15.8%	4/10.3%	0.517
PAmax normal	38/100%	36/92.3%	
PAmax abnormal	0/0.0%	3/7.7%	0.240
TPT normal	23/60.5%	11/28.2%	
TPT abnormal	15/39.5%	28/71.8%	**0.006***
PCR normal	38/100%	36/92.3%	
PCR abnormal	0/0.0%	3/7.7%	0.240

Note: Statistically significant result is marked with an asterisk (*).
Abbreviations: ACD, Airway closure duration; PAhold, Pharyngeal area when bolus held in oral cavity; PAmax, Pharyngeal area at the point of maximum pharyngeal constriction; PCR, Pharyngeal construction ratio; TPT, Total pharyngeal transit time.

line with previous studies, we found that the agreement between the two raters varied when analyzing normal-abnormal findings.[3,22,23] The best agreement between the subjective ratings was in pharyngeal residues and in the detection of penetration or aspiration. These are presumably the most important parameters reflecting the efficacy and safety of swallowing. The agreement between the objective measurements was stronger than that between the subjective ratings. The statistically significant correlations in the VFS findings were between TPT and pharyngeal phase abnormality, normal PCR and lack of penetration/aspiration, and normal PESmax and normal cricopharyngeal relaxation. Other objective measurements did not seem to correlate with the clinical VFS findings.

The agreement of subjective ratings in the oral and esophageal findings was mostly weak. These results might improve with training,

TABLE 7 Comparisons of objective measurements and subjective evaluation of cricopharyngeal relaxation ($n = 77$).

	Cricopharyngeal relaxation normal *n* /% within normal	Cricopharyngeal relaxation abnormal *n* /% within abnormal	*p* value
PESmax normal	57/98.3%	15/83.3%	
PESmax abnormal	1/1.7%	3/16.7%	**0.039***
POD normal	50/86.2%	14/77.8%	
POD abnormal	8/13.8%	4/22.2%	0.463
PCR normal	58/100%	16/88.9%	
PCR abnormal	0/0%	2/11.1%	0.054

Note: Statistically significant result is marked with an asterisk (*).
Abbreviations: PCR, Pharyngeal construction ratio; PESmax, Maximum pharyngoesophageal segment opening; POD, Pharyngoesophageal segment opening duration.

TABLE 8 Comparisons of objective timing measurements and pharyngeal construction ratio and penetration or aspiration findings ($n = 77$).

	No penetration or aspiration *n* /% within normal	Penetration or aspiration *n* /% within abnormal	*p* value
ACD normal	63/90.0%	7/100%	
ACD abnormal	7/10.0%	0/0%	1.000
POD normal	60/85.7%	5/71.4%	
POD abnormal	10/14.3%	2/28.6%	0.298
TPT normal	32/45.7%	2/28.6%	
TPT abnormal	38/54.3%	5/71.4%	0.455
PCR normal	69/98.6%	5/71.4%	
PCR abnormal	1/1.4%	2/28.6%	**0.021***

Note: Statistically significant result is marked with an asterisk (*).
Abbreviations: ACD, Airway closure duration; PCR, Pharyngeal construction ratio; POD, Pharyngoesophageal segment opening duration; TPT, Total pharyngeal transit time.

as concluded in a study in which the agreement was higher for semi-solid (moderately thick) consistencies than liquid ones and the highest after group discussions.[3] However, in our study, no difference between the agreement on the consistencies was evident. Our Kappa values, indicating better inter-rater reliability, were generally higher than those published by Stoeckli et al., especially in the pharyngeal evaluation.[23] Many findings, such as laryngeal elevation abnormality, which was not detected in any of our patients, may better indicate a risk of swallowing problems. In these small movements, objective, and preferably automated, measurements would be useful.

The objective measurements had fair to good inter-rater reliability. However, interpretation of the measurements requires expertise. Leonard and Kendall have extensively studied objective measurements during swallowing and introduced several connections to clinical practice. Pharyngeal dilatation, indicating an increase in PAhold and PAmax, has been associated with cricopharyngeal muscle dysfunction and Zenker's diverticulum.[24,25] Cricopharyngeal myotomy decreased PESmax and reduced PCR, but PAhold did not change.[25] When PESmax has been measured in patients with unilateral vocal fold paralysis, a significant decrease in PESmax and an increase in PCR have been reported, indicating pharyngeal weakness.[26] Prolonged TPT has shown to increase the risk of aspiration.[27–29] However, delayed TPT has also been reported in healthy elderly people.[30,31] PCR has been introduced to estimate pharyngeal strength and even predict aspiration.[18,32] In our study, the reproducibility of PCR was poor because PCR was calculated by dividing the fairly reproducible variable PAmax with PAhold. However, we did find a statistically significant connection between normal PCR and a lack of penetration/aspiration. As reported above, we found no statistically significant connection between TPT and penetration/aspiration, but we did find a connection between normal PCR and a lack of penetration/aspiration. In our study, normal PESmax seemed to indicate normal cricopharyngeal relaxation. However, we found no statistically significant connections between other measurements in the clinical findings, so whether these measurements could be useful in clinical practice remains unclear. Measurements might represent more of a risk to safety or impairment of swallow efficacy than actual clinical findings.

In the objective measures, we used the normative values published by Leonard and Kendall, that is ±2 SD from the normal mean.[18,19] Lee et al. used a normal/abnormal consideration of ±1 SD in their publication.[22] In our study, the agreements between the objective measurements' normality remained between minimal and moderate and seemed to differ from the results of Lee et al., although they only published a few of their individual κ-values for inter-rater correlations.[22] We also calculated the correlations in comparison to ±1 SD, but this did not significantly affect the results (data not shown).

The Swallowtail software integrates all the quantitative measurement variables with other spatial and temporal measurement tools. The software is easy to use, especially for temporal measurements, but most spatial measurements are also easy to measure in almost all other Picture Archiving and Communication Systems (PACS). Although the measurement of the seven variables mentioned was easy, their interpretation was difficult, because the Swallowtail software does not automatically provide normative value ranges for different variables. The results of the different measurements were difficult to link to clinically relevant subjective findings. For example, PCR and other pre- and intrapharyngeal phase measurements did not seem to correlate in post-swallow pyriform sinus residues, which might cause clinically relevant safety or efficacy problems in swallowing. These different pharyngeal phase variables also fail to take into consideration the potential clinically relevant esophageal phase abnormalities that could be the main culprit of dysphagia.

Automated analysis would improve the feasibility of the objective measurement analysis. Recently, new, promising deep learning analyses have been introduced to highly accurately detect penetration or aspiration in VFS in particular.[33] Other novel methods, such as automated impedance manometry with VFS, are likely to improve diagnostic performance. This method has high intra-rater and inter-rater reproducibility.[34,35] The Swallowtail software and objective measurements only detect pharyngeal phenomena. However, (oro)pharyngeal evaluation of swallowing alone is not enough, because esophageal etiologies often cause oropharyngeal symptoms.[36–41]

Our study had some limitations. The subjective raters did not discuss the evaluations nor were they trained together, which presumably affected the results, as also shown by Scott et al.[3] However, even well-trained evaluators have shown inaccurate agreements.[22] Moreover, they received no specific training in objective measurement, which has also shown to improve the accuracy of measurements.[42] Nevertheless, the ICC between our two experienced radiologists was in line with previously published agreements among trained speech and language pathologists.[42] Our results obtained by two radiologists with no joint training likely reflected a realistic clinical scenario and thus suggests reasonably good generalizability. This study also only examined the inter-rater agreement and not the actual accuracy of the results. Three or more raters would have been a good addition to the study. We did not assess intra-rater reproducibility.

5 | CONCLUSIONS

Subjective analysis of VFS resulted in moderate to strong agreement in the pharyngeal findings, especially concerning the detection of pharyngeal residues and penetration/aspiration, which are the most important findings when evaluating the efficacy and safety of swallowing. The objective spatial and temporal measurements had moderate to good inter-observer reproducibility and could therefore improve the reliability of VFS diagnostics. However, the clinical utility of objective measurements still needs further research.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Tero Vahlberg for their assistance with planning statistical analyses, Dominic A. Bellino for providing the Swallowtail software, and Professor Jérôme Lechien for providing the GLUPS score.

FUNDING INFORMATION

This study was funded by the Finnish ORL-HNS Foundation, the Finnish Society for Laryngology, the Turku University Foundation (the Kosti Hämmärö Fund), the Päivikki and Sakari Sohlberg Foundation, and Turku University Hospital Research Funds.

CONFLICT OF INTEREST STATEMENT

The authors have no other funding, financial relationships, or conflict of interest to disclose.

ORCID

Jonna Kuuskoski https://orcid.org/0000-0003-4614-2744

REFERENCES

1. Ekberg O, Hamdy S, Woisard V, Wuttge-Hannig A, Ortega P. Social and psychological burden of dysphagia: its impact on diagnosis and treatment. *Dysphagia*. 2002;17(2):139-146. doi:10.1007/s00455-001-0113-5
2. Langmore SE, Schatz K, Olsen N. Fiberoptic endoscopic examination of swallowing safety: a new procedure. *Dysphagia*. 1988;2(4):216-219. doi:10.1007/BF02414429
3. Scott A, Perry A, Bench J. A study of interrater reliability when using Videofluoroscopy as an assessment of swallowing. *Dysphagia*. 1998;13(4):223-227. doi:10.1007/PL00009576
4. McCullough GH, Wertz RT, Rosenbek JC, Mills RH, Webb WG, Ross KB. Inter- and Intrajudge reliability for Videofluoroscopic swallowing evaluation measures. *Dysphagia*. 2001;16(2):110-118. doi:10.1007/PL00021291
5. Kim J, Oh BM, Kim JY, Lee GJ, Lee SA, Han TR. Validation of the videofluoroscopic dysphagia scale in various etiologies. *Dysphagia*. 2014;29(4):438-443. doi:10.1007/s00455-014-9524-y
6. Chang MC, Lee C, Park D. Validation and inter-rater reliability of the modified videofluoroscopic dysphagia scale (mVDS) in dysphagic patients with multiple etiologies. *J Clin Med*. 2021;10(13):2990. doi:10.3390/jcm10132990
7. Lechien JR, Cavelier G, Thill MP, et al. Validity and reliability of the French version of eating assessment tool (EAT-10). *Eur Arch Otorhinolaryngol*. 2019;276(6):1727-1736. doi:10.1007/s00405-019-05429-1
8. Martin-Harris B, Brodsky MB, Michel Y, et al. MBS measurement tool for swallow impairment-MBSImp: establishing a standard. *Dysphagia*. 2008;23(4):392-405. doi:10.1007/s00455-008-9185-9
9. Chun SW, Lee SA, Jung IY, Beom J, Han TR, Oh BM. Inter-rater agreement for the clinical dysphagia scale. *Ann Rehabil Med*. 2011;35(4):470-476. doi:10.5535/arm.2011.35.4.470
10. Kendall KA. Timing of events in normal swallowing: a Videofluoroscopic study. *Dysphagia*. 2000;16(2):74-83. doi:10.1007/s004550010004
11. Leonard RJ, Kendall KA, McKenzie S, Gonçalves MI, Walker A. Structural displacements in Normal swallowing: a Videofluoroscopic study. *Dysphagia*. 2000;15(3):146-152. doi:10.1007/s004550010017
12. Cichero JAY, Lam P, Steele CM, et al. Development of international terminology and definitions for texture-modified foods and thickened fluids used in dysphagia management: the IDDSI framework. *Dysphagia*. 2017;32(2):293-314. doi:10.1007/s00455-016-9758-y

13. IDDSI. IDDSI - Home. Accessed March 23, 2024. https://iddsi.org/
14. Kuuskoski J, Vanhatalo J, Rekola J, Aaltonen LM, Järvenpää P. The water swallow test and EAT-10 as screening tools for referral to Videofluoroscopy. *Laryngoscope*. 2024;134(3):1349-1355. doi:10.1002/lary.31038
15. Regueiro MRKB, Nascimento WV, Parreira LC, Dantas RO. Videofluoroscopic analysis of different volumes of liquid bolus swallowing in healthy individuals: comparison between height and sex. *Clinics (Sao Paulo)*. 2017;72(11):693-697. doi:10.6061/clinics/2017(11)08
16. Park T, Kim Y, Ko DH, McCullough G. Initiation and duration of laryngeal closure during the pharyngeal swallow in post-stroke patients. *Dysphagia*. 2010;25(3):177-182. doi:10.1007/s00455-009-9237-9
17. Leonard R. Predicting aspiration risk in patients with dysphagia: evidence from fluoroscopy. *Laryngoscope Investig Otolaryngol*. 2019;4(1):83-88. doi:10.1002/lio2.226
18. Leonard R, Rees CJ, Belafsky P, Allen J. Fluoroscopic surrogate for pharyngeal strength: the pharyngeal constriction ratio (PCR). *Dysphagia*. 2011;26(1):13-17. doi:10.1007/s00455-009-9258-4
19. Leonard R, Kendall K. *Dysphagia Assessment and Treatment Planning: A Team Approach, Fourth Edition*. Plural Publishing, Incorporated; 2017 http://ebookcentral.proquest.com/lib/kutu/detail.action?docID=5509493
20. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-282.
21. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
22. Lee JW, Randall DR, Evangelista LM, Kuhn MA, Belafsky PC. Subjective assessment of Videofluoroscopic swallow studies. *Otolaryngology–Head and Neck Surgery*. 2017;156(5):901-905. doi:10.1177/0194599817691276
23. Stoeckli SJ, Huisman TAGM, Seifert BAGM, Martin Harris BJW. Interrater reliability of Videofluoroscopic swallow evaluation. *Dysphagia*. 2003;18(1):53-57. doi:10.1007/s00455-002-0085-0
24. Belafsky PC, Rees CJ, Allen J, Leonard RJ. Pharyngeal dilation in cricopharyngeus muscle dysfunction and Zenker diverticulum. *Laryngoscope*. 2010;120(5):889-894. doi:10.1002/lary.20874
25. Allen J, White CJ, Leonard R, Belafsky PC. Effect of cricopharyngeus muscle surgery on the pharynx. *Laryngoscope*. 2010;120(8):1498-1503. doi:10.1002/lary.21002
26. Domer AS, Leonard R, Belafsky PC. Pharyngeal weakness and upper esophageal sphincter opening in patients with unilateral vocal fold immobility. *Laryngoscope*. 2014;124(10):2371-2374. doi:10.1002/lary.24779
27. Johnson ER, McKenzie SW, Rosenquist CJ, Lieberman JS, Sievers AE. Dysphagia following stroke: quantitative evaluation of pharyngeal transit times. *Arch Phys Med Rehabil*. 1992;73(5):419-423.
28. Johnson ER, McKenzie SW, Sievers A. Aspiration pneumonia in stroke. *Arch Phys Med Rehabil*. 1993;74(9):973-976.
29. Rademaker AW, Pauloski BR, Logemann JA, Shanahan TK. Oropharyngeal swallow efficiency as a representative measure of swallowing function. *J Speech Hear Res*. 1994;37(2):314-325. doi:10.1044/jshr.3702.314
30. Butler SG, Maslan J, Stuart A, et al. Factors influencing bolus dwell times in healthy older adults assessed endoscopically. *Laryngoscope*. 2011;121(12):2526-2534. doi:10.1002/lary.22372
31. Kendall KA, Leonard RJ. Bolus transit and airway protection coordination in older dysphagic patients. *Laryngoscope*. 2001;111(11 Pt 1):2017-2021. doi:10.1097/00005537-200111000-00028
32. Yip H, Leonard R, Belafsky PC. Can a fluoroscopic estimation of pharyngeal constriction predict aspiration? *Otolaryngology–Head and Neck Surg*. 2006;135(2):215-217. doi:10.1016/j.otohns.2006.03.016
33. Kim JK, Choo YJ, Choi GS, Shin H, Chang MC, Park D. Deep learning analysis to automatically detect the presence of penetration or aspiration in Videofluoroscopic swallowing study. *J Korean Med Sci*. 2022;37(6):e42. doi:10.3346/jkms.2022.37.e42
34. Omari TI, Papathanasopoulos A, Dejaeger E, et al. Reproducibility and agreement of pharyngeal automated impedance manometry with Videofluoroscopy. *Clin Gastroenterol Hepatol*. 2011;9(10):862-867. doi:10.1016/j.cgh.2011.05.026
35. Szczesniak MM, Maclean J, Zhang T, et al. Inter-rater reliability and validity of automated impedance manometry analysis and fluoroscopy in dysphagic patients after head and neck cancer radiotherapy. *Neurogastroenterology & Motility*. 2015;27(8):1183-1189. doi:10.1111/nmo.12610
36. Jones B, Ravich WJ, Donner MW, Kramer SS, Hendrix TR. Pharyngoesophageal interrelationships: observations and working concepts. *Gastrointest Radiol*. 1985;10(1):225-233. doi:10.1007/BF01893105
37. Edwards DAW. History and symptoms of esophageal disease. *Diseases of the Esophagus*. Springer Verlag; 1974:103-118.
38. Gullung JL, Hill EG, Castell DO, Martin-Harris B. Oropharyngeal and esophageal swallowing impairments: their association and the predictive value of the modified barium swallow impairment profile and combined multichannel intraluminal impedance-esophageal manometry. *Ann Otol Rhinol Laryngol*. 2012;121(11):738-745. doi:10.1177/000348941212101107
39. Wilcox CM, Alexander LN, Clark WS. Localization of an obstructing esophageal lesion. Is the patient accurate? *Dig Dis Sci*. 1995;40(10):2192-2196. doi:10.1007/BF02209005
40. Smith DF, Ott DJ, Gelfand DW, Chen MY. Lower esophageal mucosal ring: correlation of referred symptoms with radiographic findings using a marshmallow bolus. *AJR Am J Roentgenol*. 1998;171(5):1361-1365. doi:10.2214/ajr.171.5.9798879
41. Madhavan A, Carnaby GD, Crary MA. "Food sticking in my throat": Videofluoroscopic evaluation of a common symptom. *Dysphagia*. 2015;30(3):343-348. doi:10.1007/s00455-015-9605-6
42. Nordin NA, Miles A, Allen J. Measuring competency development in objective evaluation of Videofluoroscopic swallowing studies. *Dysphagia*. 2017;32(3):427-436. doi:10.1007/s00455-016-9776-9

How to cite this article: Kuuskoski J, Vanhatalo J, Hirvonen J, Rekola J, Aaltonen L-M, Järvenpää P. Inter-rater reliability and clinical relevance of subjective and objective interpretation of videofluoroscopy findings. *Laryngoscope Investigative Otolaryngology*. 2024;9(4):e1298. doi:10.1002/lio2.1298