AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY FOUNDATION

Otolaryngology–Head and Neck Surgery
2014, Vol. 150(4) 538–541

DOI: 10.1177/0194599814521761
http://otojournal.org
SAGE

The Interrater and Intrarater Reliability of the Philpott-Javer Staging System Based on Level of Training

Harman S. Parhar[1], Andrew Thamboo, MD, MHSc[1], Al-Rahim Habib[1], Brent Chang, MD[1], Eng Cern Gan, MBBS, MRCS, MMED[1], and Amin R. Javer, MD, FRCSC, FARS[1]

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

Abstract

Objective. The Philpott-Javer postoperative endoscopic mucosal staging system for allergic fungal rhinosinusitis has previously demonstrated acceptable interrater reliability among rhinologists. There are, however, numerous learners involved in patient care at tertiary centers. This study aims to analyze the interrater and intrarater reliability of this system among learners in otolaryngology at different stages in training.

Study Design. A prospective analysis of retrospectively collected endoscopic photographs.

Setting. A tertiary care teaching hospital (January 2013).

Subjects. Fifty patients undergoing routine follow-up.

Method. Three photographs from each of 50 patients undergoing routine postsurgical nasoendoscopy were reviewed. Images were played twice, 1 week apart, in 2 differently randomized cycles and scored according to Philpott-Javer criteria by a rhinologist, a rhinology fellow, a senior otolaryngology resident, a junior otolaryngology resident, and a medical student. Interobserver reliability was assessed using the intraclass correlation coefficient, while intrarater reliability was assessed by Shrout-Fleiss κ values. Agreement between each learner and the rhinologist was also assessed using κ values.

Results. The interclass correlation among the 5 raters was 0.7600 (95% confidence interval, 0.6917-0.8161) for the Philpott-Javer scoring system, suggesting substantial reliability. Intrarater data showed substantial to almost-perfect reliability (κ values between 0.668 and 0.815) among all raters using this system. There was also moderate to substantial agreement between the learners and the rhinologist (κ values between 0.534 and 0.710).

Conclusion. Results suggest that the Philpott-Javer staging system has acceptable intrarater and interrater reliability among learners of differing levels of clinical experience and is suitable for evaluating progress following surgery.

Keywords

allergic fungal rhinosinusitis, endoscopy, endoscopic sinus surgery, residency training

Received September 18, 2013; revised November 11, 2013; accepted January 9, 2014.

Allergic fungal rhinosinusitis (AFRS), a subtype of chromic sinusitis with nasal polyposis that also has allergic mucin containing fungal hyphae, is a recalcitrant disease diagnosed using the Bent and Kuhn criteria (**Table 1**).[1-4] It requires aggressive and complete endoscopic sinus surgery to remove mucin and debris and allow for drainage.[5] Follow-up is essential to prevent polyp reformation and mucin reaccumulation.[6] Several scoring systems have been created to describe and communicate the patients' disease status and assess changes in a consistent manner. The Philpott-Javer postoperative endoscopic mucosal staging system (**Table 2**) was created to expand upon the existing Kupferberg postoperative mucosal scoring system and has been validated for the endoscopic follow-up of patients with AFRS postoperatively.[6] The Kupferberg scoring system scores the sinus cavities as a collective, based on 4 endoscopic findings—no mucosal edema, mucosal edema, polypoid edema, or sinus polyps—and on either the presence or absence of allergic mucin.[6] The more detailed Philpott-Javer system grades the mucosal edema, polypoid edema, and frank polyps as mild, moderate, or severe and allocates points to each sinus cavity as well as gives additional points for allergic mucin.[6]

At our center, patients are seen postoperatively to monitor for the possibility of polyp reformation and mucin

[1]St Paul's Sinus Centre, Vancouver, British Columbia, Canada

This article was presented at the 2013 AAO-HNSF Annual Meeting & OTO EXPO; September 29–October 3, 2013; Vancouver, British Columbia, Canada.

Corresponding Author:
Harman S. Parhar, St Paul's Sinus Centre, 1081 Burrard St, Vancouver, BC, V6Z1Y6, Canada.
Email: harman.parhar@gmail.com

Table 1. Bent and Kuhn Diagnostic Criteria for Allergic Fungal Rhinosinusitis.

1. Type I hypersensitivity confirmed by history, skin tests, or serology
2. Nasal polyposis
3. Characteristic computed tomography scan (double density sign)
4. Eosinophilic mucus without fungal invasion into sinus tissue
5. Positive fungal stain of sinus contents removed intraoperatively or during office endoscopy
6. Immunocompetence (replaces number 1 at St Paul's Sinus Centre)

Table 2. Philpott-javer Endoscopic Staging System for Allergic Fungal Rhinosinusitis.[a]

Grading	State of Mucosa
0	No edema
1-3	Mucosal edema (mild/moderate/severe)
4-6	Polypoid edema (mild/moderate/severe)
7-9	Frank polyps (mild/moderate/severe)

[a]+1 Point for each sinus that contains mucin.

accumulation. It is, however, unclear whether the routine postsurgical endoscopic evaluations have interrater and intrarater reliability. Smith et al[7] recently published work demonstrating acceptable interrater reliability based on a number of endoscopic scoring parameters (middle turbinate position, synechiae/adhesions, inflammation, and crusting), but we wanted to evaluate this on the standardized Philpott-Javer system. In most tertiary care centers, a variety of learners (fellows, residents, and medical students) are involved in the care of patients. This study aims to analyze both the interrater and intrarater reliability of the Philpott-Javer scoring system among learners in otolaryngology at different stages to gain insight into whether this system can be used reliably by learners who are becoming increasingly involved in patient care as they gain experience. In addition, if the different learner evaluators were found to score reliably in this study when compared with each other, it would suggest reliable scoring among the various learners who regularly rotate in and out of the rhinology service. We suspect, based on our anecdotal experience thus far, that the learners will achieve reasonable reliability when using this scoring system.

Materials and Methods

This study was approved by the University of British Columbia ethical review committee. Fifty consecutive patients, who had all previously undergone complete bilateral endoscopic sinus surgery and visited our center in January 2013 for follow-up, were included in the study. All photographs examined came from patients previously diagnosed with AFRS according to the Bent and Kuhn diagnostic criteria. To begin, 3 photographs of each of the 50 patients (150 photographs total) were randomly selected from the routine set of nasoendoscopic images (bilateral ethmoid, frontal, sphenoid, and maxillary sinuses) taken during standard postoperative rigid nasoendoscopy. The photographs were coded and randomized in 2 differently randomized cycles (the same 150 photographs in each cycle but in a different order) by a nonclinical investigator and played in a loop for anonymous review. Evaluators were blinded to the identity of the patients, and none of the images showed regions external to the nostrils. Each of the differently randomized cycles was played once, 1 week apart, and, with a descriptive chart as a guide (**Figure 1**), scored according to Philpott-Javer criteria by a staff rhinologist, a rhinology fellow, a senior otolaryngology resident, a junior otolaryngology resident, and a medical student.

Interobserver variability was assessed using the intraclass correlation coefficient while intrarater reliability was assessed by Fleiss κ values. Agreement between each learner and the staff rhinologist was also assessed using κ values.

Results

A total of 150 photographs collected from 50 subjects were pooled and analyzed together in 2 different sittings 1 week apart. The interrater reliability of the 5 raters was assessed using the interclass correlation coefficient (ICC). On the first sitting, the ICC was 0.6553 (95% confidence interval [CI], 0.5545-0.7374) for the Philpott-Javer scoring system, indicating substantial agreement. The second sitting showed an ICC of 0.7600 (95% CI, 0.6917-0.8161), suggesting improved reliability.

Intrarater data were assessed by comparing each learner's scores in the 2 sittings. This was quantified by calculating the κ values. Overall, all participants showed substantial reliability with κ values between 0.668 and 0.815 among all raters. In particular, the Philpott-Javer system had κ values of 0.718 (95% CI, 0.659-0.776), 0.732 (95% CI, 0.661-0.803), 0.700 (95% CI, 0.636-0.764), 0.815 (95% CI, 0.770-0.861), and 0.668 (95% CI, 0.600-0.737) achieved by the staff rhinologist, rhinology fellow, senior resident, junior resident, and medical student, respectively.

Both interrater and intrarater reliability were assessed as described above, but we were also curious to see how each learner would compare with the staff rhinologist with years of experience evaluating postoperative patients. To do so, κ values were calculated comparing each learner with the rhinologist for Philpott-Javer scoring. Again, there was moderate to substantial agreement demonstrated by κ values between 0.534 and 0.710 for all learners. In particular, the κ scores were 0.580 (95% CI, 0.504-0.656), 0.670 (95% CI, 0.605-0.736), 0.710 (95% CI, 0.652-0.767), and 0.534 (95% CI, 0.456-0.612) for the rhinology fellow, senior resident, junior resident, and medical student, respectively.

Discussion

It is known that patients with AFRS benefit from close follow-up postoperatively to monitor disease progression

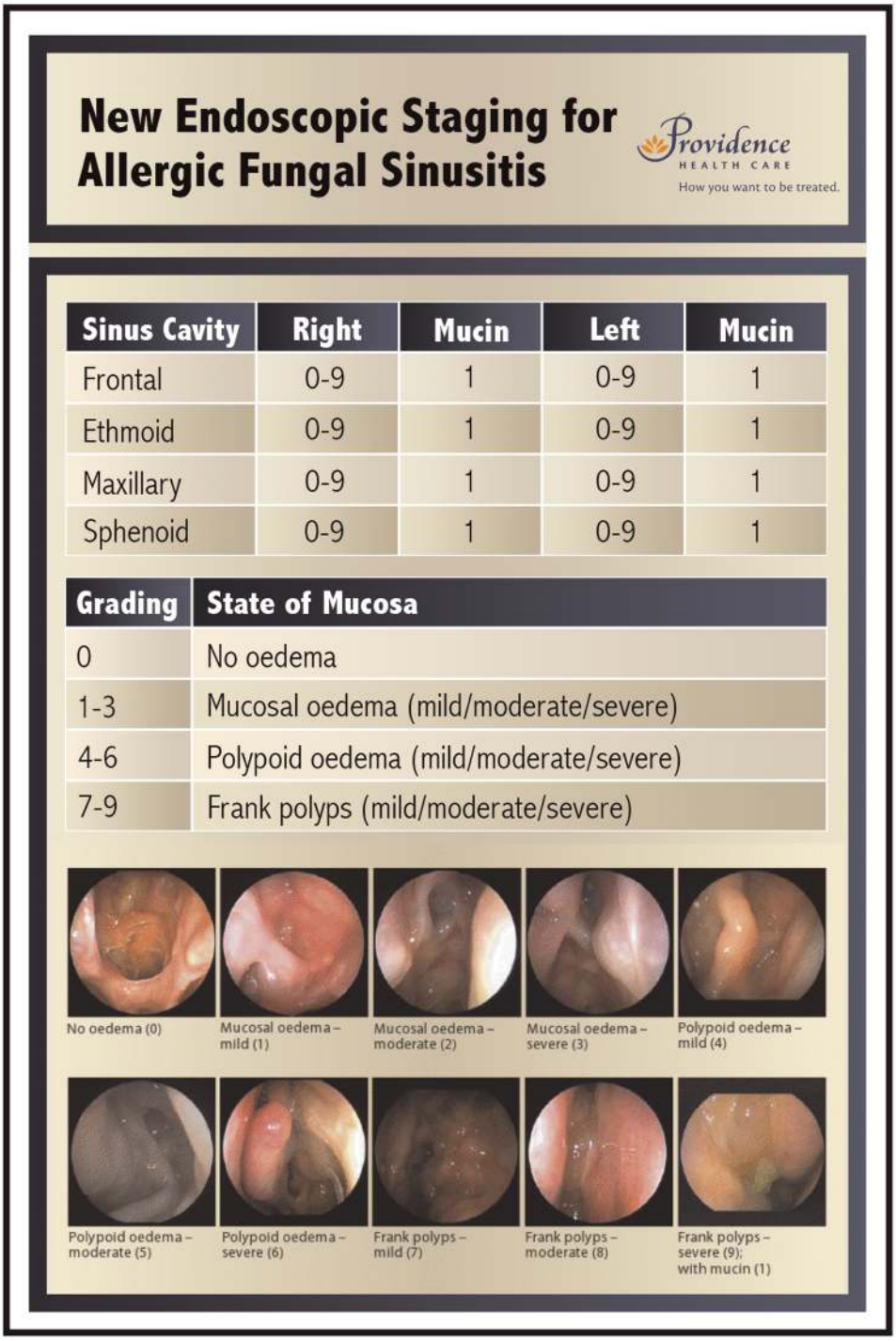

Figure 1. Wall-mounted poster showing the Philpott-Javer endoscopic mucosal staging system. Each sinus cavity is graded as having mucosal edema, polypoid edema, or frank polyps and further subdivided as being mild, moderate, or severe. An additional point is given for the visual appearance of allergic mucin.

and guide management. At our center, patients are seen at 6- to 12-week intervals. Like most academic tertiary care teaching hospitals, numerous learners, including local and visiting medical students, residents in otolaryngology and other fields, and clinical fellows who are often international, are involved in the care of patients. While all patient care falls under the close supervision of an experienced staff rhinologist, the learners tend to enjoy becoming active in disease evaluation and management. Prior to the implementation of standardized systems, such as the Philpott-Javer endoscopic mucosal staging system, it was very difficult to assess and communicate disease progression reliably. Anecdotal evidence at this center supported the notion that learners were reasonably good at being able to use this

scoring system, but this study was our attempt to actually investigate and quantify this notion.

Interrater reliability among the 5 raters of differing clinical experience was good using the Philpott-Javer scoring system. This has been demonstrated previously among staff rhinologists but never among less experienced learners.[6] Interestingly, the interrater reliability increased even between the first and second sittings from 0.6553 to 0.7600 for the Philpott-Javer scoring system. Although this is only 1 trial, it speaks to how learnable the system may be. Intrarater reliability was also very good, having correlations of between 0.668 and 0.815 for all evaluators using the system. Of note, the more experienced evaluators did not always do better than those with less clinical experience, and there appeared to be no proportional increase in reliability with more training. An additional interest of ours was how well the learners could do compared with the most experienced member of the team. Correlations between the staff rhinologist and the learners again showed moderate to substantial agreement with values of 0.580, 0.670, 0.710, and 0.534 for the rhinology fellow, senior resident, junior resident, and medical student, respectively. First, this suggests that the learners had reasonable reliability in scoring the patients compared with the staff rhinologist, who is usually the individual responsible to do so. This also suggests that the staff rhinologist does not have to modify learner scores in the patient record very often.

A primary objective we had when beginning this study was to establish whether the various learners in our center were reliably staging patient disease progression. Each patient examination room at our center has a wall-mounted chart (**Figure 1**) to guide team members when evaluating postoperative AFRS patients. These charts show a single-example photograph of each stage of Philpott-Javer scoring along with written descriptors. These were the same charts made available to evaluators in this study. Given the moderate to substantial reliability achieved and discussed above, it would appear that this system is relatively easy to learn and apply even by those with very little exposure and experience in the field. As learners are becoming increasingly involved in patient care, especially in academic teaching hospitals, we believe simple tools such as the Philpott-Javer system can help learners discover the course of AFRS and allow them to contribute to patient evaluation in a reliable manner.

Future work in this area would strengthen our conclusions. Our study showed that there was no proportional increase in reliability with more training, but it remains unclear whether there is a certain amount of training above which increases in reliability become insignificant. In addition, our second trial showed improvements in the interrater reliability over our first, particularly among less experienced evaluators, and it would be interesting to see how much improvement we might see with more trials to assess the learnability of the system. Our study involved a modest number of self-selected evaluators, and while we believe they were representative individuals for their levels of training, it remains unclear whether the same results would have been obtained with multiple raters of the same level of training. Last, our study used a set of still photographs, but future studies might use real-time videos in their assessments to better replicate the clinical setting.

Conclusions

This article demonstrates that the Philpott-Javer scoring system has acceptable interrater and intrarater reliability for scoring the disease progression of postsurgical AFRS patients even when used by evaluators of differing levels of clinical experience. Moreover, these results suggest that this system is relatively easy to learn and apply in a reliable manner even for learners.

Acknowledgment

We acknowledge the contributions of Dr C. Philpott for his efforts in developing the scoring system evaluated in this study.

Author Contributions

Harman S. Parhar, research design, clinical evaluation, data analysis, manuscript preparation; **Andrew Thamboo**, research design, clinical evaluation, data analysis, manuscript review; **Al-Rahim Habib**, research design, experiment setup, data analysis, manuscript review; **Brent Chang**, research design, clinical evaluation, manuscript review; **Eng Cern Gan**, research design, clinical evaluation, manuscript review; **Amin R. Javer**, research design, group supervision, clinical evaluation, manuscript review.

Disclosures

Competing interests: Amin R. Javer is a consultant and/or speaker for Honeydoc, Merck, Takeda, Sinusys, and Laurimed.

Sponsorships: None.

Funding source: None.

References

1. Scadding GK, Durham SR, Mirakian R, et al. BSACI guidelines for the management of rhinosinusitis and nasal polyposis. *Clin Exp Allergy*. 2008;38:260-275.
2. Hutcheson PS, Schubert MS, Slavin RG. Distinctions between allergic fungal rhinosinusitis and chronic rhinosinusitis. *Am J Rhinol Allergy*. 2010;24:405-408.
3. Bent JP III, Kuhn FA. Diagnosis of allergic fungal sinusitis. *Otolaryngol Head Neck Surg*. 1994;111:580-588.
4. Kuhn FA, Swain R. Allergic fungal sinusitis: diagnosis and treatment. *Curr Opin Otolaryngol Head Neck Surg*. 2003;11:1-5.
5. Glass D, Amedee AG. Allergic fungal rhinosinusitis: a review. *Ochsner J*. 2011;11:271-275.
6. Philpott CM, Clark A, Javer AR. Allergic fungal rhinosinusitis—a new staging system. *Rhinology*. 2011;49:318-323.
7. Smith TL, Hwang PH, Murr AH, et al. Interrater reliability of endoscopic parameters following sinus surgery. *Laryngoscope*. 2012;122:230-236.