Original Contribution

Interexaminer reliability of pharyngeal injection and palatine tonsillar hypertrophy in a pediatric emergency department

Soyun Hwang, Jae Yun Jung *, Joong Wan Park, Do Kyun Kim, Young Ho Kwak

Department of Emergency Medicine, Seoul National University Hospital, 101 Daehak-Ro Jongno-Gu, Seoul 03080, Republic of Korea

ARTICLE INFO

ABSTRACT

Objectives: To evaluate the interrater reliability of throat examinations in children according to the major and training stage.
Study design: We performed a prospective observational study of interrater reliability. The participants included physicians with various amounts of experience and majors who were working in an urban, tertiary hospital. We collected 20 photos of the throats of children who presented to the pediatric emergency department (ED) and performed 2 surveys (with or without medical history). The primary outcome was the interrater agreement for pharyngeal injection (PI) and palatine tonsillar hypertrophy (PTH), and the secondary outcome was the interrater agreement for PI and PTH in subgroups of examiners divided by major and duration of clinical experience.
Results: Thirty-three examiners participated in this study. The overall percent agreement for PI was 0.669, and Fleiss' kappa was 0.296. The interrater reliability was similar before and after providing patients' medical history. The overall percent agreement for PTH was 0.408, and Kendall's W was 0.674. When the patients' medical history was provided, Kendall's W increased (0.692). In the subgroup analysis, Fleiss' kappa for PI ranged from 0.257 to 0.33, and Kendall's W for PTH ranged from 0.593 to 0.711.
Conclusion: Examiners' agreement for PTH was more reliable than that for PI when evaluating children who visited the ED. The interrater reliability did not improve with increased clinical experience. These findings should be considered in the examination of pharyngeal pathology.

© 2019 Elsevier Inc. All rights reserved.

1. Introduction

In the emergency department (ED), patient flow is mainly determined by physical examination findings, and the reliability and reproducibility of these findings is very important [1,2]. However, physical examinations can be subjective. In training hospitals in particular, interpretation of physical examination findings can differ based on the examiner's level of clinical experience, which depends on the training stage, so treatment directions may vary by examiner.

Unlike in adult populations, in pediatric patients, physicians tend to minimize unnecessary radiographic imaging and laboratory tests. Radiography tends to be performed as little as possible because of the risk of radiation exposure [3,4], and blood and urine tests are expensive, time consuming and invasive [5]. In short, accurate physical examination is essential for accurate decision making to practice pediatric emergency medicine (EM).

In some previous studies, the interrater reliability of abdominal examinations in children who visited the ED with abdominal pain was evaluated, and the results showed that physical examination findings were different according to the examiner's major and training stage [6-8]. Another study evaluated the inter- and intrarater reliability of the overall clinical appearance of febrile infants and showed modest agreement; the provider's level of experience had little effect on agreement [9]. However, few studies have evaluated the interrater reliability of physical examinations in the pediatric ED, and further studies are necessary to provide quality control for physicians' physical examinations. Thus, improvements in standardization of physical examination methods are necessary.

Fever is the most common chief complaint of infants and children in the pediatric ED. Every child who presents to the ED with fever undergoes a physical examination to identify the cause of the fever. In particular, pharyngitis and tonsillitis are common causes of fever in children, and blood and urine tests are not required in addition to physical examination in most cases [10,11]. Therefore, the proper diagnosis of pharyngitis and tonsillitis may prevent unnecessary testing. Furthermore, if pharyngitis and tonsillitis are definitely excluded, appropriate screening for the correct diagnosis is indicated.

In this study, we evaluated the interrater reliability of pharyngeal injection (PI) and palatine tonsillar hypertrophy (PTH) in children according to each examiner's major (emergency medicine (EM) or pediatrics) and training stage using photographs of the throats of children who visited the pediatric ED.

2. Methods

2.1. Study design and setting

This was a prospective observational study of interrater reliability. The study was conducted from August 2017 to October 2017. This study was performed at an urban, tertiary teaching hospital ED with residencies and fellowships in EM and pediatrics. This hospital also has a distinctive pediatric EM fellowship and faculty. This study was approved by Institutional Review Board of Seoul National University Hospital (IRB No. 1706-188-864). Our IRB did not require written consent from the study participants.

2.2. Study participants and sample size

The participants included residents, fellows and faculty in EM and pediatrics. We recruited eleven EM residents, eleven board-certified general and pediatric EM physicians, and eleven pediatrics residents or board-certified pediatricians.

For this interrater reliability study, we assumed that if the relative error was 30%, and the overall agreement was 70%, 20 subjects were needed [12]. Additionally, if the desired coefficient of variation was 20%, the required number of raters was 10 for each group [13]. Considering the subgroup analysis and possible drop-outs, we recruited a total of 33 doctors including eleven physicians in each group.

2.3. Study protocol

We extracted medical photos of the throats of children who presented to the pediatric ED from the electronic medical record (EMR) system. These photos were taken by EM residents with a small endoscope camera and uploaded to the EMR. Twenty photos were selected, and the research personnel developed Google survey forms. The photos were selected to include various ages of patients (from 18-month-old to 18-year-old patients) and clinical severities (from grade 0 to grade 4). The photos were selected by 2 board-certified pediatric emergency physicians, both with >10 years of clinical experiences in pediatric EM. These 2 physicians did not participate in the survey. They ensured that photos of various grades of PTH, as assessed by a previously standardized system for evaluating tonsillar size, were included [14]. However, since there were no clear criteria for PI, photos were selected based on the medical experience of these two physicians, and the principal investigator determined which photo to include if the two physicians did not agree on a single photo.

There were two sets of Google forms: set A or set B. Set A contained 20 photos of children's throats, and for each photo, the sex and age of the child was provided followed by 2 questions. The first question was regarding the presence of PI, and the answer was either yes or no. The second question was regarding presence and severity of PTH, and the answer ranged from grade 0 to grade 4. Set B was similar to set A, but some clinical information was also provided, such as the duration of fever and whether the patient complained of a sore throat. The order of the photos was not different between the two sets. Other basic characteristics of the examiners were also collected including their majors (EM, pediatrics, or both) and the number of years of clinical experience (set A: https://goo.gl/forms/8PkhVtVU1c9TSVHH3; set B: https://goo.gl/forms/a2h0HDFcYDbjhGBP2).

When eligible participants were identified, the research personnel recruited the participant, and the Google survey form was sent via e-mail. Set A was sent first, and then when the participant

answered the survey, set B was sent three days after their response. The participants were encouraged to answer the survey after their duty and not to discuss the answers with other people.

2.4. Outcomes

Our primary outcome was the interrater agreement of the examiners regarding PI and PTH. Our secondary outcome was the interrater agreement regarding PI and PTH in subgroups of examiners that were divided according to their majors and duration of clinical experience (residents vs board-certified physicians).

2.5. Statistical methods

Data were entered into an Excel spreadsheet (Microsoft, 2016), and analysis was performed with STATA (version 14.0, STATA Corp., College Station, TX, USA). Proportions were calculated for categorical variables. For interrater agreement regarding PI, Fleiss' kappa coefficient was mainly used because this was a nominal variable in the reproducibility test, and three or more assessors were compared [15,16]. For interrater agreement regarding PTH, Kendall's coefficient of concordance (Kendall's W) was used because this was an ordinal variable (grade 0 to 4) in the reproducibility test [17,18]. Both Fleiss' kappa coefficient and Kendall's W range from −1 to +1, and +1 indicates perfect agreement. Descriptive terms, such as 'poor agreement' and 'moderate agreement', were also used, according to previously published studies [19,20] (Supplemental Table 1). We additionally calculated the percent agreement and Gwet's first-order agreement coefficient (Gwet's AC1) for interrater agreement regarding PI.

3. Results

3.1. Participant characteristics

There were total 33 examiners who participated in this study. The distribution of their majors and duration of clinical experience in years is shown in Table 1. There were 20 EM physicians, 11 pediatricians and 2 double-boarded physicians (both in EM and pediatrics; these physicians were considered EM physicians in our analysis because they were currently working in the ED as emergency physicians). The majority of physicians had >5 years but <10 years of clinical experience. Ten doctors had >3 years but <5 years of clinical experience, and four doctors had >1 year but <3 years of clinical experience. Two doctors had >10 years of experience. The distribution of clinical experience in years in each subgroup is shown in Table 2. No participants in this study were colorblind.

3.2. Interrater reliabilities for PI and PTH

The interrater reliabilities for PI and PTH is shown in Table 3, respectively. The overall percent agreement for PI was 0.669 (substantial agreement), and Fleiss' kappa was 0.296, indicating fair agreement. When Gwet's AC1 was calculated, the overall agreement for PI was slightly increased (0.377) but still indicated fair

Table 1
Participants characteristics.

Total	N	%
	33	
Major of the participants		
PED	11	33.33
EM	20	60.61
EM + PED	2	6.06
Clinical experiences		
1–3 years	4	12.12
3–5 years	10	30.3
5–10 years	17	51.52
>10 years	2	6.06

Table 2
Distribution of clinical experience of each subgroup.

	EM residents	EM boards	PED
1–3 years	3	0	1
3–5 years	6	0	4
5–10 years	2	9	6
>10 years	0	2	0

agreement. In cases of PI, the interrater reliability was similar before and after the patients' medical history was provided.

The overall percent agreement for PTH was 0.408 (moderate agreement), which was lower than that for PI. Kendall's W was 0.674 (substantial agreement). When the patients' medical history was provided, Kendall's W increased (0.692).

3.3. Subgroup analysis of interrater reliability

We analyzed the interrater reliability in each subgroup (Table 4). For PI, Fleiss' kappa for EM residents, board-certified EM physicians, and pediatricians was 0.289, 0.332 and 0.264, respectively, when the patients' medical history was not provided. These results demonstrated fair agreement, despite small differences among the three subgroups. After the patients' medical history was provided, Fleiss' kappa for EM residents and pediatricians decreased slightly (0.272 and 0.257, respectively) but improved a little in board-certified EM physicians; however, all the groups still had fair agreement.

For PTH, Kendall's W of each subgroup showed substantial agreement (0.690, 0.699 and 0.711) when the patients' medical history was not provided. When the patients' medical history was provided, the agreement improved among EM residents and board-certified EM physicians but decreased among pediatricians from substantial agreement (0.711) to moderate agreement (0.593).

4. Discussion

In this study, we evaluated the interrater reliabilities for PI and PTH in children among EM physicians and pediatricians. The agreement for PI was generally fair, and the agreement for PTH was substantial in our analysis. The agreement for PTH increased slightly when the patients' medical history was provided, and it remained substantial. In the subgroup analyses, the results were generally similar: the agreement for PI was fair, but the agreement for PTH was substantial, with the exception of moderate agreement among pediatricians when the patients' medical history was provided.

In previous studies assessing interrater reliability, the agreement for physical examination was fair to moderate, and slight agreement was reported for pediatric abdominal examination [6,8,21]. No studies have previously evaluated the interrater reliability for throat examinations in children, but the agreement seems to be similar to that of abdominal examination. In previous studies, some authors mentioned that the low agreement regarding physical examination might be the result of differences in the examiners' training stages and majors [8]. However, other studies have showed no improvement in interrater reliability with increasing experience with clinical assessments [9,22]. Those previous studies evaluated gestalt impressions of overall clinical appearance [9] and gut feelings about serious infections [22]. There are no clearly predefined guidelines or gold standards for 'overall clinical appear-

Table 4
Subgroup analysis of inter-rater reliability of PI and PTH.

		N	without history	with history
PI	Total	33	0.298 (0.189–0.407)	0.292 (0.148–0.436)
	EM resident	11	0.289 (0.166–0.413)	0.272 (0.112–0.432)
	EM board	11	0.332 (0.174–0.490)	0.370 (0.178–0.563)
	PED	11	0.264 (0.116–0.413)	0.257 (0.116–0.398)
PTH	Total	33	0.692 (N/A)	0.680 (N/A)
	EM resident	11	0.690 (N/A)	0.752 (N/A)
	EM board	11	0.699 (N/A)	0.754 (N/A)
	PED	11	0.711 (N/A)	0.593 (N/A)

ance' or 'gut feelings' in contrast to abdominal examination, which has relatively more established criteria. Our study showed no differences in agreement between examiners regarding throat examination according to their training stages and majors, which is similar to the findings of the abovementioned studies. This result may be because clinicians may not have clear definitions of PI and PTH in their minds; thus, more education about throat examinations in children is necessary.

The doctors who participated in this study seemed to agree less on PI than on PTH. While there were only two choices in case of PI (yes or no), there were five grades for PTH; thus, this result was different from the expectation. The possible reason for this unexpected result is that there is a well-known and widely used classification system for PTH [14], while there is no established definition for PI. Thus, evaluation of PI is more subjective to each individual's perspectives than evaluation of PTH. Even among people who are not colorblind, perceptions of the degree of 'redness' could be different. It is known that when discrimination is clear, better agreement can be obtained [20].

There are some limitations in this study. First, this study was a single-centered study and included only EM physicians and pediatricians in a training hospital. However, we tried to enroll variety of participants with different majors and durations of clinical experience. Thus, these results may be applicable to other hospitals or medical providers.

Additionally, this study did not measure a clinical endpoint, such as antibiotic prescriptions. Most cases of acute pharyngitis in children do not requires antibiotics because they are caused by viral organisms [10,23], but in cases of group A streptococcal pharyngitis, antibiotics are indicated [10]; thus, it is important to differentiate streptococcal pharyngitis from benign, self-limiting viral pharyngitis by physical examination. However, differentiation of streptococcal pharyngitis from viral pharyngitis includes assessment of other physical examination findings, such as cervical lymphadenopathy [11,24], and our survey did not include this information. Therefore, it would have been difficult and meaningless to have the examiners determine whether to prescribe antibiotics solely based on a single photograph of throat. In addition, our study primarily emphasizes the interrater agreement, rather than the intrarater agreement. Because the internal threshold for prescribing antibiotics may vary from physician to physician depending on the experiences and available resources, measuring the clinical endpoint would be less meaningful.

Although this study has some limitations, our results indicated that physical exam itself can be subjective depending on the examiner. This problem may be caused by individual differences in both perception of the physical exam finding and description of it. Furthermore, this low reliability of physical examination cannot be

Table 3
The inter-rater reliability of PI and PTH.

	PI			PTH	
	Percent agreement	Fleiss' kappa	Gwet's AC1	Percent agreement	Kendall's W
Overall	0.669 (0.622–0.717)	0.296 (0.210–0.381)	0.377 (0.252–0.502)	0.408 (0.384–0.432)	0.674 (N/A)
Without history	0.666 (0.603–0.729)	0.298 (0.189–0.407)	0.362 (0.191–0.533)	0.427 (0.392–0.463)	0.692 (N/A)
With history	0.673 (0.596–0.750)	0.292 (0.148–0.436)	0.392 (0.194–0.590)	0.389 (0.355–0.423)	0.680 (N/A)

limited to the visual examination alone. Auscultation finding were also inconsistent among physicians [25], the palpable ratings may not be reliable [26], and interrater reliability of olfactory and taste sense was only moderate to good [27].

However, further efforts are needed to overcome this discrepancy of the physical examination. Proper training for each physical examination is important, but the communication of clinical findings also can be improved. One possible way to improve communication is the use of more specific expressions in describing physical examination findings. For example, in one study regarding a new grading scale for gross hematuria, the authors introduced a more specific grading scale using CYMK color codes [28]. This study showed excellent agreement among the urologists as well as the laypeople because of an objective and easy-to-use grading tool.

Another breakthrough may come from the evolution of technology. Due to advances in examination room equipment and medical recording systems, communication can be augmented via audio and video media, without relying solely on writing. Although there is no international standard for video and audio media, it will be helpful for more accurate delivery and evaluation of the patient's clinical findings if the medical record technology using supplementary media becomes more generalized in the future.

5. Conclusions

In conclusion, among children visiting the ED, the interrater reliability for PI was fair, and that for PTH was good. The interrater reliability did not improve with increased clinical experience. These findings should be considered in the examination of pharyngeal pathology.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ajem.2019.01.016.

Submission declaration

All authors declare that our paper is not published or under consideration for publishing elsewhere.

Declaration of interest

All authors disclose any financial and personal relationships with other people or organizations that could inappropriately influence their work.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] McCarthy PL, Lembo RM, Fink HD, Baron MA, Cicchetti DV. Observation, history, and physical examination in diagnosis of serious illnesses in febrile children less than or equal to 24 months. J Pediatr 1987;110(1):26–30. Epub 1987/01/01. PubMed PMID: 3540248.

[2] Reynolds SL, Jaffe DM. Diagnosing abdominal pain in a pediatric emergency department. Pediatr Emerg Care 1992;8(3):126–128. Epub 1992/06/01. PubMed PMID: 1614900.

[3] Kwon H, Jung JY. Effectiveness of a radiation reduction campaign targeting children with gastrointestinal symptoms in a pediatric emergency department. Medicine (Baltimore). 2017;96(3):e5907. Epub 2017/01/19. https://doi.org/10.1097/MD.0000000000005907. PubMed PMID: 28099351; PubMed Central PMCID: PMCPMC5279096.

[4] Jennings RM, Burtner JJ, Pellicer JF, Nair DK, Bradford MC, Shaffer M, et al. Reducing head CT use for children with head injuries in a community emergency department. Pediatrics. 2017;139(4). Epub 2017/03/04. https://doi.org/10.1542/peds.2016-1349. PubMed PMID: 28255067.

[5] Ouellet-Pelletier J, Guimont C, Gauthier M, Gravel J. Adverse events following diagnostic urethral catheterization in the pediatric emergency department.

[6] Kharbanda AB, Fishman SJ, Bachur RG. Comparison of pediatric emergency physicians' and surgeons' evaluation and diagnosis of appendicitis. Acad Emerg Med 2008;15(2):119–125. https://doi.org/10.1111/j.1553-2712.2008.00029.x. PubMed PMID: 18275440.

[7] Hunter BR, Seupaul RA. Interrater reliability of history and physical examination is limited among children with possible appendicitis. J Pediatr 2012;161(3):566. https://doi.org/10.1016/j.jpeds.2012.07.005. PubMed PMID: 22916978.

[8] Kharbanda AB, Stevenson MD, Macias CG, Sinclair K, Dudley NC, Bennett J, et al. Interrater reliability of clinical findings in children with possible appendicitis. Pediatrics. 2012;129(4):695–700. https://doi.org/10.1542/peds.2011-2037. PubMed PMID: 22392173; PubMed Central PMCID: PMCPMC3313636.

[9] Walsh P, Thornton J, Asato J, Walker N, McCoy G, Baal J, et al. Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. PeerJ. 2014;2:e651. https://doi.org/10.7717/peerj.651. PubMed PMID: 25401054; PubMed Central PMCID: PMCPMC4230550.

[10] Bisno AL. Acute pharyngitis. N Engl J Med 2001;344(3):205–211. Epub 2001/02/15. https://doi.org/10.1056/NEJM200101183440308. PubMed PMID: 11172144.

[11] Bisno AL, Gerber MA, Gwaltney JM, Jr., Kaplan EL, Schwartz RH, Infectious Diseases Society of A. Practice guidelines for the diagnosis and management of group A streptococcal pharyngitis. Infectious Diseases Society of America. Clin Infect Dis 2002;35(2):113–125. Epub 2002/06/28. https://doi.org/10.1086/340949. PubMed PMID: 12087516.

[12] Cantor AB. Sample-size calculations for Cohen's kappa. Psychol Methods 1996;1(2):150.

[13] Gwet KL. Handbook of inter-rater reliability. 3rd ed. Maryland, USA: Advanced Analytics, LLC; 2012.

[14] Brodsky L. Modern assessment of tonsils and adenoids. Pediatr Clin N Am 1989;36(6):1551–1569. Epub 1989/12/01. PubMed PMID: 2685730.

[15] Cicchetti DV, Volkmar F, Sparrow SS, Cohen D, Fermanian J, Rourke BP. Assessing the reliability of clinical scales when the data have both nominal and ordinal features: proposed guidelines for neuropsychological assessments. J Clin Exp Neuropsychol 1992;14(5):673–686. Epub 1992/09/01. https://doi.org/10.1080/01688639208402855. PubMed PMID: 1474138.

[16] Rucker G, Schimek-Jasch T, Nestle U. Measuring inter-observer agreement in contour delineation of medical imaging in a dummy run using Fleiss' kappa. Methods Inf Med 2012;51(6):489–494. Epub 2012/11/20. https://doi.org/10.3414/ME12-01-0005. PubMed PMID: 23160666.

[17] Bottcher HF, Posthoff C. Mathematical treatment of rank correlation–a comparison of Spearman's and Kendall's coefficients. Z Psychol Z Angew Psychol 1975;183(2):201–217. Epub 1975/01/01. PubMed PMID: 3056.

[18] Baumgartner R, Somorjai R, Summers R, Richter W. Assessment of cluster homogeneity in fMRI data using Kendall's coefficient of concordance. Magn Reson Imaging 1999;17(10):1525–1532. Epub 1999/12/28. PubMed PMID: 10610002.

[19] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174. Epub 1977/03/01. PubMed PMID: 843571.

[20] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276–282. Epub 2012/10/25. PubMed PMID: 23092060; PubMed Central PMCID: PMCPMC3900052.

[21] Yen K, Karpas A, Pinkerton HJ, Gorelick MH. Interexaminer reliability in physical examination of pediatric patients with abdominal pain. Arch Pediatr Adolesc Med 2005;159(4):373–376. https://doi.org/10.1001/archpedi.159.4.373. PubMed PMID: 15809393.

[22] Van den Bruel A, Thompson M, Buntinx F, Mant D. Clinicians' gut feeling about serious infections in children: observational study. BMJ. 2012;345:e6144. Epub 2012/09/28. https://doi.org/10.1136/bmj.e6144. PubMed PMID: 23015034; PubMed Central PMCID: PMCPMC3458229.

[23] Putto A. Febrile exudative tonsillitis: viral or streptococcal? Pediatrics 1987;80 (1):6–12. Epub 1987/07/01. PubMed PMID: 3601520.

[24] McIsaac WJ, Kellner JD, Aufricht P, Vanjaka A, Low DE. Empirical validation of guidelines for the management of pharyngitis in children and adults. JAMA 2004;291(13):1587–1595. Epub 2004/04/08. https://doi.org/10.1001/jama.291.13.1587. PubMed PMID: 15069046.

[25] Florin TA, Ambroggio L, Brokamp C, Rattan MS, Crotty EJ, Kachelmeyer A, et al. Reliability of examination findings in suspected community-acquired pneumonia. Pediatrics 2017;140(3):. https://doi.org/10.1542/peds.2017-0310%Je20170310.

[26] Gordon JK, Girish G, Berrocal VJ, Zhang M, Hatzis C, Assassi S, et al. Reliability and validity of the tender and swollen joint counts and the modified rodnan skin score in early diffuse cutaneous systemic sclerosis: analysis from the prospective registry of early systemic sclerosis cohort. J Rheumatol 2017. https://doi.org/10.3899/jrheum.160654%J. jrheum.160654.

[27] Rawal S, Hoffman HJ, Honda M, Huedo-Medina TB, Duffy VBJCP. The taste and smell protocol in the 2011–2014 US National Health and Nutrition Examination Survey (NHANES): test–retest reliability and validity testing. Chemosens Percept 2015;8(2):138–48. https://doi.org/10.1007/s12078-015-9194-7.

[28] Lee JY, Chang JS, Koo KC, Lee SW, Choi YD, Cho KS. Hematuria grading scale: a new tool for gross hematuria. Urology 2013;82(2):284–9. https://doi.org/10.1016/j.urology.2013.04.048.