Published in final edited form as:

Pediatrics. 2010 November ; 126(5): 925–930. doi:10.1542/peds.2010-1039.

Reliability of Clinical Examinations for Pediatric Skin and Soft-Tissue Infections

Jennifer R. Marin, MD, MSc^{a,b}, Warren Bilker, PhD^d, Ebbing Lautenbach, MD, MPH, MSCE^{c,d}, and Elizabeth R. Alpern, MD, MSCE^{a,b,d}

^aDivision of Emergency Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania

^bDepartment of Pediatrics, Department of Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania

^cDivision of Infectious Diseases, Department of Medicine, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

^dCenter for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania

Abstract

OBJECTIVE—To determine the interrater reliability of clinical examination by pediatric emergency medicine physicians for the diagnosis of skin and soft-tissue infections (SSTIs).

METHODS—A cross-sectional study of patients presenting to a pediatric emergency department with SSTIs was performed. Each lesion was examined by a treating physician and a study physician (from a pool of 62 physicians) at the bedside during the emergency department visit. The primary outcome was reliability, as measured with the weighted κ statistic, for determining whether the lesion was an abscess and whether the lesion required a drainage procedure.

RESULTS—A total of 371 lesions were analyzed for interrater reliability. The weighted κ value for diagnosis of the lesion as an abscess was 0.39 (95% confidence interval: 0.32–0.47), and that for assessment of the need for drainage was 0.43 (95% confidence interval: 0.36–0.51). Agreement was statistically more likely for lesions in children ≥ 4 years of age but was not more likely for lesions in nonblack patients, lesions in patients with a history of or exposure to a close contact with a SSTI, or lesions examined by 2 experienced pediatric emergency medicine physicians.

CONCLUSIONS—Among the 62 participating physicians at our site, the reliability of the clinical examination was poor. This may indicate that improved education and/or more-objective means for diagnosing these infections in the acute care setting are warranted. Additional studies are needed to determine whether these results are generalizable to other settings.

Keywords

skin and soft-tissue infections; reliability

Copyright © 2010 by the American Academy of Pediatrics

Address correspondence to Jennifer R. Marin, MD, MSc, Children's Hospital of Pittsburgh, Division of Emergency Medicine, 4401 Penn Ave, AOB, Suite 2400, Pittsburgh, PA 15224. marinj@email.chop.edu.

Dr Marin's current affiliations are Division of Emergency Medicine, Children's Hospital of Pittsburgh, Pittsburgh, PA; and Department of Pediatrics, University of Pittsburgh, School of Medicine, Pittsburgh, PA.

FINANCIAL DISCLOSURE: Dr Marin has received research support in the form of loaned ultrasound equipment from SonoSite. Dr Lautenbach has received research funding from Merck, AstraZeneca, Cubist, and Ortho-McNeil.

In recent years, there have been dramatic increases in the numbers of patients with skin and soft-tissue infections (SSTIs) presenting to emergency departments (EDs). In 2005, 3.4 million ED visits were attributable to SSTIs, compared with 1.2 million visits in 1993.¹ Currently, most physicians rely on clinical examination results in diagnosing these infections, particularly in distinguishing cellulitis from abscesses. This distinction is important, because treatment of most abscesses includes a drainage procedure,^{2,3} whereas cellulitis is treated with systemic antibiotic therapy alone.⁴ Because of the overlap in clinical symptoms between these 2 entities, however, accurate discrimination between abscesses and cellulitis is difficult.⁵ Errors in diagnosis can lead to several types of adverse consequences. Misdiagnosis of cellulitis as an abscess can result in an unnecessary drainage procedure, with subsequent risks, costs, and trauma. Conversely, failure to differentiate an abscess from isolated cellulitis can result in disease progression and the extra costs and time associated with a return visit for definitive care, with incision and drainage.

Given the differing treatment options for these 2 clinical entities and the risks of misdiagnosis, it is important that physicians be able to diagnose these infections reliably. Because clinicians primarily use clinical examination results to make these diagnoses, rigorous evaluation of the performance of examinations is necessary.⁶ One aspect of this evaluation is the assessment of reliability, which measures the extent to which repeated measurements or evaluations yield the same results.⁷ Although no studies have examined the reliability of clinical examinations for SSTIs specifically, previous studies showed that physician interrater reliability of clinical examinations for pharyngitis, abdominal pain, and ankle injuries and bimanual examinations is no more than moderate.^{8–13} The objective of this study was to determine the interexaminer reliability of clinical examinations by pediatric emergency medicine (PEM) physicians for the diagnosis of SSTIs. Specifically, we sought to determine agreement in clinical opinions regarding the presence of an abscess, as well as whether the physician thought that the lesion required a drainage procedure. Secondary objectives included measurements of interrater agreement within specific subgroups, including older patients, patients with lighter skin (nonblack), and patients with a history of or exposure to a close contact with a history of a SSTI, and between physician pairs with greater experience in PEM.

METHODS

Study Design and Setting

We performed a cross-sectional study of a convenience sample of patients who presented to a pediatric ED with a SSTI during an 18-month period (June 1, 2008, through November 30, 2009). Each patient was examined by 2 independent physicians, as part of a larger cohort study evaluating diagnostic options for SSTIs. This study was performed in the ED of an urban, tertiary care, pediatric hospital with an annual census of ~90 000 patients.

Patient Selection

Patients were eligible for enrollment if they were between 2 months and 19 years of age and the treating physician considered the diagnosis of an isolated SSTI requiring treatment with systemic antibiotic therapy (ie, folliculitis was excluded). Of note, at our institution, it is standard practice currently to treat SSTIs with systemic antibiotic therapy regardless of whether a drainage procedure is performed. If >1 lesion was present, then we enrolled up to 3 lesions per patient. If an individual patient had >3 lesions, then determination of which 3 to enroll was at the discretion of the treating physician. We excluded patients without an English-speaking parent or guardian available to provide consent, patients who had been enrolled in the study previously, patients who underwent imaging (such as ultrasonography or computed tomography) before arrival at the ED, and immunocompromised patients. We

did not include lesions involving the face, genital region, or perirectal region, surgical wound infections, or felons, because we determined a priori that these lesions likely would be managed by a sub-specialty consultant rather than the emergency physician. In addition, we excluded lymphadenitis, paronychia, and infections surrounding indwelling catheters, because these likely would represent different infectious processes, compared with simple SSTIs. Trained research coordinators screened the electronic tracking board in the ED for potentially eligible patients 17 hours per day (between 7 AM and midnight), 7 days per week. Parents/guardians and patients provided written consent and assent, respectively. The study was approved by the hospital's institutional review board.

Physician Raters

Physician examiners consisted of 62 PEM attending physicians, PEM fellows, and urgent care attending physicians (board-certified pediatricians without formal PEM training). We documented physicians' levels of experience in PEM on the basis of the date of PEM board certification, which was obtained from the American Board of Pediatrics. Physicians provided verbal consent for documentation of experience levels for use in comparisons of interrater reliability.

Measurements

We obtained information regarding historical and clinical features of the patients and lesions in question with a standardized physician questionnaire. Each patient was examined by 2 independent physicians linked to each other's assessment, that is, the treating physician and a study physician. After the history-taking and physical examination, the physicians completed the questionnaire, documenting their overall clinical impressions, including the presence or absence of an abscess and the need for a drainage procedure, defined as manual manipulation (squeezing), needle aspiration, or incision and drainage. In addition, we asked the treating physicians whether they had information regarding imaging or drainage before they completed the questionnaire, because such information might bias their impressions.

The primary outcome measures were interrater agreements between the 2 physicians in their opinions regarding whether the lesion was an abscess and whether drainage was needed. Opinions regarding diagnosis of an abscess were measured as definitely no abscess, probably no abscess, uncertain, probably abscess, or definitely abscess. Opinions regarding the need for a drainage procedure were measured as no drainage needed, drainage needed, or uncertain.

Covariates

To determine the effects of patient and clinician characteristics on agreement among practitioners, we performed analyses with stratification according to patient age (<4 years versus ≥ 4 years), race (black versus nonblack), history of or exposure to a close contact with a history of a SSTI within the past year, and physician training/experience (≥ 3 years since board certification in PEM versus <3 years since board certification in PEM or non-PEM pediatrician). These variables and categories were determined a priori and on the basis of previous similar analyses, when possible.⁷ We assessed the associations between these covariates and agreement by using χ^2 tests.

Statistical Analyses

For interrater reliability, we calculated the rate of agreement and Cohen's weighted κ statistic,¹⁴ a measure of the extent to which agreement is greater than expected on the basis of chance alone. The κ statistic is affected by the diagnostic base rates of each rater and therefore accounts for the proportion of cases to which each rater gives a negative or

positive diagnosis.¹⁵ The weighted κ uses linear weights to account for the equal-appearing intervals that separate rank-ordered categories.¹⁵ Weighting allows for partial credit for discordant observations when the categories assigned by the different observers are closer together, with more weight being given to rater discrepancies that are closer in rank order than to those that are farther away. Table 1 presents the exact weights determined a priori and used in the analysis. We calculated 95% confidence intervals (CIs) for the weighted κ values by using the bootstrap method,¹⁶ with 1000 replications. We interpreted κ results as indicating poor agreement ($\kappa < 0.00$), slight agreement ($\kappa = 0.00$ – 0.20), fair agreement ($\kappa = 0.21$ – 0.40), moderate agreement ($\kappa = 0.41$ – 0.60), substantial agreement ($\kappa = 0.61$ – 0.80), or nearly perfect agreement ($\kappa = 0.81$ – 1.00).¹⁷ Patients with missing data were excluded from the κ analysis for that variable or, if the physician opinion regarding drainage was missing, from the overall κ analysis.

In the primary analysis, we included multiple lesions for individual patients. To evaluate the impact of this decision on interrater reliability, we repeated the analysis by using 1 randomly selected lesion per patient. We assessed the potential for selection bias by comparing the demographic features of eligible patients who were missed with those of patients who were enrolled.

A sample size of 335 lesions was needed to establish a 95% CI for the κ statistic of ± 0.1 , given a hypothesized κ of 0.4 (the upper limit of the range considered fair agreement¹⁷) and a suspected proportion of lesions rated as requiring drainage for each of the raters of 0.6. We analyzed the data by using Stata 10.0 (Stata Corp, College Station, TX).

RESULTS

A total of 394 lesions in 349 patients were evaluated (Fig 1). Table 2 lists the patient and lesion characteristics for the enrolled population. For 23 lesions, there was not appropriate documentation from one or both raters regarding their diagnoses. These lesions were not substantively different in terms of patient demographic features. A total of 371 lesions were analyzed for inter-rater reliability. Patient characteristics (age, gender, and race/ethnicity) of the patients who were eligible for enrollment but were missed (presented to the ED outside times of screening and enrollment) were similar to those of the patients who were enrolled.

Table 3 shows the 2 measurements of physician agreement. For diagnosis of a lesion as an abscess, the weighted κ statistic was 0.39. For determination of whether a lesion required a drainage procedure, the κ statistic was 0.43.

Physicians were more likely to agree on whether a lesion required a drainage procedure for patients ≥ 4 years of age, compared with patients < 4 years of age (relative risk [RR]: 1.24 [95% CI: 1.04–1.48]). However, agreement was not more likely for lesions among nonblack patients (RR: 1.02 [95% CI: 0.86–1.21]), for lesions among patients with a history of or exposure to a close contact with a SSTI (RR: 0.95 [95% CI: 0.80–1.14]), or when 2 PEM physicians with ≥ 3 years of experience examined the lesion (RR: 0.98 [95% CI: 0.84–1.15]).

We performed a second analysis in which we allowed for only 1 randomly selected lesion per patient; 330 lesions were evaluated. For diagnosis of a lesion as an abscess, the weighted κ statistic was 0.40 (95% CI: 0.33–0.48). For determination of whether a lesion required a drainage procedure, the κ statistic was 0.46 (95% CI: 0.37–0.54). When we analyzed agreement according to each of the 4 aforementioned covariates, the results were similar to those of the primary analysis, with agreement varying only according to patient age (RR: 1.24 [95% CI: 1.04–1.49]).

For 6 lesions, treating physicians indicated that they had information on procedures or drainage before documentation of their diagnostic and management impressions on the questionnaire. When we analyzed the data by excluding those 6 lesions, the results were unchanged.

DISCUSSION

Our results demonstrate that clinical examinations by PEM physicians for the diagnosis and management of SSTIs are unreliable. Similar results were demonstrated in other studies of bedside clinical examinations. In a study evaluating adult patients with sore throats, physicians demonstrated slight to moderate agreement in physical examination findings.⁸ Only fair agreement was demonstrated when 2 attending physicians examined adult patients for conjunctival pallor.¹⁸ For the evaluation by emergency physicians of adult patients with ankle injuries, the reliability of several components of the physical examination that were incorporated in the decision to obtain radiographs was poor to fair.¹³ A study of physical examinations for patients with suspected appendicitis between senior surgical residents and PEM physicians showed slight to moderate agreement, depending on the physical examination component.¹¹ Finally, in a study of children with abdominal pain, the reliability between PEM attending physicians and surgical residents was poor to moderate.⁹

When we stratified our results, we found little to no improvement in agreement among our selected covariates. We theorized that the clinical evaluation of younger children might be more variable and, although we did find that physicians were statistically more likely to agree regarding older children, the clinical significance of this finding is questionable and should be pursued further. The level of training has been shown to be a factor in reliability and correlates with how much specific attention raters pay to relevant cues and how much interest they actually have in the activity being assessed.¹⁵ However, we did not demonstrate that 2 physicians who were more experienced, according to our definition, were more reliable than pairs who were not more experienced or were of different experience levels. It is possible that greater experience does not lead to more-consistent examination results in the case of SSTIs, which suggests that the examination of these lesions is difficult for all practitioners and they may require a more-objective means of assessment.

Possible explanations for our findings of a lack of reliability include a lack of consistent clinical criteria and indications for diagnosing and treating these lesions and potential subjectivity in interpreting clinical examination results. Therefore, more-standardized, more-objective methods of diagnosis should be investigated as a means to improve reliability, such as focused education and teaching regarding examination of SSTIs and evaluation of bedside imaging studies, such as clinician-performed ultrasonography for this indication.

There are several limitations to this study. Although the 2 examiners were independent and blinded to each others' opinions, we cannot exclude the possibility that the study physician obtained clues to the treating physician's opinion, such as topical anesthesia applied to the area, a nurse preparing for intravenous sedation, or parents discussing the treatment plan. However, any such clues would serve to increase agreement; therefore, our findings would represent an overestimation of the true agreement. We also cannot exclude the possibility that the second physician performed a less-careful history and physical examination than the treating physician, because he or she might have been less concerned than the treating physician with making an accurate diagnosis. This would lead to an underestimation of the true κ , although we expect the impact to be minimal. We considered standardizing the history and physical examination, to overcome this limitation, but we wanted our study to reflect true practice conditions. In addition, we considered multiple lesions on a single patient to be independent, which might inflate the κ statistic if physicians were more willing

to assign a similar diagnosis and plan to a second or third lesion. However, our results were the same when we selected randomly 1 lesion per patient. When we stratified our analysis according to physician experience, we determined a priori that ≥ 3 years of practice after fellowship training would define experience. It is possible that this cutoff point does not represent enough experience that there would be an effect on agreement. Also, non-PEM physicians were included in the group of physicians with < 3 years since PEM training. Some of those physicians might have had more experience, in terms of number of years of practice. However, we think that fellowship training represents a different level of training and experience, compared with the training and experience of those without sub-specialty training, regardless of years of practice. If some of these physicians were misclassified, it is not clear how the misclassification would have affected our results. Because this was a single-center study, it is possible that our results cannot be generalized to other practice settings; however, the large number of patients and physicians who participated in the study, each with individual practice patterns, might serve to mitigate this limitation. Moreover, because we enrolled a convenience sample of patients during times when a research associate and a study physician were available, there might be a selection bias regarding the patients who were enrolled in the study. However, when we evaluated this by comparing demographic characteristics of missed and enrolled patients, the 2 populations were similar. In addition, it is unlikely that patients who were missed would represent a different population with a different disease process. Finally, in the case of some labile or dynamic disease processes, such as abdominal pain, it is possible that patients' examination results change between examiners; therefore, the κ statistic may be underestimated. SSTIs typically are stable within the window of our examinations, however, and our estimate of reliability should not have been affected.

CONCLUSIONS

Overall, at our institution, reliability among PEM physicians for diagnosis of a SSTI as an abscess was poor. Because there was such variability in the diagnosis of this common disease process, there should be further investigation into the reasons for this lack of reliability, such as the need for improved education and more-objective means for diagnosing these infections. Studies involving other institutions and other settings should be performed to evaluate whether these findings are generalizable.

Acknowledgments

This study was supported by a grant from the Nicholas Crognale Chair for Emergency Medicine, Children's Hospital of Philadelphia (Philadelphia, PA). We thank Jeremy Kahn, MD, MSc, for critical review of this manuscript.

ABBREVIATIONS

SSTI	skin and soft-tissue infection
ED	emergency department
PEM	pediatric emergency medicine
RR	relative risk
CI	confidence interval

References

1. Pallin DJ, Egan DJ, Pelletier AJ, Espinola JA, Hooper DC, Camargo CA Jr. Increased US emergency department visits for skin and soft tissue infections, and changes in antibiotic choices,

- during the emergence of community-associated methicillin-resistant *Staphylococcus aureus*. *Ann Emerg Med*. 2008; 51(3):291–298. [PubMed: 18222564]
2. Daum RS. Skin and soft-tissue infections caused by methicillin-resistant *Staphylococcus aureus*. *N Engl J Med*. 2007; 357(4):380–390. [PubMed: 17652653]
 3. Fitch MT, Manthey DE, McGinnis HD, Nicks BA, Pariyadath M. Videos in clinical medicine: abscess incision and drainage. *N Engl J Med*. 2007; 357(19):e20. [PubMed: 17989377]
 4. Swartz MN. Cellulitis. *N Engl J Med*. 2004; 350(9):904–912. [PubMed: 14985488]
 5. Butler, K. Manifestations of abscess formation. In: Roberts, JR., editor. *Clinical Procedures in Emergency Medicine*. 4. Philadelphia, PA: Saunders; 2004. p. 717-726.
 6. Fitzgerald FT. Physical diagnosis versus modern technology: a review. *West J Med*. 1990; 152(4):377–382. [PubMed: 2190412]
 7. Stevens MW, Gorelick MH, Schultz T. Interrater agreement in the clinical evaluation of acute pediatric asthma. *J Asthma*. 2003; 40(3):311–315. [PubMed: 12807175]
 8. Schwartz K, Monsur J, Northrup J, West P, Neale AV. Pharyngitis clinical prediction rules: effect of interobserver agreement: a MetroNet study. *J Clin Epidemiol*. 2004; 57(2):142–146. [PubMed: 15125623]
 9. 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. [PubMed: 15809393]
 10. Pines J, Uscher Pines L, Hall A, Hunter J, Srinivasan R, Ghaemmaghami C. The interrater variation of ED abdominal examination findings in patients with acute abdominal pain. *Am J Emerg Med*. 2005; 23(4):483–487. [PubMed: 16032616]
 11. 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. [PubMed: 18275440]
 12. Close RJ, Sachs CJ, Dyne PL. Reliability of bimanual pelvic examinations performed in emergency departments. *West J Med*. 2001; 175(4):240–244. [PubMed: 11577050]
 13. Stiell IG, McKnight RD, Greenberg GH, Nair RC, McDowell I, Wallace GJ. Interobserver agreement in the examination of acute ankle injury patients. *Am J Emerg Med*. 1992; 10(1):14–17. [PubMed: 1736906]
 14. Cohen J. Weighted κ : nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968; 70(4):213–220. [PubMed: 19673146]
 15. Cicchetti D, Bronen R, Spencer S, et al. Rating scales, scales of measurement, issues of reliability: resolving some critical issues for clinicians and researchers. *J Nerv Ment Dis*. 2006; 194(8):557–564. [PubMed: 16909062]
 16. Stata Corp. *Stata Reference Manual Release 7*. Vol. 1. College Station, TX: Stata Corp; 2001. bstrap: bootstrap sampling and estimation; p. 164-174.
 17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33(1):159–174. [PubMed: 843571]
 18. Wallace DE, McGreal GT, O'Toole G, et al. The influence of experience and specialisation on the reliability of a common clinical sign. *Ann R Coll Surg Engl*. 2000; 82(5):336–338. [PubMed: 11041034]

WHAT'S KNOWN ON THIS SUBJECT

Studies have demonstrated fair to moderate reliability of clinical examinations for several diagnoses. Skin and soft-tissue infections represent a common reason for acute care visits, and physicians rely on clinical acumen to determine the need for drainage of these lesions.

WHAT THIS STUDY ADDS

As a poorly reliable method for determining management of skin and soft-tissue infections, clinical examinations should not be the sole means through which physicians evaluate these lesions, and more-objective methods should be investigated.

FIGURE 1.
Patient enrollment diagram. HSV indicates herpes simplex virus.

TABLE 1

Weights Assigned to Each Rating Combination for Abscess and Drainage Opinions

Rater 1	Rater 2	κ Weight (0–1.0)^a
Abscess		
Probably abscess	Definitely abscess	0.8
Probably no abscess	Definitely no abscess	0.8
Probably abscess	Uncertain	0.5
Probably no abscess	Uncertain	0.5
Definitely abscess	Uncertain	0.25
Definitely no abscess	Uncertain	0.25
Definitely no abscess	Probably abscess	0
Probably no abscess	Probably abscess	0
Definitely no abscess	Definitely abscess	0
Probably no abscess	Definitely abscess	0
Drainage		
No drainage needed	Uncertain	0.5
Drainage needed	Uncertain	0.5
Drainage needed	No drainage needed	0

^aFull weight of 1.0 was given to ratings that were in complete agreement.

TABLE 2**Patient and Lesion Characteristics**

Patients (<i>N</i> = 349)	
Age	
Mean $\pm$ SD, y	7.7 $\pm$ 6
Minimum, mo	2
Maximum, y	19
Ethnicity, <i>n</i> (%)	
Black	225 (64)
White	92 (26)
Other	32 (9)
Female, <i>n</i> (%)	197 (56)
History of/exposure to contact with SSTI, <i>n</i> (%)	157 (45)
Lesion location (<i>N</i> = 394), <i>n</i> (%)	
Leg	119 (30)
Buttock	98 (25)
Abdomen	27 (7)
Arm	25 (6)

TABLE 3

Measurements of Clinician Agreement

Assessment Variable	N	Agreement, %	κ (95% CI)
Diagnosis of abscess	371	70.4	0.39 (0.32–0.47)
Need for drainage	371	61.7	0.43 (0.36–0.51)