Reliability of Examination Findings in Suspected Community-Acquired Pneumonia

Todd A. Florin, MD, MSCE,[a,b] Lilliam Ambroggio, PhD, MPH,[b,c,d] Cole Brokamp, PhD,[c] Mantosh S. Rattan, MD,[b,e] Eric J. Crotty, MD,[b,e] Andrea Kachelmeyer, BS, CCRP,[a] Richard M. Ruddy, MD,[a,b] Samir S. Shah, MD, MSCE[a,b,d,f]

abstract

BACKGROUND: The authors of national guidelines emphasize the use of history and examination findings to diagnose community-acquired pneumonia (CAP) in outpatient children. Little is known about the interrater reliability of the physical examination in children with suspected CAP.

METHODS: This was a prospective cohort study of children with suspected CAP presenting to a pediatric emergency department from July 2013 to May 2016. Children aged 3 months to 18 years with lower respiratory signs or symptoms who received a chest radiograph were included. We excluded children hospitalized ≤14 days before the study visit and those with a chronic medical condition or aspiration. Two clinicians performed independent examinations and completed identical forms reporting examination findings. Interrater reliability for each finding was reported by using Fleiss' kappa (κ) for categorical variables and intraclass correlation coefficient (ICC) for continuous variables.

RESULTS: No examination finding had substantial agreement (κ/ICC > 0.8). Two findings (retractions, wheezing) had moderate to substantial agreement (κ/ICC = 0.6–0.8). Nine findings (abdominal pain, pleuritic pain, nasal flaring, skin color, overall impression, cool extremities, tachypnea, respiratory rate, and crackles/rales) had fair to moderate agreement (κ/ICC = 0.4–0.6). Eight findings (capillary refill time, cough, rhonchi, head bobbing, behavior, grunting, general appearance, and decreased breath sounds) had poor to fair reliability (κ/ICC = 0–0.4). Only 3 examination findings had acceptable agreement, with the lower 95% confidence limit >0.4: wheezing, retractions, and respiratory rate.

CONCLUSIONS: In this study, we found fair to moderate reliability of many findings used to diagnose CAP. Only 3 findings had acceptable levels of reliability. These findings must be considered in the clinical management and research of pediatric CAP.

Divisions of [a]Emergency Medicine, [c]Biostatistics and Epidemiology, [d]Hospital Medicine, and [f]Infectious Diseases, and [e]Department of Radiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; and [b]Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, Ohio

Dr Florin conceptualized and designed the study, interpreted the data, and drafted the initial manuscript; Drs Ambroggio, Ruddy, and Shah played substantial roles in study design and data interpretation and critically reviewed the manuscript; Dr Brokamp conducted the initial data analyses, interpreted the data, and critically reviewed the manuscript; Drs Rattan and Crotty interpreted all chest radiographs in this study (data acquisition) and critically reviewed the manuscript; Ms Kachelmeyer coordinated and supervised data collection and critically reviewed the manuscript; and all authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work. Dr Florin and Dr Brokamp had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

WHAT'S KNOWN ON THIS SUBJECT: The authors of national guidelines emphasize the use of history and examination findings to diagnose community-acquired pneumonia (CAP) in outpatient children. Little is known about the interrater reliability of the physical examination in children with suspected CAP.

WHAT THIS STUDY ADDS: In this study, we found fair to moderate reliability of many findings used to diagnose CAP. Only 3 of the 19 findings examined had acceptable reliability. The fair reliability of the examination must be considered in the clinical management and research of pediatric CAP.

To cite: Florin TA, Ambroggio L, Brokamp C, et al. Reliability of Examination Findings in Suspected Community-Acquired Pneumonia. *Pediatrics*. 2017;140(3):e20170310

Children with a respiratory illness commonly require emergency department (ED) evaluation. Distinguishing community-acquired pneumonia (CAP) from other causes of respiratory illness can be challenging. Consequently, there is substantial variation in ED evaluation of children with CAP because clinicians attempt to distinguish upper from lower respiratory tract infection (LRTI) and viral from bacterial causes.[1,2]

The Pediatric Infectious Diseases Society and Infectious Diseases Society of America published consensus guidelines for CAP management in children to minimize unnecessary variation and encourage judicious use of testing and antibiotics.[3] Citing high-quality evidence, the authors of the guidelines provide a strong recommendation that routine chest radiographs (CXRs) are not necessary to confirm suspected CAP in patients well enough to be cared for at home after office or ED evaluation, stressing that the clinician should diagnose CAP by using historical and examination findings.

Given the importance of the physical examination in the diagnosis of children with suspected CAP, it is critical that examination findings are reliable. Interrater reliability (IRR), the ability to obtain similar results by different examiners under similar conditions, is a key component of any examination measure. If a measure has poor IRR, it cannot be used to make an accurate diagnosis that is generalizable across providers and practice settings. Among adults with CAP, the IRR of examination findings was highly variable, with most demonstrating poor to fair IRR.[4] Studies of children with cough, asthma, and bronchiolitis have revealed mixed results, ranging from poor to good reliability of examination findings.[5–8] Given the variability in these respiratory processes and the paucity of evidence in pediatric CAP, our objective was to evaluate the IRR of examination findings specifically for children with suspected CAP. We hypothesized that examination measures frequently used to diagnose CAP in children would demonstrate only fair to moderate IRR.

METHODS

Study Design

This study is part of an ongoing prospective cohort study of children with suspected CAP called Catalyzing Ambulatory Research in Pneumonia Etiology and Diagnostic Innovations in Emergency Medicine (CARPE DIEM). Patients who presented to the ED at Cincinnati Children's Hospital Medical Center (CCHMC) and were enrolled in CARPE DIEM from July 2013 to May 2016 were eligible for inclusion in this analysis. The study was approved by the CCHMC Institutional Review Board. Informed consent was obtained from all legal guardians of patients and assent from children ≥11 years of age.

Study Population

Children 3 months to 18 years of age with signs and symptoms of LRTI who received a CXR for suspicion of CAP were enrolled. Signs and symptoms of LRTI were defined as having one or more of the following: new or different cough, new or different sputum production, chest pain, dyspnea and/or shortness of breath, documented tachypnea, or abnormal findings consistent with LRTI on physical examination (eg, rales/crackles, wheezing).[9] There are no uniform standards for use of CXRs with suspected pneumonia at our institution; in 2016, 83% of children with a diagnosis of pneumonia received a CXR in our ED. We excluded children hospitalized ≤14 days before the index ED visit to exclude possible hospital-acquired pneumonia. Children with immunocompromising or chronic medical conditions known to predispose to severe or recurrent pneumonia (eg, immunodeficiency, chronic corticosteroid use, cystic fibrosis, chronic lung disease, malignancy, sickle cell disease, congenital heart disease, tracheostomy-dependent patients, and neuromuscular disorders affecting the lungs) were not included, nor were children with a history of aspiration or aspiration pneumonia. These criteria were developed to include otherwise healthy children with suspected CAP. Patients who were enrolled within 30 days before the study ED visit were excluded to ensure a distinct episode of infection during the study visit.

Study Protocol

The current study was composed of a convenience sample of patients from the CARPE DIEM cohort for whom 2 clinicians were able to perform independent physical examinations. The patient's primary treating clinician (attending physician, pediatric emergency medicine [PEM] fellow, or nurse practitioner), the first assessor, completed a standardized physical examination on a case report form. A second assessor completed an examination on the same child and an identical form independent of the first assessor. The second assessor was identified on the basis of available clinicians in the ED at the time of enrollment; it could have been a fellow working with the first assessor or another clinician working elsewhere in the ED. It was required that the examination was not discussed between the 2 assessors before completion of the case report forms. Ideally, both examinations would be performed before knowledge of the CXR results; however, this was challenging for practical reasons of clinical flow in a busy ED. In our study, 71% of first assessors and 63% of second assessors knew radiograph

results before the examination. The research coordinators encouraged completion of the 2 examinations within 20 minutes of each other. In addition, no respiratory treatments, including nebulized therapies or nasal suctioning, should have occurred between the 2 examinations to mitigate the effect of these treatments on examination findings.

Measurements

The case report forms recorded specific examination findings relevant to a child presenting with suspected CAP. These findings were selected on the basis of an extensive literature review and expert opinions of faculty physicians in emergency medicine, hospital medicine, and infectious diseases.[3,10–13] Examination findings included general appearance (ie, how the patient appeared to the clinician on first observation), behavior, perfusion (skin color, cool extremities, capillary refill, and peripheral pulse quality), respiratory signs (cough, nasal flaring, retractions, grunting, crackles/rales, rhonchi, wheezing, head bobbing, decreased breath sounds, pleuritic chest pain, abdominal pain, tachypnea, and respiratory rate), and overall clinical impression (ie, the clinician's final impression after taking all history and examination factors into account). Respiratory rate was counted by the individual clinicians at the time of the examination. To capture real-world interpretation of the physical examination, we did not provide detailed definitions of findings. For certain findings, such as pleuritic chest pain, assessment is challenging because of developmental immaturity. We included options such as "too young to assess," "unable to assess," or "unknown" for these findings.

To define radiographic pneumonia, CXRs were independently reviewed by 2 radiologists (M.S.R. and E.J.C.) blinded to all clinical information and outcomes. A patient was considered to have radiographic pneumonia when both radiologists agreed that radiograph findings favored pneumonia (as opposed to atelectasis or normal findings). In cases of disagreement, the attending radiologist's impression on the clinical report at the time of the study visit acted as a tiebreaker.

Data Analysis

IRR for categorical findings was assessed by using Fleiss' kappa (κ), and its 95% confidence interval (CI) was estimated by using 1000 bootstrap replicates.[14–16] κ accounts for variability across many raters because many different physicians performed physical examinations as part of CARPE DIEM. κ is reported on a 0 to 1 scale, with 0 indicating poor agreement and 1 indicating near-perfect agreement (poor to slight agreement: κ = 0.00–0.20; fair agreement: κ = 0.21–0.40; moderate agreement: κ = 0.41–0.60; substantial agreement: κ = 0.61–0.80; and near-perfect agreement: κ = 0.81–1.00).[17] On the basis of previous literature, we defined acceptable agreement as at least moderate agreement, with the lower 95% confidence limit of $\kappa \geq 0.4$.[17–21] IRR for continuous findings was assessed by using the intraclass correlation coefficient (ICC). We repeated the above analyses for 2 subgroups: (1) those with and without radiographic pneumonia and (2) those younger than and older than 5 years of age. We anticipated that some examination findings would rarely be abnormal (eg, peripheral pulses) and thus have high measures of raw agreement while not necessarily achieving at least moderate agreement with the lower 95% confidence limit of κ. With our sample size calculations, we determined that to achieve a κ of 0.7 with a lower limit of the 95% CI of 0.4 and conservatively assuming the prevalence of a finding is 5%, 122 paired examinations were required. Analyses were performed by using R (v3.3).[22]

RESULTS

Paired assessments were performed on 128 children. Assessments were completed within 20 minutes of each other in 96.5% of children; 94.7% had no breathing treatments between evaluations. The study cohort for IRR was similar to the overall CARPE DIEM cohort in age, sex, medical history, presence of radiographic pneumonia, and primary billing diagnoses (Supplemental Table 3). For the first assessment, 2 assessments (2%) were performed by nurse practitioners, 46 (36%) by PEM fellow physicians, 40 (31%) by attending physicians with <5 years of practice after completion of training, and 40 (31%) by attending physicians with ≥5 years of practice after completion of training. For the second assessment, 3 (2%) were performed by nurse practitioners, 58 (45%) by PEM fellow physicians, 21 (17%) by attending physicians with <5 years of practice after completion of training, and 46 (36%) by physicians with ≥5 years of practice after completion of training.

The prevalence of findings as determined by the clinician caring for the patient (ie, first assessor) is reported in Table 1. Raw overall agreement for physical examination findings ranged from 52% (behavior) to 96% (cool extremities and peripheral pulses) (Fig 1). No examination finding had substantial to near-perfect agreement. Two findings (retractions and wheezing) had moderate to substantial agreement (κ = 0.6–0.8). Eight findings (abdominal pain, pleuritic pain, nasal flaring, skin color, overall impression, cool extremities, tachypnea, and crackles/rales) had moderate agreement (κ = 0.4–0.6). The ICC for respiratory rate was 0.58 (95% CI 0.41 to 0.72), also indicating

TABLE 1 Prevalence of Examination Findings by Primary Clinician

Finding	Presence of Finding by Primary Assessor, *n* (%)
General examination	
General appearance	
Well	35 (27)
Mildly ill or distressed	40 (31)
Moderately ill or distressed	49 (38)
Severely ill or distressed	4 (3)
Behavior	
Playing and appropriate	33 (26)
Quiet but appropriate	57 (45)
Sleeping but easily arousable	11 (9)
Fussy but consolable	20 (16)
Irritable	6 (5)
Lethargic, confused, or reduced response to pain	1 (1)
Respiratory examination	
Crackles	49 (38)
Decreased breath sounds	55 (43)
Wheezing	41 (32)
Retractions	73 (57)
Rhonchi	43 (34)
Tachypneic	89 (70)
If tachypneic, respiratory rate, mean breaths per minute (SD)	46 (12)
Nasal flaring	30 (23)
Pleuritic chest pain	14 (27)
Grunting	14 (11)
Head bobbing	9 (7)
Observed cough	68 (53)
Cardiovascular and perfusion examination	
Cool extremities	4 (3)
Skin color	
Pink and/or normal	109 (85)
Pale or dusky	18 (14)
Mottled	1 (1)
Capillary refill time (s)	
1–2	109 (85)
3	19 (15)
4	0 (0)
≥5	0 (0)
Peripheral pulses	
Normal	124 (98)
Bounding	3 (2)
Abdominal examination	
Abdominal pain or tenderness	14 (11)
Overall impression	
Mild	55 (43)
Moderate	63 (50)
Severe	9 (7)

moderate agreement. Eight findings (capillary refill time, cough, rhonchi, head bobbing, behavior, grunting, general appearance, and decreased breath sounds) had poor to fair reliability (κ or ICC = 0–0.4). Only 3 examination findings had acceptable agreement: wheezing, retractions, and respiratory rate (Fig 1). One finding, peripheral pulses, was rated as normal in 97% (*n* = 124) of patients; therefore, we only report the raw agreement of 96%.

If auscultatory findings were present, examiners were asked whether findings were diffuse or focal. Analyses of focal findings were limited by sample size. For wheezing, there were no focal findings. For rhonchi, there was 1 focal finding. For crackles (*n* = 28), the raw agreement on diffuse versus focal was 86% with a κ of 0.52 (95% CI, −0.06 to 0.9). For decreased breath sounds (*n* = 34), raw agreement was 71% with a κ of 0.33 (95% CI, −0.02 to 0.67). Once it was agreed that focal findings were present, the reliability of location was moderate to substantial. For crackles (*n* = 21), left-sided findings had a raw agreement of 86% with a κ of 0.69 (95% CI, 0.3 to 1.0), and right-sided findings (*n* = 21) had a raw agreement of 90% with a κ of 0.81 (95% CI, 0.53 to 1.0). For decreased breath sounds (*n* = 18), left-sided findings had a raw agreement of 89% with a κ of 0.77 (95% CI, 0.4 to 1.0), and right-sided findings had a raw agreement of 78% with a κ of 0.54 (95% CI, 0.11 to 0.89).

Table 2 illustrates the prevalence of examination findings by the primary treating clinician (ie, first assessor) for the 2 subgroup analyses: those with and without radiographic pneumonia and those <5 years of age or ≥5 years of age. There were no substantial differences in IRR (ie, there is overlap of all CIs) of findings in those with and without radiographic pneumonia (Supplemental Fig 2). There were no substantial differences in IRR of findings in those <5 years old and those ≥5 years of age, with the exception of retractions, for which the IRR appears substantially greater in older children (κ = 0.81; 95% CI, 0.61 to 0.96) compared with younger children (κ = 0.42; 95% CI, 0.20 to 0.63) (Supplemental Fig 3).

DISCUSSION

We found fair to moderate IRR for most physical examination findings in children who presented to the ED with suspected CAP. Three findings (wheezing, retractions, and respiratory rate) met the definition of acceptable reliability, with a lower confidence limit of >0.4.[17–21] IRR was largely similar between the strata

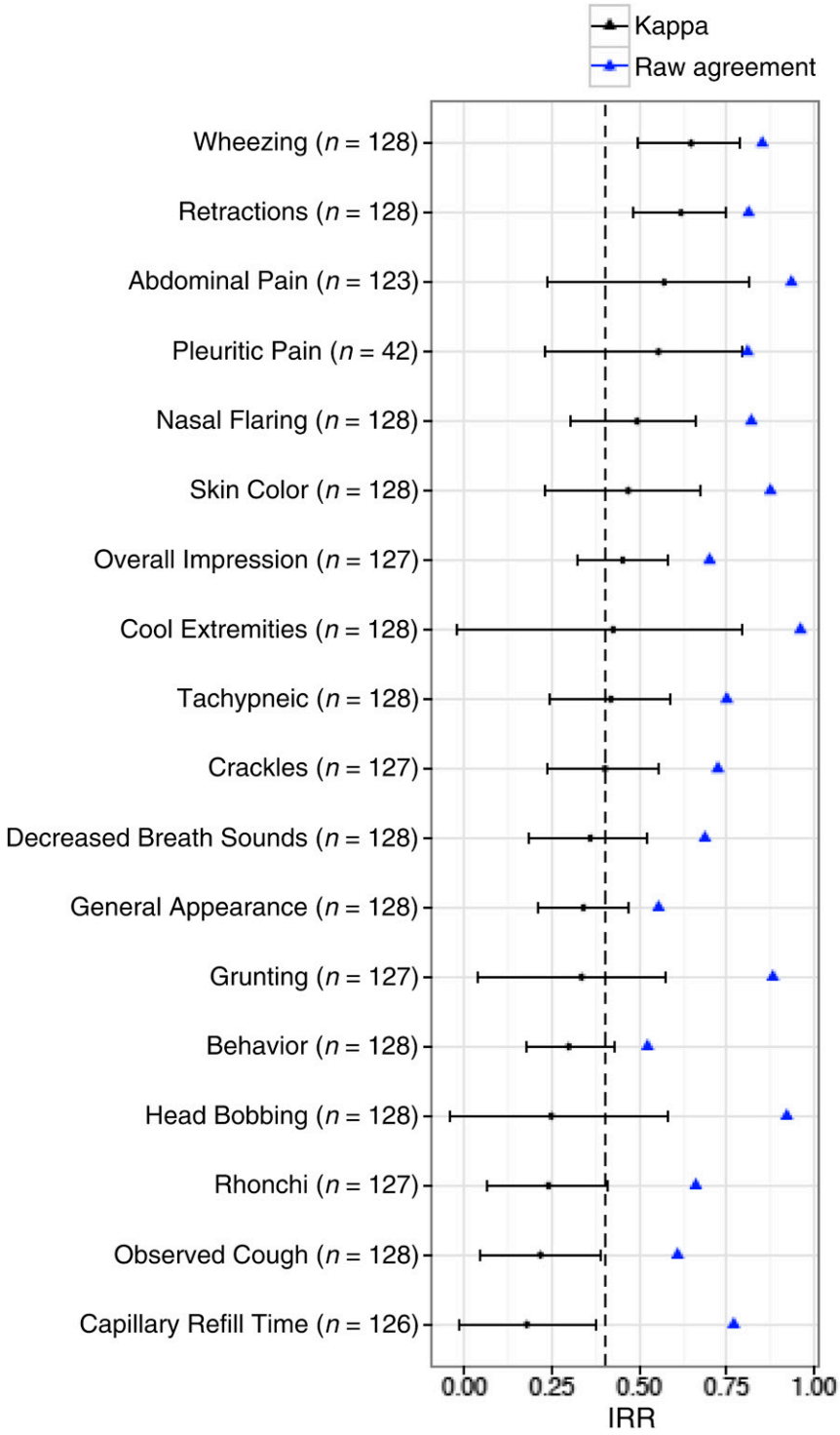

FIGURE 1

Agreement and IRR of examination findings in children with suspected CAP. Dots with bars represent the κ statistic with 95% CIs. Triangles represent raw agreement. The dotted line represents the acceptable level of agreement for the lower 95% confidence limit (κ < 0.40).

(pneumonia versus not, age <5 years vs ≥5 years).

Authors of previous studies who examined the IRR of pediatric respiratory examination findings focused on children with asthma, bronchiolitis, or wheezing. The authors of these previous studies demonstrate conflicting results, reflecting the heterogeneity of respiratory findings and populations examined.[5–8,23] Our results in children with suspected CAP are generally consistent with previous studies in other respiratory conditions. The IRR for respiratory rate ranges from κ or ICC of 0.58 to 0.95, although values as low as 0.12 have been reported.[5–8,23–27] Similarly, retractions and wheezing also have large ranges of reported κ statistics across studies (κ = 0.25–0.77 for retractions, 0.31–0.78 for wheezing).[5–8,23–27]

Several other examination findings that are considered important in making the clinical diagnosis of pneumonia (eg, crackles/rales and decreased breath sounds) demonstrated poor to moderate IRR. In adults with suspected CAP, the IRR for crackles/rales and rhonchi ranged from 0 to 0.64.[4] The reliability of auscultatory sounds other than wheezing is not often reported in children. The authors of 1 study of examination findings in preschoolers with acute cough found a κ of 0.39 (95% CI, 0.26 to 0.53) for abnormal chest findings, which included wheezing or crackles.[7] Decreased breath sounds had only fair IRR in our study, which is in contrast to reported κ statistics in children with asthma (0.67–0.76).[28] Differences between the IRR in children with suspected CAP and those with other conditions, such as asthma, likely reflect differences in pathophysiology. Given the importance placed on these findings in clinically diagnosing CAP in children, it is concerning that they demonstrated only fair reliability.

There are several possible reasons for the modest IRR of examination findings in children with suspected CAP. Clinical providers have different descriptions for the adventitious sounds discovered on lung auscultation (eg, coarse crackles versus rhonchi). For

TABLE 2 Prevalence of Examination Findings by Radiographic Pneumonia Diagnosis and Age

Finding	Radiographic Pneumonia, *n* (%)		Age, *n* (%)	
	No Pneumonia (*n* = 105)	Pneumonia (*n* = 23)	<5 y (*n* = 81)	≥ 5 y (*n* = 47)
General examination				
General appearance				
Well	31 (30)	4 (17)	19 (23)	16 (34)
Mildly ill or distressed	39 (37)	10 (43)	34 (42)	15 (32)
Moderately ill or distressed	32 (30)	8 (35)	25 (31)	15 (32)
Severely ill or distressed	3 (3)	1 (4)	3 (4)	1 (2)
Behavior				
Playing and appropriate	28 (27)	5 (22)	17 (21)	16 (34)
Quiet but appropriate	45 (43)	12 (52)	31 (38)	26 (55)
Sleeping but easily arousable	10 (10)	1 (4)	9 (11)	2 (4)
Fussy but consolable	17 (16)	3 (13)	18 (22)	2 (4)
Irritable	5 (5)	1 (4)	6 (7)	0
Lethargic, confused, or reduced response to pain	0	1 (4)	0	1 (2)
Respiratory examination				
Crackles	35 (33)	14 (61)	35 (43)	14 (30)
Decreased breath sounds	36 (34)	19 (83)	24 (30)	31 (66)
Wheezing	36 (34)	5 (22)	30 (37)	11 (23)
Retractions	60 (57)	13 (57)	55 (68)	18 (38)
Rhonchi	38 (36)	5 (22)	35 (43)	8 (17)
Tachypneic	72 (69)	17 (74)	61 (75)	28 (60)
Pleuritic chest pain	7 (20)	7 (41)	0 (0)	14 (33)
Nasal flaring	24 (23)	6 (26)	21 (26)	9 (19)
Grunting	11 (10)	3 (13)	10 (12)	4 (9)
Head bobbing	9 (9)	0 (0)	8 (10)	1 (2)
Observed cough	51 (49)	17 (74)	44 (54)	24 (51)
Cardiovascular and perfusion examination				
Cool extremities	2 (2)	2 (9)	4 (5)	0 (0)
Skin color				
Pink and/or normal	91 (87)	18 (78)	69 (85)	40 (85)
Pale or dusky	13 (12)	5 (22)	11 (14)	7 (15)
Mottled	1 (1)	0	1 (1)	0
Capillary refill time (s)				
1–2	91 (87)	18 (78)	69 (85)	40 (85)
3	14 (13)	5 (22)	12 (15)	7 (15)
Peripheral pulses				
Normal	102 (97)	22 (96)	80 (99)	44 (96)
Bounding	2 (2)	1 (4)	1 (1)	2 (4)
Abdominal examination				
Abdominal pain	10 (10)	4 (17)	4 (5)	10 (22)
Overall impression				
Mild	51 (49)	4 (17)	35 (43)	20 (44)
Moderate	46 (44)	17 (74)	40 (49)	23 (50)
Severe	8 (8)	1 (4)	6 (7)	3 (6)

Findings reflect the examination of the primary treating clinician (ie, first assessor).

example, the authors of 1 study of 12 physicians classifying lung audiovisual recordings found poor to fair agreement ($\kappa < 0.4$) when using detailed descriptions of adventitious sounds, such as coarse crackles versus fine crackles. Interrater agreement improved when findings were combined in more general terms (crackles overall [$\kappa = 0.62$] and wheezes overall [$\kappa = 0.59$]), suggesting that descriptions of auscultation findings differ across physicians.[29] To address this, we provided general descriptions of examination findings (eg, "crackles/rales" instead of "coarse crackles/rales" vs "fine crackles/rales") to capture real-world clinical practice. Despite keeping terminology general, clinicians likely labeled auscultation findings differently, contributing to the observed fair to moderate IRR. It is also likely that physicians use different terms (eg, rhonchi versus coarse crackles) for the same auscultatory findings or that findings are misinterpreted or indistinguishable (eg, transmitted upper airway sounds versus lower respiratory rhonchi). The poor to fair reliability for rhonchi and crackles suggests that this might be the case.

Age may also play a role in explaining our results because certain findings may be more reliable on the basis of age because of differences in size and pathophysiology. We found the reliability of retractions was substantially higher in older children compared with those <5 years of age, likely because of retractions being more pronounced in larger children and adolescents. Certain examination findings were also more prevalent by age, such as retractions and rhonchi in younger children, likely reflecting the higher prevalence of viral lower respiratory tract disease in this age group.

The fair reliability that we observed may also be due to changes in physical examination education, practice, and precision over time. Advanced imaging and laboratory testing has potentially supplanted the examination for the current generation of physicians.[30] Researchers for 1 study who examined the competency of the cardiac examination across years of training found that cardiac examination skills improve from the beginning to the end of medical school, but skills do not improve after the third year of medical school and may decline after years in practice.[31] The conflict over advocating that clinical skills are of less importance because of the improvements in diagnostic technologies competes with others stressing that clinical history and examination findings are still critical.[32,33] In the case of pediatric pneumonia, the reference diagnostic standard (chest radiography) has moderate IRR and is imprecise.[34] Thus, the diagnosis is often a clinical one rather than one based on imaging or laboratory testing, further emphasizing the importance of a reliable physical examination.

These findings have several important implications for clinical care and research. Authors of national guidelines recommend against the use of radiography to confirm suspected CAP in children treated in the outpatient setting.[3] If the examination is not sufficiently reliable, it will have limited use in clinical practice, thus contributing to the wide variation in management of children who present to the ED with febrile respiratory illnesses.[1,2] Furthermore, because the Infectious Diseases Society of America and Pediatric Infectious Diseases Society pneumonia guidelines do not recommend radiography in the outpatient setting, the reliance of examination findings with limited reliability to diagnose CAP has the potential to increase antibiotic overuse, resulting in the spread of antimicrobial resistance, antibiotic-associated adverse effects, and increased cost. It is reassuring that the findings of respiratory rate and retractions, often abnormal in pneumonia, had acceptable reliability. However, the limited reliability of many other findings that are hallmarks of the clinical diagnosis of CAP suggests that either interventions to improve examination skills are necessary, a more standardized approach to the diagnosis is required, or more objective tools are needed to aid in CAP diagnosis in children.

This study has several limitations. The study was conducted in an academic pediatric ED; thus, our results may not be generalizable to other clinical settings. Several facets of the respiratory examination may change, even in a brief time, which may underestimate κ. Given that almost all of our paired examinations occurred within 20 minutes of each other, this is unlikely to substantially alter our results. The κ statistic is affected by the prevalence of the examination finding; for uncommon findings, low κ values may not necessarily reflect low levels of agreement.[35] An example of a finding that had a low prevalence of positive findings, resulting in a high agreement with lower κ, is head bobbing. High agreement with low κ is a known phenomenon in cases in which there are imbalances in the prevalence of findings, which can occur if the prevalence of a finding approaches 0% or 100%. It has been shown that low values of κ because of marginal imbalances in the prevalence of findings, even when absolute agreement is high, cannot be dismissed as an unfair penalty and should be interpreted as truly indicating lesser degrees of agreement beyond those expected by chance.[36] Our sample size calculation was based on a prevalence rate of 5%; therefore, for a few findings with lower prevalence (ie, cool extremities, head bobbing, and peripheral pulses), the study was not adequately powered. For findings with prevalence <5% and those of our subanalyses with sample sizes <122, they may have been classified as unacceptable, but with an increased prevalence or sample size, it may be that these findings are indeed acceptable because the lower CI potentially may increase.

Some children may not have had clear pneumonia on a CXR at time of visit, causing us to misclassify these patients as not having pneumonia. This would not affect the overall κ because all patients were included in those calculations. Similarly, ∼20% of our cohort had radiographic pneumonia, limiting our ability to provide definitive results of IRR in those with radiographic pneumonia. It was our objective, however, to examine the IRR of examination findings in suspected pneumonia, which reflects real-world conditions of performing the examination first and using these findings to decide if a radiograph or treatment of pneumonia is warranted. There were 7 patients who received breathing treatments between examinations, which was a protocol deviation and not an exclusion criterion. We

performed a sensitivity analysis removing those 7 children, and results did not change substantively, suggesting that these treatments did not alter the IRR of the examination findings (Supplemental Fig 4). A large proportion of assessors knew the CXR results, which may have influenced reporting of certain examination findings. The proportion of first and second assessors privy to radiograph results was similar, however, and thus we would not expect differential bias between assessors. In addition, some second assessors may have been privy to more clinical information than others; however, given the random selection of the second assessor, knowledge of clinical history and radiograph results would be randomly distributed. Thus, there is a low likelihood that systematic bias would be introduced as a result. Finally, we did provide detailed definitions of examination findings to examiners because we intended to capture real-world interpretation of the physical examination.

Despite these limitations, our study has several notable strengths, namely its prospective, real-world approach. The large number of raters involved in our study replicates the real-world setting and contributes to generalizability. In addition, despite the competing demands of a busy ED, nearly all examinations occurred within 20 minutes and without intervening treatments. Our ability to incorporate a large number of examination findings that are commonly used to diagnose CAP clinically makes this the most extensive study to our knowledge to address the reliability of examination findings in suspected CAP.

CONCLUSIONS

We found fair to moderate reliability of many findings thought to be important to the clinical diagnosis of pneumonia. Only 3 findings (retractions, wheezing, and respiratory rate) had acceptable levels of IRR. The reliability of these findings must be considered in the clinical management and research of children with CAP.

ACKNOWLEDGMENTS

We acknowledge Shreya Reddy for her assistance with the literature review; Judd Jacobs and Jessi Lipscomb for data management and programming support; Kerry Aicholtz for administrative assistance; Mekibib Altaye, PhD, and Heidi Sucharew, PhD, for statistical supervision; the Clinical Research Coordinator team, physicians, and advanced practice and staff nurses of the Division of Emergency Medicine at CCHMC who assisted with study enrollment and procedures; and all the patients and families who participated in CARPE DIEM.

ABBREVIATIONS

CAP: community-acquired pneumonia
CARPE DIEM: Catalyzing Ambulatory Research in Pneumonia Etiology and Diagnostic Innovations in Emergency Medicine
CCHMC: Cincinnati Children's Hospital Medical Center
CI: confidence interval
CXR: chest radiograph
ED: emergency department
ICC: intraclass correlation coefficient
IRR: interrater reliability
LRTI: lower respiratory tract infection
PEM: pediatric emergency medicine

This study was presented in abstract form at the annual meeting of the Pediatric Academic Societies; April 30, 2016; Baltimore, MD.

DOI: https://doi.org/10.1542/peds.2017-0310

Accepted for publication Jun 1, 2017

Address correspondence to Todd A. Florin, MD, MSCE, Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, ML 2008, Cincinnati, OH 45229. E-mail: todd.florin@cchmc.org

PEDIATRICS (ISSN Numbers: Print, 0031-4005; Online, 1098-4275).

FINANCIAL DISCLOSURE: The authors have indicated they have no financial relationships relevant to this article to disclose.

FUNDING: This work was supported by a pediatric research grant from The Gerber Foundation (to T.A.F.); the National Center for Research Resources and the National Center for Advancing Translational Sciences (grant 8 KL2 TR000078-05 to T.A.F.); the National Institute for Allergy and Infectious Diseases and the National Institutes of Health (grant 1 K23 AI121325-01 to T.A.F.); and a Trustee Award from Cincinnati Children's Hospital Medical Center (to L.A.). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. Funded by the National Institutes of Health (NIH).

POTENTIAL CONFLICT OF INTEREST: The authors have indicated they have no potential conflicts of interest to disclose.

REFERENCES

1. Florin TA, French B, Zorc JJ, Alpern ER, Shah SS. Variation in emergency department diagnostic testing and disposition outcomes in pneumonia. *Pediatrics*. 2013;132(2):237–244
2. Shah S, Bourgeois F, Mannix R, Nelson K, Bachur R, Neuman MI. Emergency department management of febrile respiratory illness in children. *Pediatr Emerg Care*. 2016;32(7):429–434
3. Bradley JS, Byington CL, Shah SS, et al; Pediatric Infectious Diseases Society and the Infectious Diseases Society of America. The management of community-acquired pneumonia in infants and children older than 3 months of age: clinical practice guidelines by the Pediatric Infectious Diseases Society and the Infectious Diseases Society of America. *Clin Infect Dis*. 2011;53(7):e25–e76
4. Wipf JE, Lipsky BA, Hirschmann JV, et al. Diagnosing pneumonia by physical examination: relevant or relic? *Arch Intern Med*. 1999;159(10):1082–1087
5. Biondi EA, Gottfried JA, Dutko Fioravanti I, Schriefer JA, Aligne CA, Leonard MS. Interobserver reliability of attending physicians and bedside nurses when using an inpatient paediatric respiratory score. *J Clin Nurs*. 2015;24(9–10):1320–1326
6. Gajdos V, Beydon N, Bommenel L, et al. Inter-observer agreement between physicians, nurses, and respiratory therapists for respiratory clinical evaluation in bronchiolitis. *Pediatr Pulmonol*. 2009;44(8):754–762
7. Hay AD, Wilson A, Fahey T, Peters TJ. The inter-observer agreement of examining pre-school children with acute cough: a nested study. *BMC Fam Pract*. 2004;5:4
8. Wang EE, Milner RA, Navas L, Maj H. Observer agreement for respiratory signs and oximetry in infants hospitalized with lower respiratory infections. *Am Rev Respir Dis*. 1992;145(1):106–109
9. Jain S, Williams DJ, Arnold SR, et al; CDC EPIC Study Team. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med*. 2015;372(9):835–845
10. Lynch T, Platt R, Gouin S, Larson C, Patenaude Y. Can we predict which children with clinically suspected pneumonia will have the presence of focal infiltrates on chest radiographs? *Pediatrics*. 2004;113(3, pt 1). Available at: www.pediatrics.org/cgi/content/full/113/3/e186
11. Mahabee-Gittens EM, Grupp-Phelan J, Brody AS, et al. Identifying children with pneumonia in the emergency department. *Clin Pediatr (Phila)*. 2005;44(5):427–435
12. Bilkis MD, Gorgal N, Carbone M, et al. Validation and development of a clinical prediction rule in clinically suspected community-acquired pneumonia. *Pediatr Emerg Care*. 2010;26(6):399–405
13. Neuman MI, Monuteaux MC, Scully KJ, Bachur RG. Prediction of pneumonia in a pediatric emergency department. *Pediatrics*. 2011;128(2):246–253
14. Fung KP, Lee J. Bootstrap estimate of the variance and confidence interval of kappa. *Br J Ind Med*. 1991;48(7):503–504
15. Marin JR, Bilker W, Lautenbach E, Alpern ER. Reliability of clinical examinations for pediatric skin and soft-tissue infections. *Pediatrics*. 2010;126(5):925–930
16. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? *BMC Med Res Methodol*. 2016;16:93
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174
18. Gorelick MH, Atabaki SM, Hoyle J, et al; Pediatric Emergency Care Applied Research Network. Interobserver agreement in assessment of clinical variables in children with blunt head trauma. *Acad Emerg Med*. 2008;15(9):812–818
19. Holmes JF. Clinical prediction rules. In: Li G, Baker SP, eds. *Injury Research: Theories, Methods, and Approaches*. New York, NY: Springer; 2012:317–336
20. Kuppermann N, Holmes JF, Dayan PS, et al; Pediatric Emergency Care Applied Research Network (PECARN). Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. *Lancet*. 2009;374(9696):1160–1170
21. Yen K, Kuppermann N, Lillis K, et al; Intra-abdominal Injury Study Group for the Pediatric Emergency Care Applied Research Network (PECARN). Interobserver agreement in the clinical assessment of children with blunt abdominal trauma. *Acad Emerg Med*. 2013;20(5):426–432
22. *R: A Language and Environment for Statistical Computing* [computer program]. Version 3.3. Vienna, Austria: R Foundation for Statistical Computing; 2016
23. Angelilli ML, Thomas R. Inter-rater evaluation of a clinical scoring system in children with asthma. *Ann Allergy Asthma Immunol*. 2002;88(2):209–214
24. Liu LL, Gallaher MM, Davis RL, Rutter CM, Lewis TC, Marcuse EK. Use of a respiratory clinical score among different providers. *Pediatr Pulmonol*. 2004;37(3):243–248
25. Eggink H, Brand P, Reimink R, Bekhof J. Clinical scores for dyspnoea severity in children: a prospective validation study. *PLoS One*. 2016;11(7):e0157724
26. Bekhof J, Reimink R, Bartels IM, Eggink H, Brand PL. Large observer variation of clinical assessment of dyspnoeic wheezing children. *Arch Dis Child*. 2015;100(7):649–653
27. Stevens MW, Gorelick MH, Schultz T. Interrater agreement in the clinical evaluation of acute pediatric asthma. *J Asthma*. 2003;40(3):311–315
28. Ducharme FM, Chalut D, Plotnick L, et al. The pediatric respiratory assessment measure: a valid clinical score for assessing acute asthma severity from toddlers to teenagers. *J Pediatr*. 2008;152(4):476–480, 480.e1
29. Melbye H, Garcia-Marcos L, Brand P, Everard M, Priftis K, Pasterkamp H. Wheezes, crackles and rhonchi: simplifying description of lung sounds increases the agreement on their classification: a study of 12 physicians' classification of lung sounds from video recordings. *BMJ Open Respir Res*. 2016;3(1):e000136

30. Feddock CA. The lost art of clinical skills. *Am J Med*. 2007;120(4):374–378

31. Vukanovic-Criley JM, Criley S, Warde CM, et al. Competency in cardiac examination skills in medical students, trainees, physicians, and faculty: a multicenter study. *Arch Intern Med*. 2006;166(6):610–616

32. Reilly BM. Physical examination in the care of medical inpatients: an observational study. *Lancet*. 2003;362(9390):1100–1105

33. Patel N, Ngo E, Paterick TE, Chandrasekaran K, Tajik J. Should doctors still examine patients? *Int J Cardiol*. 2016;221:55–57

34. Lynch T, Bialy L, Kellner JD, et al. A systematic review on the diagnosis of pediatric bacterial pneumonia: when gold is bronze. *PLoS One*. 2010;5(8):e11989

35. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43(6):543–549

36. Gorelick MH, Yen K. The kappa statistic was representative of empirically observed inter-rater agreement for physical findings. *J Clin Epidemiol*. 2006;59(8):859–861