Impact of Training on Observer Variation in Chest Radiographs of Children with Severe Pneumonia

Archana B. Patel[1,2], Amit Amin [2], S. Z. Sortey[3], Ambarish Athawale[2] and Hemant Kulkarni[2]

From the [1]Department of Pediatrics and Clinical Epidemiology Unit, Indira Gandhi Government Medical College, Nagpur, and [2]Lata Medical Research Foundation, Nagpur, [3]Department of Radiology, Indira Gandhi Government Medical College, Nagpur, India.

Correspondence to: Dr. Archana Patel, 125, Opp. Tidke Vidyalay, Katol Road, Nagpur 440 013, India. E-mail: archana_patel@vsnl.com

Manuscript received: June 14, 2006; Initial review completed: May 18, 2007; Revision accepted: June 12, 2007.

Background : *Pneumonia diagnosed using chest radiographs is often used as a study end point in trials and epidemiological studies. We studied whether training of the end-users in 172 standardized chest radiographic features will decrease variability in the interpretation.* ***Methods:*** *Inter-observer variation of 3 observers in recognizing standardized radiographic features for pneumonia was studied in 172 chest radiographs of children with clinical severe pneumonia. (as per WHO definition). The observers were then trained using a software with a repository of normal and abnormal films showing a spectrum of radiological changes in pneumonia. The inter-observer variation in recognizing the same standardized radiographic features was recorded after this training. For each radiographic feature, Cohen's kappa statistics to assess the between-observer agreement and Fleiss's multiple rater kappa statistics to assess agreement among all three clinicians was used.* ***Results :*** *The 'uniterpretable' films reduced from 16.6% (95% CI 0%-34.1%) before training to 8.1% (95% CI0%-17.7%) after training. The 'adequate' films increased from 54.2% (95% CI 12.5%-95.9%) before training to 70% (95% CI 46.5%-93.4%) after training. For all features, agreement between observers 1 with 2 and 1 with 3, the Cohen's kappa improved from poor to moderate agreement. The Fleiss's kappa values before training were 0.1 to 0.2 and after training ranged from 0.37 to 0.52 indicating moderate to good agreement after training.* ***Conclusions****: Training of the doctors using standardized features with the help of a software improves agreement substantially in identifying radiological pneumonia.*

Key words: *Pneumonia, Radiography, Software, Training.*

PNEUMONIA is the leading cause of childhood death in developing countries contributing globally to 21% of deaths in under-five children(1). World Health Organization's (WHO) clinical case definitions for diagnosing pneumonia are sensitive and appropriate when the consequences of missed diagnosis are serious and useful for treatment decisions, but have low positive predictive value (2,3). Therefore, for epidemiological pur-poses, a method for diagnosing pneumonia can be chest radiography which also allows the readers to be blinded to the intervention or clinical course of the patient. However inter-observer variation in the interpretation of chest radiographs of children with ARI is well known as there exists no strict radiological definition of pneumonia(4-6). Standardization, simplification and categorization of radiological features along with training can help mitigate this problem(7). For the purposes of vaccine efficacy trial, WHO established a working group to standardize the categorization of radiological pneumonia, for the purpose of establishing the burden estimates of likely bacterial pneumonia and estimating vaccine impact(7). Using this categorization of radiologic features of pneumonia, we undertook this study to estimate the inter-observer agreement in interpretation of chest radiographs in children with severe pneumonia and to examine whether the WHO training intervention increases the inter-observer agreement.

Subjects and Methods

The present study was conducted at Indira Gandhi Government Medical College and Hospital (IGGMC), a tertiary care center at Nagpur, India. This site participated in the Amoxicillin Penicillin Pneumonia International Study (APPIS) which was a multicenter randomized, study conducted in 8 countries to determine whether oral amoxicillin and parenteral penicillin were equivalent in the treatment

of WHO-defined severe pneumonia (fast breathing with lower chest wall indrawing) in children aged 3-59 months(8). The chest radiographs of 172 of 200 children with WHO-defined severe pneumonia were assigned a unique code number to maintain confidentiality and to blind the observers.

Three observers *i.e.* a pediatric faculty (ABP), a radiology faculty (SZS) and a radiology resident (APA), blinded to each other's observations, independently read the chest radiographs before undergoing training with the standardized software. Based on their experience they recorded what they thought was: film adequacy, significant pathology in lung fields, consolidation on the left side, consolidation on the right side, other infiltrates/abnormality on the left side, other infiltrates/abnormality on the right side, pleural effusion on the left side and pleural effusion on the right side. These were recorded as "before-training readings". Standardized training by the WHO expert, using the software began a week after these readings. After training, the three clinicians re-interpreted the 172 chest radiographs in random order, under the same conditions. These were the "after-training readings".

The training intervention

AiMS multimedia is a commercial software with a repository of normal and abnormal films showing a spectrum of radiological changes in pneumonia. It contains training, teaching and assessment tutorials. A team of experts for WHO vaccine trials standardized the observed radiological changes of these chest radiographs (*Table I*). Each observer (ABP, SZS and APA) was trained for 3 days to recognize the standard radiological features under the supervision of a WHO representative who was also a member of the team for development of the software and its training program for the vaccine trials. The software has several pneumonia tutorial sets (20 chest radiographs in each). The observers have to read these radiographs and their responses are compared to that provided by the software. The responses of the observers had to have a sensitivity and specificity of 80% with the software responses to be successfully trained.

Statistical analysis

Using the statistical method of latent class analysis which uses the consensus among the observers as a measure of the estimated 'true' prevalence, we estimated the prevalence of 'uniterpreatable' and 'adequate' films before and after training. For each described radiographic feature (*Table I*), its Cohen's kappa statistic was used to assess between-observer agreement between pairs of observers, and, to assess agreement among all three clinicians Fleiss's multiple rater kappa statistic was used(9,10). Unweighted kappa was used for radiological features with two outcome categories (*e.g.* 'Yes' or 'No' outcome) while weighted kappa was used for features with more than two ordered categories (*e.g.* 'uninterpretable', 'suboptimal' and 'normal' category for "Film Adequacy"). We defined 'unanimity' as a complete agreement by all observers on each category of a radiographic feature. For instance, 'unanimity' was said to exist if all three observers agreed on the presence or absence of consolidation on the left side. To estimate the strength of association between training and 'unanimity' for each radiographic feature, univariate logistic regression models were used. Agreement 1.0 (Lata Medical Research Foundation, Nagpur, India) and Stata 7.0 (Stata Corp, College Station, TX, USA) software programs were used for analysis.

Results

The prevalence of 'uniterpreatable' films was 16.6% before training and significantly reduced to 8.1% after training (P = 0.000). For 'adequate' films, it was 54.22% before training and significantly increased to 70% after training (P = 0.000). There was 'unanimity' for absence of pleural effusion on the left side before and after training. Observers ABP, SZS and APA identified 2, 2 and 1 films, respectively as having pleural effusion on the right side before training. This lack of pre-training 'unanimity' improved after training and right pleural effusion was now identified by all, in just one film. For further agreement analysis, we only excluded the films classified as uninterpretable by at least one observer, and, those of pleural effusion as the numbers were small.

A comparison of pair-wise Cohen's kappa values before and after training is shown in *Table II.* For agreement between observers ABP and SZS, there

was a significant improvement for all features except infiltrates on the left side. For observers ABP and APA the improvement in agreement was significant for all the features. For the pair of observers SZS and APA, a significant improvement in agreement was seen only for the feature of primary end point consolidation on the left side – all other kappa values for this pair of observers being moderate-to-high even before training.

Figure 1 shows the multiple rater Fleiss's kappa estimates before and after training. It was observed that the training intervention contributed to an improved agreement among the three observers for all the radiographic features. The Fleiss's kappa values after training ranged from 0.37 to 0.52 indicating moderate to good agreement and were highly significant at Z values ranging from 7.9 to 11.4. Maximum improvement in Fleiss's kappa was observed for primary end point consolidation on the left side followed by the one for infiltrates on right side.

Finally, we assessed whether 'unanimity' improved significantly after training by using a logistic regression model. *Table III* summarizes the results of the logistic regression analysis. It was again observed that the best improvement in 'unanimity' was for the feature of primary end point consolidation on left side followed by the one for the feature of infiltrates on the right side.

Discussion

This study emphasizes the importance and benefit of standardizing the interpretation of the chest radiograph using a training intervention. Chest radiographs are often used in epidemiological studies and for antimicrobial or vaccine clinical trials, to determine the outcome of pneumonia(11,12). The reporting of this important study outcome is often difficult and ambiguous if there is a lack of agreement between observers. This can contribute to bias and misclassification(13). Reported agreements between readers also vary from study to study (4,6,14).

There are many possible reasons for a lack of agreement between observers. Firstly, a wide spectrum of radiological findings are observed in children such as a typical appearance of lobar consolidation of bacterial pneumonia to mild interstitial and perihilar changes often associated with bacterial, viral infections, asthma or normal children. The varied radiological manifestations in patients of Human Immunodeficiency Virus may further complicate the issue(6). Secondly, clinicians can describe the radiographic features in different terminologies and also a single feature can have different grades. So standardizing definitions for radiological features and simplification of their grades can improve agreement(15,16). Thirdly, the observers can be of different specialization and experience(6). In this study, radiologist observers (SZS and APA) had less disagreement whereas the pediatrician's observations were more in disagreement with radiologist colleagues.

In this study, there was a significant agreement for all radiological features subsequent to training. Although the outcome of infilterates were scaled on a simple two point scale as presence or absence, training further enhanced agreement. Overall the reporting of "adequate" films improved and the diagnosis of "significant pathology" decreased. This shows that training even experienced clinicians could help reduce the number of repeat orders for chest radiographs, when judged as "uninterpretable" in clinical practice. It also helped to decrease an over interpretation of abnormality, which often can occur if the observers know that they are reading chest radiographs of children already diagnosed with pneumonia.

Previously, the agreement for these simplified and standardized WHO defined radiographic features was studied on 20 radiologists and clinicians with a reference reading but improvement with training was not assessed(17). The agreement for any abnormality ranged from 71-85% with a range of kappa from 0.31-0.68. The post training agreement between observers in our study is similar to that reported between trained readers and the reference standard in the WHO study. This simplified standardized method of reporting chest radiographic features without training has been used in a double blind randomized trial of 9–valent pneumococcal conjugate vaccine enrolling 39,836 children in South Africa(11). However our study showed that merely standardization and simplification may not be enough to achieve even

TABLE I–*WHO Standardized Chest Radiographic Features in Pneumonia*

Film description	Definition
Quality	
Uninterpretable	Image are not interpretable in terms of presence or absence of "primary end-point" without additional images
Suboptimal	Interpretation of primary end-point but not of other infiltrates or findings
Adequate	Confident interpretation of end-point as well as other infiltrates
Radiological features	
Significant pathology	Presence of consolidation, infiltrates or effusion.
Primary end-point(PEP) consolidation	Dense, often homogeneous, confluent alveolar infiltrate sometimes may encompass an entire lobe or large segment;: fluffy, mass-like, cloud-like density, erases heart and diaphragm borders (silhouette sign); often contains air bronchograms and sometimes associated with pleural effusion
Other infilterates of abnormality	Linear and patchy densities (interstitial infiltrate) in a lacy pattern involving both lungs, featuring peribronchial thickening and multiple areas of atelectasis. Lung inflation is normal to increased. It also includes minor patchy infiltrates that are not of sufficient magnitude to constitute primary end-point consolidation, and small areas of atelectasis which in children can be difficult to distinguish from consolidation
Pleural effusion	Fluid in the pleural space between the lung and chest wall, at the costo-phrenic angle or as a layer adjacent to the lateral chest wall, but not in the horizontal or oblique fissures. It is considered as "primary end-point" (PEP) if it is in the lateral pleural space and is spatially associated with a pulmonary parenchymal infiltrate (including other infiltrate) OR if the effusion obliterates enough of the hemithorax to obscure an opacity.

TABLE II–*Comparison of Pair-wise Agreement (Cohen's kappa) Before and After Training for all Radiographic Features*

Radiographic feature	Before training			After training		
Pairs of Observers*	1 & 2	1 & 3	2 & 3	1 & 2	1 & 3	2 & 3
Film Adequacy	0.035	0.056	0.39	0.35	0.37	0.37
Significant pathology	0.12	0.064	0.41	0.63	0.44	0.50
PEP consolidation left	–0.01	–0.013	0.11	0.39	0.66	0.50
PEP consolidation right	–0.07	–0.03	0.36	0.51	0.54	0.46
Other infiltrates left	0.21	0.06	0.18	0.47	0.34	0.43
Other infiltrates right	0.13	–0.00	0.19	0.50	0.36	0.31
Conclusion	0.06	0.05	0.38	0.57	0.46	0.50

* Observer 1 is pediatric faculty (ABP), Observer 2 is radiology faculty (SZS) and Observer 3 is radiology resident (APA),

moderate agreement and training helps improve agreement.

Finally, in spite of a significant improvement in agreement for all the radiographic features, the post-training kappa values indicated only moderate-to-good agreement. The training was provided using computer images but chest radiographs of patients were read on the viewbox, which could be one of the limitations in achieving better agreement. Also it endorses the fact that radiological diagnosis of chest radiographs in children with severe pneumonia are

TABLE III–*Effect of Training on Agreement for Various Radiographic features.*

Feature	Training		Odds ratio*	95% CI
	No	Yes		
Film adequacy				
Non-unanimous	164	69		
Unanimous	68	103	2.28	1.48-3.52
Significant pathology				
Non-unanimous	96	60		
Unanimous	76	112	2.36	1.53-3.64
PEP consolidation left				
Non-unanimous	15	3		
Unanimous	157	169	5.38	1.53-18.95
PEP consolidation right				
Non-unanimous	34	20		
Unanimous	138	152	1.87	1.03- 3.41
Other infiltrates left				
Non-unanimous	108	60		
Unanimous	64	112	3.15	2.03-4.89
Other infiltrates right				
Non-unanimous	123	71		
Unanimous	49	101	3.57	2.28-5.60
Conclusion				
Non-unanimous	123	76		
Unanimous	49	96	3.17	2.03-4.96
Total	172	172		

Odds ratio represents the times-likelihood of unanimous agreement consequent to the training intervention.

inherently subject to a substantial inter-observer variation which reinforces the importance of standardization and training.

Acknowledgements

We are thankful to the APPIS project for providing the chest radiographs, Dr. Shamim Qazi for prwoviding the WHO standard training software package and to Dr. Thomas Cherian, WHO for imparting training to observers. We are also thankful to Dr. Jon Simon, Director, Center for International Health, Boston University for reviewing the manuscript and to Elizabeth Bertone, DSc for her help in analysis and reviewing the manuscript. We are indebted to Smita Puppalwar for her help in the study.

Contributors: ABP and AA designed the study, collected the data and wrote the manuscript. AA and HK analyzed data and also participated in writing manuscript. SZS and AAt assisted in data collection and its management.

Funding: World Health Organization

Competing interests: None

REFERENCES

1. Black RE, Morris SS, Bryce J. Where and why are 10 million children dying every year? Lancet 2003; 361: 2226-2234.
2. World Health Organization (WHO). Acute respiratory tract infections in children: case management in small hospitals in developing countries. WHO/ARI/90.5. Geneva: WHO, 1994.
3. Mulholland EK, Simoes EA, Costales MO, McGrath EI, Manalac EM, Gove S. Standardized diagnosis of pneumonia in developing countries. Pediatr Infect Dis J 1992; 11: 77-81.
4. Swingler GH. Observer variation in chest radiography of acute lower respiratory infections in children: a systematic review. BMC Med Imaging 2001; 1:1.
5. Sivit CJ, Miller CR, Rakusan TA, Ellaurie M, Kushner DC. Spectrum of chest radiographic

What this Study Adds?

- Chest radiographs in children with pneumonia are inherently subject to substantial inter-observer variation.
- Simplified standardized system of categorization of radiographic features and training using software reduces inter-observer variation.

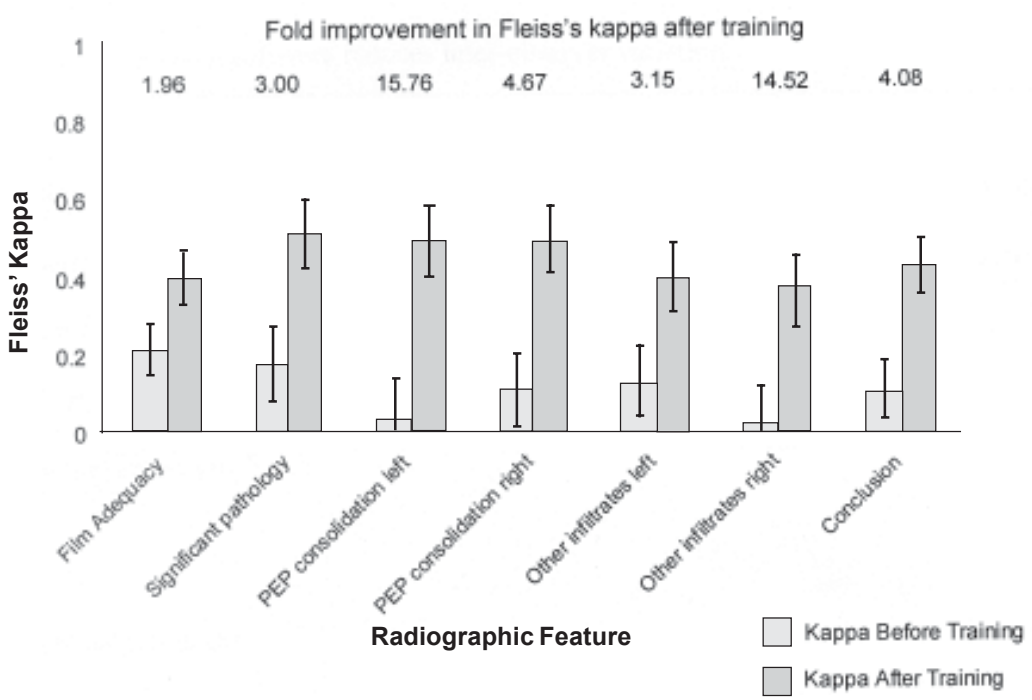

Fig. 1. Agreement among the three observers (multiple rater Fleiss's kappa estimates) for all the radiographic features

abnormalities in children with AIDS and Pneumocystis carinii pneumonia. Pediatr Radiol 1995; 25: 389-392.

6. Davies HD, Wang EE, Manson D, Babyn P, Shuckett B. Reliability of the chest radiograph in the diagnosis of lower respiratory infections in young children. Pediatr Infect Dis J 1996; 15 : 600-604.
7. World Health Organization Pneumonia Vaccine Trial Investigators Group. Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children. Geneva :WHO, 2001. WHO/V&B/0.1.35.
8. Chisaka N, Hassan M, Hibberd P, Patel A, Qazi S, Thea DM, *et al.* Oral amoxicillin versus injectable penicillin for severe pneumonia in children aged 3-59 months: a randomized multicenter equivalence study. Amoxicillin Penicillin Pneumonia International Study. Lancet 2004; 364: 1141-1148.
9. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measurement 1960; 20: 37-46.
10. Fleiss JL. Statistical methods for rates and proportions. 2nd edn. New York, John Wiley & Sons, 1981; 212-225.
11. Klugman KP, Madhi SA, Heubner RE, Kohberger R, Mbelle N, Pierce N. A trial of a 9-valent pneumococcal conjugate vaccine in children with and those without HIV infection. N Engl J Med 2003; 349: 1341-1348.
12. Madhi SA, Kuwanda L, Cutland C, Klugman K. The impact of 9-valent pneumococcal conjugate vaccine on the public health burden of pneumonia in HIV-infected and uninfected children. Clin Infect Dis 2005; 40: 1511-1519.
13. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the

estimation of relative risk. Am J Epidemiol 1977; 105: 488-495.

14. Bloomfield FH, Teele RL, Voss M, Knight DB, Harding JE. Inter- and intra-observer variability in the assessment of atelectasis and consolidation in neonatal chest radiographs. Pediatr Radiol 1999; 29: 459-462.

15. Rubenfeld GD. Interobserver variability in applying a radiographic definition for ARDS. Chest 1999; 116: 1347-1353.

16. Coakley FV, Green J, Lamont AC, Rickett AB. An investigation into perihilar inflammatory change on the chest radiographs of children admitted with acute respiratory symptoms. Clin Radiol 1996; 511: 614-617.

17. Cherian T, Mulholland E K, Carlin JB, Ostensen H, Amin R, Campo M *et al.* Standardized interpretation of pediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. Bull WHO 2005; 83: 353-359.