Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies

Thomas Cherian,[1] E. Kim Mulholland,[2] John B. Carlin,[3] Harald Ostensen,[1] Ruhul Amin,[4] Margaret de Campo,[5] David Greenberg,[6] Rosanna Lagos,[7] Marilla Lucero,[8] Shabir A. Madhi,[9] Katherine L. O'Brien,[10] Steven Obaro,[11] Mark C. Steinhoff,[12] & the WHO Radiology Working Group

Background Although radiological pneumonia is used as an outcome measure in epidemiological studies, there is considerable variability in the interpretation of chest radiographs. A standardized method for identifying radiological pneumonia would facilitate comparison of the results of vaccine trials and epidemiological studies of pneumonia.

Methods A WHO working group developed definitions for radiological pneumonia. Inter-observer variability in categorizing a set of 222 chest radiographic images was measured by comparing the readings made by 20 radiologists and clinicians with a reference reading. Intra-observer variability was measured by comparing the initial readings of a randomly chosen subset of 100 radiographs with repeat readings made 8–30 days later.

Findings Of the 222 images, 208 were considered interpretable. The reference reading categorized 43% of these images as showing alveolar consolidation or pleural effusion (primary end-point pneumonia); the proportion thus categorized by each of the 20 readers ranged from 8% to 61%. Using the reference reading as the gold standard, 14 of the 20 readers had sensitivity and specificity of $\geqslant 0.70$ in identifying primary end-point pneumonia; 13 out of 20 readers had a kappa index of > 0.6 compared with the reference reading. For the 92 radiographs deemed to be interpretable among the 100 images used for intra-observer variability, 19 out of 20 readers had a kappa index of > 0.6.

Conclusion Using standardized definitions and training, it is possible to achieve agreement in identifying radiological pneumonia, thus facilitating the comparison of results of epidemiological studies that use radiological pneumonia as an outcome.

Keywords Pneumonia/radiography; Radiography, Thoracic/standards; Sensitivity and specificity; Observer variation; Reference standards; Child (*source: MeSH, NLM*).

Mots clés Pneumonie/radiographie; Radiographie thoracique/normes; Sensibilité et spécificité (Epidémiologie); Variation liée à l'observateur; Norme; Enfant (*source: MeSH, INSERM*).

Palabras clave Neumonía/radiografía; Radiografía torácica/normas; Sensibilidad y especificidad; Variaciones dependientes del observador; Estándares de referencia; Niño (*fuente: DeCS, BIREME*).

الكلمات المفتاحية: الالتهاب الرئوي، الدراسة الشعاعية للالتهاب الرئوي، الدراسة الشعاعية، الدراسة الشعاعية للصدر، معايير الدراسة الشعاعية للصدر، الحساسية والنوعية، التفاوت بين المراقبين، معايير مرجعية (المصدر: رؤوس الموضوعات الطبية، المكتب الإقليمي لشرق المتوسط).

Bulletin of the World Health Organization 2005;83:353-359.

Voir page 358 le résumé en français. En la página 358 figura un resumen en español. يمكن الاطلاع على الملخص بالعربية في صفحة 359.

Introduction

Acute lower respiratory tract infection, primarily pneumonia, is the leading cause of death in childhood in developing countries, resulting in an estimated 1.9 million deaths annually (*1*). However, studies to determine the true burden of pneumonia and the proportion that is preventable by vaccination have been hampered by the lack of an adequate definition of pneumonia. While radiological findings are commonly accepted as the "gold standard" for defining pneumonia, there are no validated definitions for X-ray interpretation. Moreover, inter- and intra-

[1] Initiative for Vaccine Research, Department of Immunization, Vaccines and Biologicals, World Health Organization, 1211 Geneva 27, Switzerland (email: cheriant@who.int). Correspondence should be sent to Dr Cherian.

[2] Centre for International Child Health, University of Melbourne, Melbourne, Australia.

[3] Clinical Epidemiology & Biostatistics Unit, Murdoch Children's Research Institute, and Department of Paediatrics, University of Melbourne, Melbourne, Australia.

[4] Bangladesh Institute of Child Health, Dhaka Shishu Hospital, Dhaka, Bangladesh.

[5] Royal Children's Hospital, Melbourne, Australia (current affiliation: Monath Medical Centre, Southern Health, Melbourne, Australia).

[6] Soroka Medical Centre, Beer Sheva, Israel.

[7] Hospital de Ninos Roberto del Rio, Santiago, Chile.

[8] Research Institute for Tropical Medicine, Manila, Philippines.

[9] Respiratory and Meningeal Pathogens Research Unit, Johannesburg, South Africa.

[10] Center for American Indian Health, The Johns Hopkins Bloomberg School of Public Health, Baltimore, USA.

[11] Medical Research Council, The Gambia (current affiliation: Department of Pediatric Infectious Diseases, Brown Medical School, Providence, USA).

[12] Department of International Health, The Johns Hopkins Bloomberg School of Public Health, Baltimore, USA.

Ref. No. **04-014894**

(*Submitted: 21 May 2004 – Final revised version received: 12 October 2004 – Accepted: 12 October 2004*)

observer variability in the interpretation of chest radiographs is a well recognized problem (*2*) and has been studied in the diagnosis of tuberculosis (*3*, *4*), pneumoconiosis (*5*), lung cancer (*6*) and adult pneumonia (*7*, *8*). For childhood pneumonia, apart from a few studies (*9*–*12*), this problem has not been adequately addressed. More specifically, there are no studies reporting multi-observer reviews of paediatric radiographs from developing countries where the quality of X-rays may be less adequate than that in developed countries.

With the availability of effective vaccines against the two leading bacterial pathogens causing childhood pneumonia in developing countries, namely *Haemophilus influenzae* type b (Hib) and pneumococcal conjugate vaccines, the need for standardized methods to collect data on the pneumonia disease burden and the proportion of the burden preventable by vaccines has become critical.

WHO's Department of Immunization, Vaccines and Biologicals, established a working group to standardize the categorization of radiological pneumonia, for the purpose of establishing burden estimates of likely bacterial pneumonia and estimating vaccine impact (*13*). This paper describes the process undertaken by the working group to achieve this objective and reports the results of a study of inter- and intra-observer variability in interpreting chest radiographs using the standardized methods. This process was meant only to standardize the interpretation of chest radiographs and not to address the question whether certain radiological patterns represent biologically or pathologically defined pneumonia.

Methods

The process of standardization consisted of three stages. These included (1) development and modification of nomenclature for visual descriptors of the characteristics of chest radiographic image; (2) learning and calibration of radiographic image interpretation and refinement of the definitions; and (3) formal measurement of inter- and intra-observer variability in interpretation.

Study participants

The participants in this process were from nine study sites that were proposing to evaluate the impact of Hib or pneumococcal conjugate vaccine. Of these, seven sites participated in the calibration phase, whereas all nine participated in the exercise to measure inter- and intra-observer variation. At each study site there were at least two X-ray readers (one radiologist and one clinician); two sites had three readers each and two sites had one common reader, who was a radiologist.

End-points and definitions

The definitions and end-points were developed through a series of workshops that included review of a large number of chest radiographic images, and were further refined at a calibration workshop. In formulating the definitions and end-points, the group took several factors into consideration. Previous studies had shown that although there is reasonable agreement on the presence of alveolar consolidation, there is considerable disagreement on other findings (*8*, *11*, *14*). Also, the presence of significant alveolar consolidation is considered by most authorities to be the most specific radiographic predictor of bacterial pneumonia. Therefore, this was chosen as the primary end-point of interest. Of necessity, the definitions used were framed to be more specific (albeit less sensitive) for likely bacterial pneumonia than those used for clinical purposes; this was considered appropriate for epidemiological studies. To determine the quality of each image, its adequacy for allowing categorization, rather than its technical quality, was assessed. All non-pulmonary findings were ignored for this study. The definitions thus formulated are shown in Table 1.

Calibration

This phase was carried out to determine whether there was systematic variability in interpretation of chest radiographs between the participating sites, and to determine whether there was a common understanding of the visual representation of end-points and definitions. For this phase, a set of 172 digitized chest X-ray images (frontal views only), from children aged 2–60 months with clinical pneumonia, was sent to seven study sites and also to two radiologists who were to act as a WHO reference panel. Each study site was asked to provide a consensus reading from their two readers using the above definitions and end-points. The pattern of readings from this exercise was used to identify specific areas of disagreement between study sites; these were discussed and resolved at a workshop, and the definitions refined, where they were seen as the source of disagreement.

Measurement of inter- and intra-observer variation in interpretation

For measurement of inter-observer variation in interpretation, another set of 222 digitized chest X-ray images (frontal views) (set A) was distributed to each study site. These images were selected from among those obtained from children aged 2–60 months with a clinical diagnosis of pneumonia attending the Soroka University Medical Center, Israel, during 2000 and the Chris Hani Baragwanath Hospital, South Africa, during 1998–99. They were selected with the aim of having a sample in which approximately 20–25% would be categorized as primary end-point pneumonia, which was the expected proportion of primary end-point pneumonia cases among children with clinical pneumonia enrolled in the trials. However, a higher proportion of images were classified as end-point pneumonia after the images had been read by all the readers and a reference reading assigned (see section on Analysis).

X-ray readers at each study site were required to independently review the images and enter their readings in the standardized data entry program. They were given no clinical information and were unaware of the prevalence of images categorized as showing end-point pneumonia in the set. For measurement of intra-observer variability, 100 images were selected at random from the set of 222, re-coded and compiled into a separate set (set B). The X-ray readers were required to report on set B 8–30 days after reading set A.

In order to minimize differences in interpretation that could arise owing to the quality of the image as viewed on individual computer monitors, specifications were provided for the hardware and graphics software to be used. In addition, a grayscale test pattern was provided to allow readers to optimize their monitor settings (*13*).

Sample size was based on considerations of precision to be expected in estimates of sensitivity and specificity (compared with the reference reading) and of the kappa index of agreement.

Table 1. **Definitions of radiological findings and end-points of pneumonia**

	Finding	Definition
Film quality	Uninterpretable	Features of the image are not interpretable with respect to presence or absence of "primary end-point" without additional images
	Suboptimal	Features allow interpretation of primary end-point, but not of other infiltrates or findings; no entries were made for "other infiltrates" for such images
	Adequate	Features allow confident interpretation of end-point as well as other infiltrates
Classification of findings	Significant pathology	Refers specifically to the presence of consolidation, infiltrates or effusion
	End-point consolidation[a]	A dense or fluffy opacity that occupies a portion or whole of a lobe or of the entire lung, that may or may not contain air-bronchograms[b]
	Other (non-end-point) infiltrate	Linear and patchy densities (interstitial infiltrate) in a lacy pattern involving both lungs, featuring peribronchial thickening and multiple areas of atelectasis; 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 may be difficult to distinguish from consolidation
	Pleural effusion	Presence of fluid in the lateral pleural space between the lung and chest wall; in most cases, this will be seen at the costo-phrenic angle or as a layer of fluid adjacent to the lateral chest wall; this does not include fluid seen in the horizontal or oblique fissures
Conclusions	Primary end-point pneumonia	The presence of end-point consolidation (as defined above) or pleural effusion that is in the lateral pleural space (and not just in the minor or oblique fissure) and was spatially associated with a pulmonary parenchymal infiltrate (including other infiltrate) OR if the effusion obliterated enough of the hemithorax to obscure an opacity
	Other infiltrate	The presence of other (non-end-point) infiltrate as defined above in the absence of a pleural effusion
	No consolidation/infiltrate/effusion	Absence of end-point consolidation, other infiltrate or pleural effusion

[a] The choice of the term "end-point" refers to this being the end-point of interest for trials of bacterial vaccines against pneumonia.
[b] Atelectasis of an entire lobe that produces a dense opacity and a positive silhouette sign with the mediastinal border was considered to be an end-point consolidation.

Analysis

Each of the 222 images was assigned a reference reading based on the majority reading of the 20 participating readers, with adjudication of disputed cases by consensus among two radiologists and two paediatricians (TC, KM, MdC and HO); the two radiologists (MdC and HO) also contributed to the independent readings that were evaluated in this exercise, but were blinded to their original reading when assigning the reference reading. In assigning the reference reading, for 186 cases where there was a clear majority (more than two-thirds of readers agreed on the presence or absence of primary end-point pneumonia), the majority reading of the 20 readers was accepted as the reference reading in all cases except three, two of which were considered uninterpretable by the panel although a majority of readers had provided a reading for them. Similarly, for 33 images where there was only a narrow majority (more than half but fewer than two-thirds of the readers agreed on the presence or absence of end-point pneumonia), the majority reading was accepted as the reference reading, except for six images that were classified as uninterpretable and five for which the panel assigned a reading different from that of the majority. The panel also assigned readings for the two images for which readers were equally divided in opinion.

To assess inter-observer variability, the reading of each participating reader was compared with the reference reading. For intra-observer variability, the conclusion of each reader for the images in set A was compared with the conclusion for the corresponding image in set B. Given that the reference reading has high face validity as the gold standard for these images, our analysis focused on presenting descriptive summaries of the correspondence between the reading provided by each participating reader and the reference reading in terms of sensitivity and specificity. We also present measures of agreement between each reader and the reference reading in the form of overall per cent of cases in which the reader agreed with the gold standard, difference in overall per cent positive between individual reader and reference reading, and the "chance-adjusted" kappa index. The kappa calculation finds the proportion of cases on which readers would be expected to agree, given that they each diagnose end-point pneumonia in a given percentage of cases, and reports the actual agreement among the remaining proportion of cases. Although not strictly appropriate for comparison with a gold standard, it is a useful general index of agreement. Data manipulation and kappa calculations were performed using the statistical package Stata 7.0 (Stata Corporation, 2001).

Results

Twenty X-ray readers participated in this exercise. One image in the collection could not be opened with the image viewing program by a majority of the readers and was excluded from the analysis. Also excluded were 13 other images that the reference reading indicated were uninterpretable. The results of the analysis on the remaining 208 images are summarized in Table 2 (web version only, available at: http://www.who.int/bulletin), Fig.1 and Fig.2. Also summarized in Table 2 is the analysis of intra-observer agreement for 92 of the 100 images used that the reference reading indicated were interpretable. The reference reading

concluded that 43% of these X-rays indicated the presence of primary end-point pneumonia. The proportion that the individual readers categorized as having end-point pneumonia ranged from 8% to 61%; the graph of sensitivity and specificity in Fig. 1 indicates that there was substantial variation in the threshold for positivity between the 20 readers. Of the 20 X-ray readers, 11 achieved sensitivity and specificity of ⩾ 0.75 compared with the reference reading; 14 achieved sensitivity and specificity of ⩾ 0.70. The median sensitivity and specificity for the clinicians was 0.84 and 0.89, respectively, and the median sensitivity and specificity of the radiologists was 0.87 and 0.87, respectively. Thirteen readers had kappa values of > 0.6 compared with the reference reading. The median kappa indices for the clinicians and radiologists were 0.65 and 0.73, respectively. Nineteen of the twenty readers had a kappa index of > 0.6 for repeatability.

The analyses for agreement in identifying end-point consolidation on the right and left side, respectively, are summarized in Table 3 (web version only, available at: http://www.who.int/bulletin). In general, there was a higher kappa index for the right than the left side.

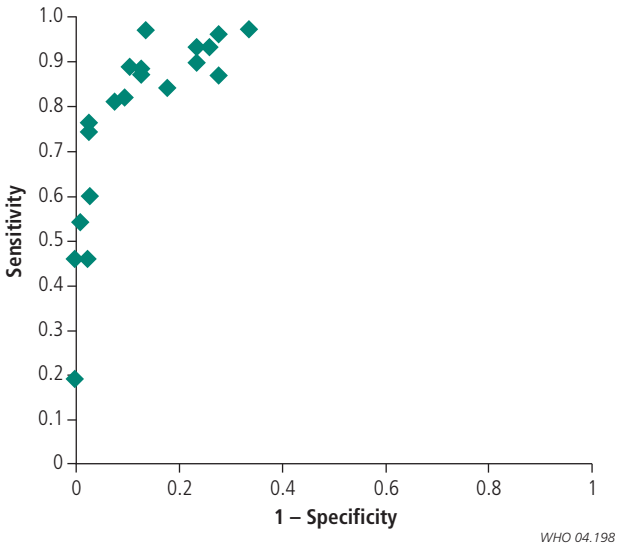
We also analysed inter-observer agreement in identifying any infiltrate on a chest radiograph, i.e. either end-point consolidation or other infiltrates. The results are summarized in Table 4 (web version only, available at: http://www.who.int/bulletin). In general, the agreement for this broader category was lower than for end-point pneumonia alone.

Discussion

Conventionally, interpretation of chest radiographs is carried out for the purposes of direct patient care; the diagnosis is seldom based on a single observation but rather on the integration of a number of related observations, and is sometimes revised on the basis of subsequent observations. In these situations, subtle changes in the radiographic findings may provide important clues towards establishing a diagnosis. For pneumonia, the bias of clinicians is towards greater sensitivity as the clinical consequences of failing to treat a possible bacterial pneumonia may be serious. On the other hand, diagnosis of pneumonia in epidemiological studies is often based on the interpretation of a single radiograph and may be uncoupled from other clinical findings. Moreover, for epidemiological studies, greater specificity is often desirable.

WHO has previously attempted to standardize the interpretation of chest radiographs for epidemiological studies. In one exercise, four paediatric radiologists met and agreed on the definitions and the reporting form to be used. The reporting form included a much more detailed description of the radiographic findings than in this exercise. The radiologist then independently read chest radiographs from a number of epidemiological studies of pneumonia. Analysis of these readings showed that while there was reasonable agreement for alveolar consolidation, agreement was low for many other findings (M. Weber, personal communication). Since many of the recorded variables were seldom used for categorization of X-rays or subsequent analysis, but added to the degree of disagreement, our aim was to develop a simplified system that would allow categorization of radiographic findings but limit variability in interpretation. Care was taken to avoid ambiguity or overlap when framing definitions and several learning and discussion sessions were used to increase consensus.

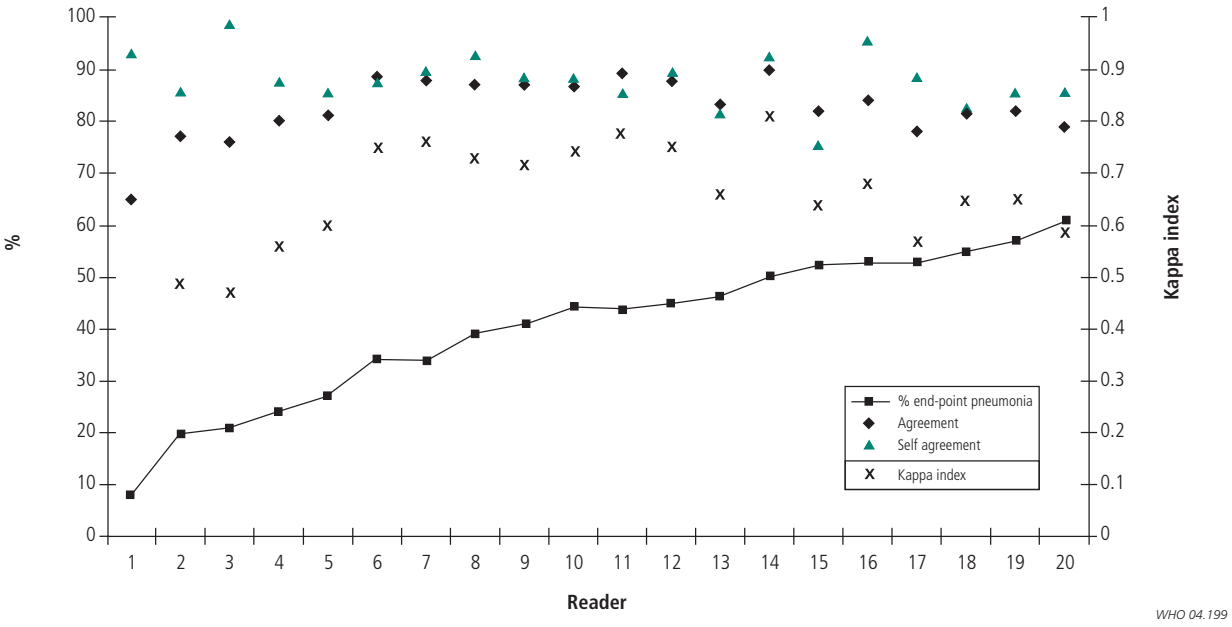
Fig. 1. **Distribution of sensitivity and specificity of the participating X-ray readers compared with the reference reading**



The data from this study suggest that with the use of simple criteria and adequate training, a reasonably high degree of agreement may be achieved in categorizing radiological pneumonia. Several other studies have evaluated agreement in interpreting chest radiographs in patients with suspected pneumonia (*8*, *11*, *15*). The results of our study compare favourably with these results, especially considering that the readers in this study were from several different countries and institutions, with varying backgrounds and, therefore, more likely to have variable interpretation. It is noted, however, that the definition of the gold standard, which is based on the majority reading of the individual readers (i.e. the reference reading), is likely to increase the level of agreement for each reader, though the increase is difficult to quantify. It has been shown that agreement may be further improved by the use of a process that required independent reading by two readers, with any discordance resolved by a third reader (*4*). Hence, the recommended plan for reading chest radiographs for epidemiological studies of pneumonia is an independent reading by a radiologist and a clinician with adjudication of discordant conclusions by a panel of two radiologists, whose consensus reading is taken as the final reading (*13*).

In general, there was high intra-observer agreement (Table 2, web version only, available at: http://www.who.int/bulletin). However, some of those readers with high intra-observer agreement had low agreement with the reference reading and vice versa. We also examined the data to see whether there was systematic variation in interpretation of radiographs that was attributable to different sites or qualification of the readers (i.e. clinician versus radiologist). At two sites, both the readers had low sensitivity but high specificity compared with the reference reading, whereas at another site all three readers had high sensitivity but low specificity (data not shown). At the start of this exercise, we had expected to find that the clinicians would call fewer images positive for primary end-point pneumonia, compared with radiologists. However, we did not find this to be the case, suggesting that once adequately trained, there may not be a difference between these two groups of observers.

Fig. 2. **Distribution of the % of images that were positive for end-point pneumonia, % agreement with reference reading, % self-agreement and Kappa index for the participating readers**



The finding that some readers had a low level of agreement with the reference reading and that some sites systematically under- or over-called primary end-point pneumonia underlines the need for further training of some readers before starting a study. Furthermore, ongoing re-calibration during the course of the study is required to assure acceptable levels of agreement. In current vaccine trials, this is being done with the help of specially-designed self-training and assessment software that contains a repository of images with an assigned reference reading. In addition, an independent blind reading by the WHO radiology panel of a sample of radiographic images for which the readings of the two site readers were concordant is conducted as a form of quality control.

The routine use of lateral views was considered by the working group and has been the subject of much debate. Existing data show that lateral views contribute to increased detection of pneumonia in only a small proportion of cases (*16–19*). Given the low yield and the additional radiation exposure, especially in some developing countries where radiation protection measures may not be optimal, it was felt that requiring lateral views only for trial purposes may not be justified.

The analysis in this study was restricted to images that were considered interpretable. However, we recognize that image quality, with respect to the original as well as the digitized images, may contribute to variability in interpretation. To limit such variability, guidelines for performing and digitizing radiographs have been prepared and distributed (*13*). In addition, site visits to review and refine radiographic procedures have been conducted at several trial sites, especially those in developing countries, and feedback on image quality is provided by the WHO radiology panel on the basis of the images submitted for adjudication.

This study was not designed to determine how predictive the definitions were of bacterial pneumonia. Preliminary unpublished data from this group suggests that cases classified as primary end-point pneumonia using the process described are enriched for bacterial pneumonia, i.e. have other manifestations of bacterial infection, such as high fever, granulocytosis, elevated erythrocyte sedimentation rate and serum C-reactive protein. The recently published results of the pneumococcal vaccine trial in South Africa showed a significant reduction in end-point pneumonia, as defined here, but not of infections of the lower respiratory tract with other radiological changes, in vaccinated children (*20*). Results from other ongoing and recently completed studies will need to be carefully analysed to determine the true value of this process.

The definitions and methods described in this paper are being used in a number of trials evaluating the effect of Hib and pneumococcal conjugate vaccines on pneumonia. It is anticipated that the same definitions and methods will also be used in other epidemiological studies of pneumonia, and that in the coming few years several studies of disease burden will be undertaken to estimate the potential impact of pneumococcal vaccination. Such studies will only be able to express the vaccine preventable burden of pneumonia with reference to the standardized definition of pneumonia that is used in the trials. Thus the definition described in this paper will form the link between disease burden studies and vaccine trials. The use of this standard method would increase the probability that any difference in the results reflect true geographic differences in disease epidemiology or vaccine performances rather than being the effect of methodological differences. ■

Acknowledgements

Names of individual Members of the WHO Vaccine Trial Investigators' Radiology Working Group are available on the web version at: http://www.who.int/bulletin.

The contributions of the following are also gratefully acknowledged: Teresa Aguado, WHO, Geneva, Switzerland;

Shams El Arifeen, Bangladesh; Steve Black, USA; Jonathan Carapetis, Australia; Gabrielle Davie, Australia; Penny Enarson, France; Godwin Enwere, The Gambia; Drora Fraser, Israel; Per-Egil Hansen, United Kingdom; Robin Huebner, South Africa; Gurauv Kumar, USA; Amanda Leach, The Gambia; Orin Levine, USA; Pirjo Helena Makela, Finland; Nontombi Mbelle, South Africa; Taneli Puumalainen, Finland; Shamim Qazi, WHO, Geneva, Switzerland; Ian Riley, Australia; Petri Ruutu, Finland; Henry R. Shinefield, USA; Montse Soriano-Gabarro, USA; and Martin Weber, WHO, Geneva, Switzerland. The contributions of the United States Agency for International Development (USAID), which has supported this process throughout its course, the Children Vaccine Programme at PATH (Program for Appropriate Technology in Health) and the International Union Against Tuberculosis and Lung Disease are also gratefully acknowledged.

Competing interests: Marilla Lucero is the principal investigator for a trial for which vaccines were supplied by Aventis Pasteur. Stephen Obaro has participated in studies for which vaccine was provided by Wyeth Lederle. Margaret de Campo is a consultant radiologist for a trial for which vaccine was supplied by Aventis Pasteur and is a member of the GlaxoSmithKline Biologicals evaluation committee. Rosanna Lagos has received financial support from manufacturers of pneumococcal vaccines to conduct surveillance projects of pneumonia. Mark Steinhoff has received honoraria and research grants from numerous vaccine manufacturers, but none connected with the subject of this study. Kim Mulholland is a co-investigator of a study in which pneumonia is an important end-point; pneumococcal polysaccharide vaccine for this trial was donated by GlaxoSmithKline Biologicals and has participated in other clinical trials that were partially supported by vaccine manufacturers. Katherine O'Brien has received research grants from Wyeth Vaccines, Merck, Aventis Pasteur, MedImmune, GlaxoSmithKline Biologicals and Chiron Vaccines and is a participant in the Advisory Board of Wyeth Vaccines and Aventis Pasteur. Shabir Madhi has received research grant support from Wyeth Vaccines.

None of the above were related to the subject of this study and the authors do not feel that they constitute a conflict of interest.

Résumé

Interprétation standardisée des radiographies pulmonaires pédiatriques servant au diagnostic de la pneumonie dans les études épidémiologiques

Objectif Bien que la pneumonie radiologique soit utilisée comme critère de jugement dans les études épidémiologiques, on relève une variabilité considérable dans l'interprétation des radiographies pulmonaires. Une méthode standardisée pour l'identification des pneumonies radiologiques faciliterait la comparaison des résultats des essais vaccinaux et des études épidémiologiques concernant la pneumonie.

Méthodes Un groupe de travail de l'OMS a mis au point des définitions de la pneumonie radiologique. Il a évalué la variabilité interobservateur dans le classement d'une série de 222 clichés radiographiques des poumons, en comparant les lectures faites par 20 radiologues et cliniciens à une lecture de référence. Il a mesuré cette variabilité par comparaison des premières lectures d'une sous-série de 100 radiographies choisies de manière aléatoire avec les nouvelles lectures réalisées 8 à 30 jours plus tard.

Résultats Parmi les 222 clichés, 208 ont été considérés comme interprétables. La lecture de référence a classé 43 % de ces radiographies comme présentant une consolidation alvéolaire ou un épanchement pleural (critère d'évaluation primaire : pneumonie), tandis que la proportion de clichés classés dans cette catégorie par les 20 lecteurs allait de 8 à 61 %. Si l'on utilise la lecture de référence comme étalon, 14 des 20 lecteurs ont fait preuve d'une sensibilité et d'une spécificité $\geqslant$ 0,70 dans l'identification du critère d'évaluation primaire, à savoir la pneumonie. Treize des 20 lecteurs obtenaient un coefficient kappa > 0,6 par rapport à la lecture de référence. Concernant la lecture des 92 radiographies jugées interprétables parmi les 100 clichés utilisés pour évaluer la variabilité interobservateur, on a déterminé un coefficient kappa > 0,6 pour 19 des 20 lecteurs.

Conclusion En recourant à des définitions et à une formation standardisées, il est possible de parvenir à un accord dans l'identification de la pneumonie radiologique, ce qui facilite la comparaison des résultats des études épidémiologiques utilisant la pneumonie radiologique comme critère de jugement.

Resumen

Interpretación normalizada de radiografías torácicas pediátricas para el diagnóstico de neumonía en estudios epidemiológicos

Objetivo Aunque la neumonía radiológica se usa en estudios epidemiológicos como medida de resultado, se observa una considerable variabilidad en la interpretación de las radiografías torácicas. Un método estandarizado de identificación de la neumonía radiológica facilitaría la comparación de los resultados de los ensayos de vacunas y los estudios epidemiológicos sobre la neumonía.

Métodos Un grupo de trabajo de la OMS elaboró definiciones de neumonía radiológica. Para medir la variabilidad interobservadores en la tarea de clasificar un conjunto de 222 imágenes de radiografías torácicas, se procedió a comparar las lecturas realizadas por 20 radiólogos y clínicos frente a una lectura de referencia. La variabilidad intraobservador se midió comparando las lecturas iniciales de un subconjunto de 100 radiografías seleccionadas al azar con las repeticiones de esas lecturas al cabo de entre 8 y 30 días.

Resultados De las 222 imágenes, 208 se consideraron interpretables. La lectura de referencia clasificó el 43% de las imágenes como indicativas de consolidación alveolar o derrame pleural (criterio principal de valoración de la existencia de neumonía); la proporción así clasificada por los 20 lectores de las imágenes osciló entre el 8% y el 61%. Comparando con la lectura de referencia, 14 de los 20 lectores presentaron una sensibilidad y especificidad $\geqslant$ 0,70 en la identificación de los criterios

principales de neumonía; 13 de los 20 lectores presentaron un índice kappa > 0,6 en comparación con la lectura de referencia. Para las 92 radiografías consideradas interpretables entre las 100 imágenes usadas para determinar la variabilidad intraobservador, 19 de los 20 lectores presentaron un índice kappa > 0,6.

Conclusión Usando definiciones y adiestramiento normalizados, es posible acordar un procedimiento para identificar la neumonía radiológica, y facilitar así la comparación de los estudios epidemiológicos que usan ese signo como resultado.

ملخص

تشخيص الالتهاب الرئوي شعاعياً

التفسير المعياري لصور الصدر الشعاعية لدى الأطفال لتشخيص الالتهاب الرئوي في الدراسات الوبائية

معلومات أساسية: رغم أن تشخيص الالتهاب الرئوي شعاعياً يستخدم في الدراسات الوبائية كمقياس للحصائل، إلا أن هناك تفاوتا ملحوظا في تفسير صور الصدر الشعاعية. وسيؤدي توافر طريقة معيارية لتعريف الالتهاب الرئوي شعاعياً إلى تسهيل المقارنة بين نتائج تجارب التطعيم والدراسات الوبائية للالتهاب الرئوي.

الطريقة: أَعَدَّ فريق عمل في منظمة الصحة العالمية تعاريف لتشخيص الالتهاب الرئوي شعاعياً، ثم أجري قياس للتفاوت بين الباحثين في تصنيف 222 من صور الصدر الشعاعية ضمن فئات تشخيصية بمقارنة القراءات التي أَعَدَّها عشرون من الاختصاصيين بالأشعة ومن الأطباء السريريين (الإكلينيكيين) مع قراءات مرجعية . كما تم قياس التفاوت ضمن المراقبين بمقارنة القراءات البدئية لمجموعة فرعية تم اختيارها عشوائيا وتتألف من مئة صورة شعاعية للصدر مع قراءات متكررة أجريت بعد مرور 8 – 30 يوماً.

الموجودات: اعتبرت 208 صور من بين مجموع الصور البالغ عددها 222 صورة أنها قابلة للتفسير . ووفقا للقراءات المرجعية، صنفت 43% من هذه الصور على أنها تبدي كثافات حويصلية أو انصبابات جنبية، مما يشير إلى الالتهاب الرئوي بشكل نهائي، وهكذا فإن النسبة المئوية للتصنيف الذي قام به كل واحد من القارئين العشرين ضمن إحدى الفئات قد تراوحت بين 8%و61%. وإذا استخدمت القراءة المرجعية كمعيار ذهبي فإن 14 من القارئين العشرين كانوا يتمتعون بحساسية ونوعية في كشف الالتهاب الرئوي البدئي والقطعي الثبوت تزيد على 0.70 ، فيما كان لدى 13 من القارئين منسب كابا يزيد على 0.6 مقارنة بالقراءة المرجعية. ومن بين الصور الشعاعية المئة اعتبر أن 92 منها قابلة للتفسير، واستخدمت لمعرفة التفاوت بين المراقبين، وكان لدى 19 من القارئين منسب كابا يزيد على 0.6 .

الاستنتاج: من الممكن الوصول إلى اتفاق لتشخيص الالتهاب الرئوي شعاعياً، وذلك باستخدام تعاريف معيارية مع التدريب على استخدامها، مما يسهل المقارنة مع نتائج الدراسات الوبائية التي تستخدم تشخيص الالتهاب الرئوي شعاعياً باعتباره من الحصائل المتمخضة عنها.

References

1. Williams BG, Gouws E, Boschi-Pinto C, Bryce J, Dye C. Estimates of world-wide distribution of child deaths from acute respiratory infections. *Lancet Infectious Diseases* 2002;2:25-32.
2. Koran LM. The reliability of clinical methods, data and judgments. *New England Journal of Medicine*, 1975;293:642-6.
3. Garland LH. Studies on the accuracy of diagnostic procedures. *American Journal of Roentgenology* 1959; 82:25-38.
4. Yerushalmy J. The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. *Radiologic Clinics of North America* 1969;3:381-92.
5. Felson B, Morgan WKC, Bristol LJ. Observations on the results of multiple readings of chest films in coal miners pneumoconiosis. *Radiology* 1973;109:19-23.
6. Kundel HL. Perception errors in chest radiograpy. *Seminars in Respiratory Medicine* 1989;10:203-10.
7. Melbye H, Dale K. Interobserver variability in the radiographic diagnosis of adult outpatient pneumonia. *Acta Radiologica* 1992;33:79-81.
8. Albaum MN, Hill LC, Murphy M, Li YH, Fuhrman CR, Britton CA, et al. Interobserver reliability of the chest radiograph in community-acquired pneumonia. PORT Investigators. *Chest* 1996;110:343-50.
9. Coblentz CL. Observer variation in detecting the radiologic features associated with bronchiolitis. *Investigative Radiology* 1991:115-8.
10. Crain EF, Bulas D, Bijur PE, Goldman HS. Is a chest radiograph necessary in the evalution of every febrile infant less than 8 weeks of age? *Pediatrics* 2001;88:821-4.
11. 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. *The Pediatric Infectious Disease Journal* 1996;15:600-4.
12. Stickler GB, Hoffman AD, Taylor WF. Problems in the clinical and roentgenographic diagnosis of pneumonia in young children. *Clinical Pediatrics* 1984;23:398-9.
13. 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 document WHO/V&B/01.35.
14. Young M, Marrie TJ. Interobserver variability in the interpretation of chest roentgenograms of patients with possible pneumonia. [comment]. *Archives of Internal Medicine* 1994;154:2729-32.
15. 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. *Pediatric Radiology* 1999;29:459-62.
16. Ely JW, Berbaum KS, Bergus GR, Thompson BH, Levy BT, Graber MA, et al. Diagnosing left lower lobe pneumonia: usefulness of the 'spine sign' on lateral chest radiographs. *The Journal of Family Practice* 1996;43:242-8.
17. Kennedy J, Dawson KP, Abbott GD. Should a lateral chest radiograph be routine in suspected pneumonia? *Australian Paediatric Journal* 1986;22:299-300.
18. Kiekara O, Korppi M, Tanska S, Soimakallio S. Radiological diagnosis of pneumonia in children. *Annals of Medicine* 1996;28:69-72.
19. Lamme T, Nijhout M, Cadman D, Milner R, Zylak C, Jacobs J, et al. Value of the lateral radiologic view of the chest in children with acute pulmonary illness. *Canadian Medical Association Journal* 1986;134:353-6.
20. Klugman KP, Madhi SA, Huebner RE, Kohberger R, Mbelle N, Pierce N, et al. A trial of a 9-valent pneumococcal conjugate vaccine in children with and those without HIV infection. *New England Journal of Medicine* 2003;349:1341-8.

Table 2. **Summary of agreement of individual readers with the reference reading for end-point pneumonia for 208/222 readable images (set A) and between the first and the second reading by these readers for 92/100 readable images (set B) used to measure repeatability**

Reader	Frequency (%)			Reader versus reference reading					Repeatability			
	Unreadable film	Suboptimal films	End-point pneumonia	Agreement (%)	Difference, reader–RR[a] (%)	Kappa[b]	Sensitivity	Specificity	J index[c]	Agreement (%)	Difference, B–A[d] (%)	Kappa
1	1	12	8	65	–35	0.21	0.19	1	0.19	93	0	0.47
2	0	3	20	77	–23	0.49	0.46	1	0.46	86	12	0.65
3	0	14	21	76	–22	0.47	0.46	0.98	0.44	99	–1	0.97
4	1	35	24	80	–19	0.56	0.54	0.99	0.53	88	10	0.71
5	8	22	27	81	–16	0.6	0.6	0.97	0.57	86	1	0.66
6	1	20	34	88	–9	0.75	0.75	0.97	0.72	88	–5	0.72
7	1	20	34	88	–9	0.76	0.76	0.97	0.73	90	8	0.79
8	0	16	39	87	–3	0.73	0.81	0.92	0.73	93	–2	0.87
9	0	12	41	87	–2	0.72	0.82	0.9	0.72	89	7	0.78
10	2	10	44	87	1	0.74	0.87	0.87	0.74	89	–4	0.78
11	1	39	44	89	1	0.78	0.89	0.89	0.78	86	8	0.72
12	0	6	45	88	2	0.75	0.88	0.87	0.75	90	–3	0.8
13	3	6	46	83	3	0.66	0.84	0.82	0.66	82	-8	0.63
14	0	28	50	90	7	0.81	0.97	0.86	0.83	93	7	0.87
15	5	30	52	82	10	0.64	0.9	0.76	0.66	76	–13	0.53
16	3	42	53	84	11	0.68	0.93	0.76	0.69	96	–4	0.91
17	1	0	53	78	10	0.57	0.87	0.72	0.59	89	–2	0.78
18	5	15	55	82	12	0.65	0.93	0.74	0.67	83	–7	0.65
19	1	13	57	82	14	0.65	0.96	0.72	0.68	86	8	0.7
20	0	40	61	79	18	0.59	0.97	0.66	0.63	86	–8	0.71
Median	1	15.5	44	82.5	1	0.655	0.855	0.88	0.675	88.5	–1.5	0.72
Maximum	8	42	61	90	18	0.81	0.97	1	0.83	99	12	0.97
Minimum	0	0	8	65	–35	0.21	0.19	0.66	0.19	76	–13	0.47

[a] RR = reference reading.

[b] The Kappa calculation finds the proportion of cases on which readers would be expected to agree, given that they diagnose end-point pneumonia in a given percentage of cases each, and reports the actual agreement among the remaining proportion of cases. Kappa calculations were performed using the statistical package Stata 7.0 (Stata Corporation, 2001).

[c] J index = Youden's J index.

[d] % difference (B–A) = % difference between set B and set A.

Table 3. **Summary of agreement of individual readers with the reference reading for end-point consolidation on the right and left side, respectively, for 208/222 readable images**

Reader	Primary end-point consolidation, right lung						Primary end-point consolidation, left lung					
	Reader versus reference reading						Reader versus reference reading					
	Frequency positive (%)	Agreement (%)	Difference, reader–RR[a] (%)	Kappa[b]	Sensitivity	Specificity	Frequency positive (%)	Agreement (%)	Difference, reader–RR (%)	Kappa	Sensitivity	Specificity
RR[a]	35	–	–	–	–	–	11	–	–	–	–	–
1	6	71	–28	0.22	0.18	1	2	90	–9	0.2	0.13	0.99
2	15	80	–20	0.5	0.43	1	5	94	–6	0.58	0.43	1
3	14	79	–21	0.47	0.4	1	7	93	–4	0.56	0.48	0.98
4	19	84	–16	0.6	0.53	0.99	5	94	–6	0.58	0.43	1
5	22	87	–13	0.69	0.61	0.99	4	91	–7	0.41	0.3	0.99
6	25	88	–10	0.71	0.68	0.98	13	89	1	0.47	0.57	0.93
7	28	92	–6	0.82	0.78	0.98	9	93	–2	0.59	0.57	0.97
8	26	89	–9	0.75	0.72	0.99	16	89	5	0.55	0.74	0.91
9	29	91	–6	0.79	0.78	0.97	17	87	6	0.48	0.7	0.9
10	33	90	–2	0.77	0.82	0.93	19	84	8	0.4	0.65	0.87
11	33	90	–1	0.77	0.83	0.93	13	92	2	0.61	0.7	0.94
12	34	91	–1	0.81	0.86	0.93	19	89	8	0.57	0.87	0.89
13	35	88	1	0.73	0.81	0.89	16	92	5	0.65	0.87	0.92
14	37	93	2	0.85	0.93	0.93	24	85	13	0.5	0.91	0.84
15	45	86	10	0.71	0.93	0.81	15	93	4	0.69	0.83	0.93
16	41	89	7	0.77	0.93	0.86	24	85	13	0.51	0.91	0.85
17	39	83	4	0.65	0.82	0.84	24	81	13	0.37	0.74	0.82
18	43	85	8	0.69	0.9	0.82	30	75	19	0.29	0.74	0.76
19	48	84	13	0.67	0.96	0.78	27	82	16	0.46	0.96	0.81
20	40	89	6	0.77	0.93	0.88	27	82	16	0.44	0.91	0.81
Median	33	88	–2	0.71	0.81	0.93	15.5	89	5	0.51	0.7	0.92
Maximum	48	93	13	0.85	0.96	1	30	94	19	0.69	0.96	1
Minimum	6	71	–28	0.22	0.18	0.78	2	75	–9	0.2	0.13	0.76

[a] RR = reference reading.

[b] The Kappa calculation finds the proportion of cases on which readers would be expected to agree, given that they diagnose end-point pneumonia in a given percentage of cases each, and reports the actual agreement among the remaining proportion of cases. Kappa calculations were performed using the statistical package Stata 7.0 (Stata Corporation, 2001).

Table 4. **Summary of agreement of individual readers with the reference reading for any infiltrate, i.e. end point consolidation or other infiltrate in 208/222 readable images and between the first and the second reading of the readers for 92/100 readable images used to measure repeatability**

Reader	Frequency any infiltrate (%)	Reader versus reference reading					Repeatability		
		Agreement (%)	Difference, reader–RR (%)	Kappa[b]	Sensitivity	Specificity	Agreement (%)	Difference, B–A[c] (%)	Kappa
RR[a]	64	–	–	–	–	–	–	–	–
1	74	86	10	0.67	0.96	0.67	86	10	0.58
2	39	75	–25	0.53	0.61	1.00	89	9	0.78
3	50	76	–14	0.52	0.71	0.85	89	0	0.78
4	58	76	–6	0.51	0.77	0.76	78	20	0.53
5	38	71	–26	0.45	0.57	0.95	79	–3	0.55
6	74	77	10	0.48	0.9	0.55	76	–11	0.48
7	54	84	–10	0.66	0.8	0.91	89	4	0.78
8	79	81	15	0.55	0.97	0.53	93	–4	0.82
9	49	77	–15	0.54	0.7	0.89	87	7	0.74
10	55	83	–9	0.64	0.80	0.88	88	–1	0.76
11	56	84	–8	0.67	0.81	0.89	89	4	0.78
12	55	79	–9	0.56	0.77	0.83	87	2	0.74
13	78	80	14	0.53	0.95	0.53	88	–8	0.65
14	71	86	7	0.68	0.95	0.71	93	2	0.85
15	64	84	1	0.65	0.88	0.77	72	–13	0.43
16	76	84	13	0.62	0.97	0.60	96	–2	0.88
17	90	73	26	0.31	0.99	0.27	92	–3	0.59
18	71	85	7	0.65	0.93	0.69	84	–5	0.63
19	72	83	8	0.61	0.92	0.65	91	4	0.78
20	65	84	1	0.64	0.88	0.76	88	–5	0.75
Median	64.5	82	1	0.585	0.88	0.76	88	–0.5	0.745
Maximum	90	86	26	0.68	0.99	1.00	96	20	0.88
Minimum	38	71	–26	0.31	0.57	0.27	72	–13	0.43

[a] RR = reference reading.

[b] The Kappa calculation finds the proportion of cases on which readers would be expected to agree, given that they diagnose end-point pneumonia in a given percentage of cases each, and reports the actual agreement among the remaining proportion of cases. Kappa calculations were performed using the statistical package Stata 7.0 (Stata Corporation, 2001).

[c] % difference (B–A) = % difference between set B and set A.

Members of the WHO Vaccine Trial Investigators' Radiology Working Group

Aliu O. Akano, Abuja, Nigeria; Abdullah Hel Baqui, Dhaka, Bangladesh; Jacob Bar-Ziv, Jerusalem, Israel; Jane Benson, Baltimore, MD, USA; Ron Dagan, Beer Sheva, Israel; Bradford Gessner, Anchorage, AK, USA; Brian Greenwood, London, England; Zahid Hossain, Dhaka, Bangladesh; Keith Klugman, Atlanta, GA, USA; Socorro Lupisan, Manila, Philippines; Jack Marvis, Soweto, South Africa; Karla Moene, Santiago, Chile; Alma Munoz, Santiago, Chile; Awaatief Musson, Johannesburg, South Africa; Hanna Nohynek, Helsinki, Finland; Terry Nolan, Parkville, Victoria, Australia; Vicente V. Romano Jr, Manila, Philippines; Mathuram Santosham, Baltimore, MD, USA; Heinz Tschäppeler, Bern, Switzerland.