

Interobserver variation in respiratory signs of severe malaria

Michael English, Steven Murphy, Isaiah Mwangi, Jane Crawley, Norbert Peshu, Kevin Marsh

Abstract

Respiratory abnormalities are common presentations of malaria and acute respiratory tract infection, both of which are major causes of childhood mortality and morbidity in sub-Saharan Africa. Appropriate management depends on accurate assessment of disease severity which for the majority of children must be based on clinical signs alone. Choosing which signs best serve this purpose remains a considerable problem particularly in malaria endemic areas. As part of a prospective study to define clinical signs indicative of life threatening malaria video recordings were used to examine the level of agreement between clinicians for potentially important respiratory signs in 51 children. Overall agreement was good for recession, severe recession, and nasal flaring ($\kappa=0.57$, 0.50 , and 0.60 respectively) and substantial for deep breathing and the summary impression of respiratory distress ($\kappa=0.70$ and 0.69 respectively). However, within this substantial variation in interpretation was apparent between individual observers from slight to almost perfect agreement (κ values 0.10 – 0.92). Video is a useful tool to demonstrate interobserver variation and it may also allow training in recognition of signs and a means of standardising clinical signs between centres.

(*Arch Dis Child* 1995; 72: 334-336)

Keywords: malaria, interobserver variation, respiratory signs.

Common and readily treatable infectious diseases remain the major cause of childhood mortality in many parts of the world. While prevention is important, effective case management strategies comprising recognition, referral, and treatment are central to attempts to improve outcome in disease. Developing protocols for the integrated management of the sick child requires agreement on the nature and interpretation of key clinical signs. Together acute respiratory tract infection (ARI) and malaria account for approximately 30% of childhood mortality in The Gambia.¹ There is increasing awareness of considerable clinical overlap in these conditions in outpatients^{2 3} and recent studies in inpatients have identified 'respiratory distress' as a major prognostic feature in all children presenting with severe malaria (K Marsh *et al*, unpublished) as well as in those with severe malarial anaemia.⁴

As respiratory features appear to be both common and of prognostic importance in malaria, working definitions are required that denote degree of severity and if possible distinguish the condition from ARI. These may then be subjected to prospective assessment in malaria endemic and non-endemic areas. In the inpatient studies referred to earlier, respiratory distress was a composite assessment of a child's respiratory status and thus liable to differences in interpretation. The first step therefore is to determine whether clinicians agree on the presence or absence of respiratory distress and on some of the signs that may be used to describe it.

Respiratory signs in a single child may vary within a short period of time confounding attempts to assess the differences in interpretation between observers. Video records may overcome this problem for examination of some respiratory signs as each observer sees the child at exactly the same point in time and has exactly the same information available to them. Any difference in reporting of the sign should reflect true differences in interpretation by the observers rather than possible changes in the signs themselves. Our aim was to use video recordings to assess interobserver agreement for individual respiratory signs and the composite sign of respiratory distress in children.

Patients and methods

Video records of children were obtained with parental consent on the day of a child's admission to the Kenya Medical Research Institute (KEMRI) malaria research ward of Kilifi District Hospital between August 1993 and January 1994. Records from 51 children with a mean age of 26 months (range 3-96 months) were of a quality thought sufficient for subsequent analysis. Edited records lasted from one and a half to two and a half minutes and all were reviewed at one sitting by five observers each of whom had five or more years' clinical experience including at least two in paediatrics. Each observer was asked to respond either yes or no if the following signs were considered present or not: nasal flaring, recession, severe recession, deep breathing, and respiratory distress. An answer of 'don't know' could be given if the observer considered the video of insufficient quality to make a judgment. No definition of the clinical signs in question was provided as these signs are part of our routine clinical examination and no discussion of the records was allowed during viewing. Each

Kenya Medical Research Institute, Kilifi, Kenya and Nuffield Department of Clinical Medicine, University of Oxford, John Radcliffe Hospital, Oxford
M English
S Murphy
J Crawley
K Marsh

Kenya Medical Research Institute, Kilifi, Kenya
I Mwangi
N Peshu

Correspondence to:
Dr M C English, KEMRI, CRC, Kilifi Unit, PO Box 428, Kilifi, Kenya.

Accepted 11 January 1995

Table 1 Yes responses made by each observer and κ values for overall agreement; values are yes responses/number of recordings of sufficient quality to make assessment (%)

Clinical sign and κ value	Observer No				
	I	II	III	IV	V
Deep breathing $\kappa=0.70$	25/49 (51)	30/51 (59)	24/51 (47)	34/51 (67)	26/50 (52)
Nasal flaring $\kappa=0.60$	29/48 (54)	29/47 (62)	31/51 (61)	27/48 (56)	28/49 (57)
Recession $\kappa=0.57$	35/51 (69)	36/51 (71)	28/51 (55)	31/51 (61)	23/51 (45)
Severe recession $\kappa=0.50$	20/51 (39)	20/51 (51)	23/51 (45)	24/51 (47)	5/51 (10)
Respiratory distress $\kappa=0.69$	32/48 (67)	34/51 (67)	24/51 (47)	23/41 (45)	26/51 (51)

video recording was repeated a maximum of three times if required by an observer.

Interobserver agreement for the five features was assessed for each possible pairing of observers using the κ statistic, which is a measure of the agreement occurring beyond that expected by chance alone.⁵ Values for the κ statistic may vary from -1 to $+1$ with -1 indicating perfect disagreement, 0 indicating the level of agreement expected by chance alone, and $+1$ perfect agreement. A simple scale for interpreting κ values has been described by Landis and Koch where any positive value for κ indicates some agreement greater than chance and values of $0.00-0.20$, $0.21-0.40$, $0.41-0.60$, $0.61-0.80$, and $0.81-1.00$ represent slight, fair, moderate, substantial, and almost perfect agreement respectively.⁶

When calculating agreement on a sign any record containing a don't know response was excluded from analysis of observer pairs and this was also done when comparing an individual observer with the consensus view of the other observers (defined as the same response from at least three of the four). In this latter group records were also excluded from analysis when no consensus was present. Overall agreement between the five observers for each sign was assessed using a modification of the κ statistic suggested by O'Connell and Dobson for use with multiple observers.⁷

Results

All observers completed the questionnaire fully and for each observer the frequency of a yes response is shown as the numerator in table 1. The denominator is the number of recordings the observer felt was of sufficient quality to make an assessment, that is total number of recordings minus don't know responses. The largest number of don't know responses was recorded for nasal flaring with 12 (5% of responses) and seven of these occurred in responses to just three records. Otherwise don't know responses were uncommon and given by only one of the five observers in assessing any clinical sign.

Differences in responses between observers reflects the difference in interpretation of the same information by clinicians (table 1). The most striking example was severe recession where observer V gave five yes responses and observer IV gave 24. Similarly for recession a yes response was much more common from

observers I and II (69% and 71% of records respectively) than observer V (45%). This would appear to represent very different observer dependent thresholds required before a yes response was given for any clinical sign.

Agreement between observer pairs for the five variables also displayed marked differences. Thus, for severe recession agreement varied from slight ($\kappa=0.10$, observer I *v* observer V) to almost perfect ($\kappa=0.88$, observer III *v* observer IV). For the other signs examined agreement was at least moderate in all cases but still displayed considerable variation. (The κ ranges were $0.43-0.84$, $0.49-0.91$, $0.47-0.76$, and $0.57-0.92$ for deep breathing, nasal flaring, recession, and respiratory distress respectively.)

Comparison of an individual observer with a consensus of the remaining four observers (complete agreement in three of the four) gives an indication of which observer is least compatible with the group in describing that sign (table 2). The difference between observer V and the others is once again obvious for severe recession while for deep breathing, nasal flaring, and recession observer I and for respiratory distress observer II show least agreement with the group. Values for κ in this analysis are generally higher because all records containing a don't know response or in which consensus was not reached were excluded from analysis and so κ values act as a comparative guide to agreement only.

Overall agreement for each physical sign by all five observers is shown in table 1. There was little agreement for don't know responses and broadly similar rates of agreement for yes and no responses. Severe recession, recession, and flaring gave moderate agreement overall while that for deep breathing and respiratory distress was substantial. Excluding the least compatible observer from analysis made a noticeable difference only in severe recession when moderate agreement became substantial (κ 0.50 to 0.69).

Discussion

Little attention has been paid to interobserver variation in the reporting of respiratory signs in childhood. In one of the few studies to address the issue Wang *et al* found poor agreement between observers examining children less than 2 years of age for the signs wheeze, retractions, and respiratory rate but slightly better agreement for general assessment.⁸ In adults several studies have shown relatively poor agreement for a broad range of respiratory signs elicited by multiple observers.⁹⁻¹¹

Table 2 Agreement between a single observer and a consensus of the other observers assessed by κ value

Clinical sign	Observer No				
	I	II	III	IV	V
Deep breathing	0.61	0.74	0.82	0.73	0.91
Nasal flaring	0.59	0.95	0.91	0.82	0.95
Recession	0.56	0.66	0.75	0.79	0.71
Severe recession	0.55	0.81	0.95	0.90	0.24
Respiratory distress	0.77	0.70	0.87	0.87	0.96

The ability of video recordings to remove the confounding effect of changing physical signs may explain why the differences found between observers in this study were generally less than those reported in previous studies. Other factors may have been that only a small number of relatively simple signs were examined and that all observers had been working in the same hospital for at least one year. However, the fact that almost perfect agreement can be achieved for some signs and between some observers means that moderate agreement should not be accepted as standard. Indeed, most striking of all is the extent of disagreement that can exist between observers who work together without its recognition.

All the signs chosen for analysis are present to greatly variable degrees in different children but clinical practice usually demands a yes or no response to the presence of the sign. Each observer develops a personal definition to guide their response as a result of training and experience. None of these definitions can be wrong but describing a sign as present or absent is only useful when those giving and those receiving the description have the same understanding of the implications. In this study the term severe recession clearly meant different things to different observers.

Much effort has been put into devising case management protocols for sick children in developing countries and these largely rely on recognition of simple clinical signs to define severely ill children for more intensive treatment. For acute respiratory infection different centres have found different signs to be of value in predicting poor outcome and these have been proposed as indicators of severity.¹²⁻¹⁵ One of the few studies to test the current protocol for management of ARI in two different centres showed reasonable sensitivity and specificity for the diagnosis of ARI in the two centres.¹⁶ However, each centre used their own clinicians to provide both training and gold standard. The findings thus suggest that health workers can be trained to agree with the trainer (although κ values for agreement on diagnosis were only 0.36 and 0.47 overall) but the extent of agreement between the two gold standards was not examined. Interestingly, the prevalences of pneumonia and severe pneumonia were very different in the two centres. While this probably reflects true differences, differences in the gold standards cannot be excluded as an important factor.

We have found respiratory distress to be of the same prognostic importance in children admitted with malaria as coma (K Marsh *et al.*, unpublished) and others have proposed that it should be the major indication for transfusion in severe anaemia.⁴ If respiratory distress is to be used as an indication of the need for urgent referral and treatment in the case management of malaria then it requires definition by a sign or signs that can be reliably recognised by health workers. Examination of video records has shown that there is substantial agreement on the presence of respiratory distress among clinicians in our unit and good or substantial agreement on the presence of some respiratory signs. Video

should enable us to further improve our level of agreement and provide a reference standard of our definition of respiratory signs and respiratory distress possibly for prospective assessment in other centres. The extent to which observers or centres agree with our reference standard could then be assessed using the κ measure of agreement. Marked lack of agreement might indicate a need for centres to cooperate in achieving a consensus, reference standard. Agreed definitions depicted as video records might then be used as teaching aids for health worker trainers if not for the health workers themselves. Measurement of interobserver variation would allow assessment of the success of training and could be a part of regular retraining as it allows observers who differ most from the consensus to be identified and targeted.

It is important to emphasise that such an approach does not describe which centre or which individual is right or wrong in their interpretation of a sign. However, until consensus is achieved it is likely that debate over which clinical signs are most valuable will continue to obscure the true value of a sign. Thus, research that relies on the reporting of clinical signs should be preceded by an assessment of the interobserver variation among study clinicians and include training until consensus is achieved. This will be particularly important in multicentre studies.

This paper is published with the permission of the Director of the Kenya Medical Research Institute. We thank our colleagues at the KEMRI Unit and Kilifi District Hospital for continued help and support. We also thank Jo Schellenberg, Bob Snow, and Dayo Forster for helpful advice and Brett Lowe for technical support. This work forms part of the KEMRI - Oxford University - Wellcome Trust collaborative programme at Kilifi. Dr Marsh is a Wellcome senior research fellow in clinical science.

- Greenwood B, Greenwood A, Bradley A, *et al.* Deaths in infancy and early childhood in a well vaccinated, rural, West African population. *Ann Trop Paediatr* 1987; 7: 91-9.
- O'Dempsey TJD, McArdle TF, Laurence BE, *et al.* Overlap in the clinical features of pneumonia and malaria in African children. *Trans R Soc Trop Med Hyg* 1993; 87: 662-5.
- Redd SC, Bloland PB, Kazembe PN, *et al.* Usefulness of clinical case-definitions in guiding therapy for African children with malaria or pneumonia. *Lancet* 1992; 340: 1140-3.
- Lackritz EM, Campbell CC, Ruebush TK, *et al.* Effect of blood transfusion on survival among children in a Kenyan hospital. *Lancet* 1992; 340: 524-8.
- Cohen JA. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; 20: 37-46.
- Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-74.
- O'Connell DL, Dobson AJ. General observer-agreement measures on individual subjects and groups of subjects. *Biometrics* 1984; 40: 973-83.
- 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: 106-9.
- Gjorup T, Bugge PM, Jensen AM. Interobserver variation in assessment of respiratory signs. *Acta Medica Scandinavica* 1984; 216: 61-6.
- Smyllie HC, Blendis LM, Armitage P. Observer disagreement in physical signs of the respiratory system. *Lancet* 1965; ii: 412-3.
- Spiteri MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet* 1988; i: 873-5.
- Campbell H, Bypass P, Lamont AC, *et al.* Assessment of clinical criteria for identification of severe acute lower respiratory tract infections in children. *Lancet* 1989; i: 297-9.
- Onyango FE, Steinhoff MC, Wafula EM, *et al.* Hypoxaemia in young Kenyan children with acute lower respiratory infection. *BMJ* 1993; 306: 612-5.
- Shann F, Barker J, Poore P. Clinical signs that predict death in children with severe pneumonia. *Pediatr Infect Dis J* 1989; 8: 852-5.
- Spooner V, Barker J, Tulloch S, *et al.* Clinical signs and risk factors associated with pneumonia in children admitted to Goroka Hospital, Papua New Guinea. *J Trop Pediatr* 1989; 35: 295-300.
- Mulholland EK, Simoes EA, Costales MO, *et al.* Standardized diagnosis of pneumonia in developing countries. *Pediatr Infect Dis J* 1992; 11: 77-81.