Wagai, J; Senga, J; Fegan, G; English, M (2009) Examining Agreement between Clinicians when Assessing Sick Children. PloS one, 4 (2). e4626. ISSN 1932-6203 DOI: https://doi.org/10.1371/journal.pone.0004626

Downloaded from: http://researchonline.lshtm.ac.uk/5384/

DOI: 10.1371/journal.pone.0004626

Usage Guidelines

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by/2.5/
Examining Agreement between Clinicians when Assessing Sick Children

John Wagai1*, John Senga1, Greg Fegan2,3, Mike English1,4

1 Child and Newborn Health Group, Centre for Geographic Medicine, Kenyan Medical Research Institute (KEMRI)–Wellcome Trust Programme, Nairobi, Kenya, 2 Kenya Medical Research Institute (KEMRI), Centre for Geographic Medicine Research-Coast, Kilifi, Kenya, 3 Infectious Disease Epidemiology Unit, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom, 4 Department of Paediatrics, University of Oxford, Oxford, United Kingdom

Abstract

Background: Case management guidelines use a limited set of clinical features to guide assessment and treatment for common childhood diseases in poor countries. Using video records of clinical signs we assessed agreement among experts and assessed whether Kenyan health workers could identify signs defined by expert consensus.

Methodology: 104 videos representing 11 clinical sign categories were presented to experts using a web questionnaire. Proportionate agreement and agreement beyond chance were calculated using kappa and the AC1 statistic. 31 videos were selected and presented to local health workers, 20 for which experts had demonstrated clear agreement and 11 for which experts could not demonstrate agreement.

Principal Findings: Experts reached very high level of chance adjusted agreement for some videos while for a few videos no agreement beyond chance was found. Where experts agreed Kenyan hospital staff of all cadres recognised signs with high mean sensitivity and specificity (sensitivity: 0.897–0.975, specificity: 0.813–0.894); years of experience, gender and hospital had no influence on mean sensitivity or specificity. Local health workers did not agree on videos where experts had low or no agreement. Results of different agreement statistics for multiple observers, the AC1 and Fleiss’ kappa, differ across the range of proportionate agreement.

Conclusion: Videos provide a useful means to test agreement amongst geographically diverse groups of health workers. Kenyan health workers are in agreement with experts where clinical signs are clear-cut supporting the potential value of assessment and management guidelines. However, clinical signs are not always clear-cut. Video recordings offer one means to help standardise interpretation of clinical signs.


Received October 17, 2008; Accepted January 26, 2009; Published February 27, 2009

Copyright: © 2009 Wagai et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Funds from a Wellcome Trust Senior Fellowship awarded to Dr. Mike English (#076827) made this work possible. The funders had no role in the design, conduct, analyses or writing of this study nor in the decision to submit for publication.

Competing Interests: Mike English has at times provided advice to WHO panels on common childhood illnesses and their management in low-income countries. There are no other conflicts of interest.

* E-mail: jwagai@nairobi.kemri-wellcome.org

Introduction

Improving child survival is a major global health priority with the focus being on low income countries where the largest burden of childhood disease is found. Within these countries pneumonia, diarrhoea, malaria, measles and malnutrition are responsible for up to 80% of infant and child mortality[1]. Case management guidance for these diseases is provided as part of the Integrated Management of Childhood Illnesses (IMCI) strategy formulated by the WHO and UNICEF. This strategy relies in part on rapid and appropriate recognition of sick children and subsequent prompt treatment and, or referral.

A relatively small set of clinical features are used for the identification and assessment of severity of illness in such approaches. [2] The majority of the recommended clinical features are included on the basis of evidence of their value accumulated over the last thirty years. The research evaluating their utility has largely been based on experts' assessments or assessments made by health workers trained by experts. In the latter case, training is employed to ensure adequate agreement between individuals involved in the research [3,4,5,6,7,8,9]. However, there have been very few attempts to examine whether experts agree [10]. We are also unaware of any specific attempts to explore agreement between experts and health workers in multiple, routine settings receiving no special training for the apparently obvious reason that it is impossible to organise for large numbers of clinicians to observe the same patient at the same time. However, use of video recordings potentially overcomes this problem, at least for selected clinical signs. Video recordings and images have been widely used in developed countries for research, teaching and for assisting patients management. In tele-dermatology for example videos and/or images have been extensively used in competency and agreement studies [11,12,13]. Our interest was, therefore, to begin to develop an approach that could explore...
agreement between experts and identify clear ‘standard’ examples of clinical signs based on their consensus. Once developed identification of these standard signs could be tested amongst health workers in non-research settings who are expected to use guidelines which are based on such clinical signs.

Methods

Videos

Video recordings of key clinical signs lasting 20–45 seconds were made in children attending Kenyan hospitals after obtaining informed, signed consent from their parents or caretakers and in such a way that treatment was not delayed or interrupted. 336 video recordings of children were made from which 104 videos, displaying one of 11 clinical features (see Table 1), were selected based on the investigators opinion of the quality of the recording. We linked to each clip a single question on the presence or absence of a single clinical sign (or where more appropriate the degree of severity). An international panel of experts (described below) reviewed all videos in January 2008 and 20 videos were identified as clear examples of a child with or without a specific clinical feature after applying criteria of strong consensus. An additional 11 high quality videos were selected for which there was no clear consensus on the presence or absence of one of the clinical features. These 31 video clips then comprised a test set for review by local health workers (see table 2). The process of video selection and presentation to the two panels is demonstrated in figure 1.

International expert panel

Thirty-two leading paediatric clinical researchers and practitioners were approached by email inviting them to review the 104 video clips. Sixteen (16) experts completed the review task using a password protected web questionnaire (http://www.cnhg-kenya.com). Data from the web based questionnaire was collected into a MySQL database available to the authors. Four (4) experts from regions with low internet band-width had to complete the same questionnaire using a memory stick version that generated a simple spreadsheet that could be emailed back to the investigators to give a final expert panel comprising 20 people (listed in the acknowledgements).

Local panels

The test set of 31 videos were embedded into a Microsoft PowerPoint presentation. Each slide showed one video and asked the same single question about the presence, absence or grade of a clinical feature that the experts had responded to. Health workers from 8 Kenyan government hospitals in relatively rural districts and from the national hospital viewed the presentation in groups and recorded their opinions on preformatted paper questionnaires. Video viewing continued until all participants had made a decision

Table 1. Demographic characteristics of the international (HE) and local panels (hospitals H1 to H8 and the national hospital, HN).

<table>
<thead>
<tr>
<th></th>
<th>HE</th>
<th>H 1</th>
<th>H 2</th>
<th>H 3</th>
<th>H 4</th>
<th>H 5</th>
<th>H 6</th>
<th>H 7</th>
<th>H 8</th>
<th>HN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>20</td>
<td>11</td>
<td>15</td>
<td>17</td>
<td>11</td>
<td>21</td>
<td>12</td>
<td>4</td>
<td>8</td>
<td>55</td>
</tr>
<tr>
<td>Female [%]</td>
<td>12 [60%]</td>
<td>5 [45%]</td>
<td>8 [53%]</td>
<td>9 [52%]</td>
<td>4 [36%]</td>
<td>10 [47%]</td>
<td>6 [50%]</td>
<td>1 [25%]</td>
<td>4 [50%]</td>
<td>30 [54%]</td>
</tr>
<tr>
<td>Number of Interns</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>12</td>
<td>0</td>
<td>6</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

1Most of these were consultant paediatricians.
doi:10.1371/journal.pone.0004626.t001

Table 2. Expert agreement observed for specific clinical sign groups within the whole panel of examples (n = 104), for signs (n = 20 examples) selected on the basis of very high proportionate agreement (Po) and for signs (n = 11) selected where proportionate agreement was low (*no example with low proportionate agreement available).

<table>
<thead>
<tr>
<th>Clinical category</th>
<th>Options available to panelists</th>
<th>All videos</th>
<th>High consensus set</th>
<th>Low consensus set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N = 104</td>
<td>Po</td>
<td>AC1</td>
</tr>
<tr>
<td>Acidotic breathing</td>
<td>[Yes] (No)</td>
<td></td>
<td>0.68</td>
<td>0.63</td>
</tr>
<tr>
<td>Capillary refilling</td>
<td>[&lt;2s] [2–3s][&gt;3s]</td>
<td>7</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td>Sunken eyes</td>
<td>[Yes] (No)</td>
<td>12</td>
<td>0.79</td>
<td>0.77</td>
</tr>
<tr>
<td>Ability to feed</td>
<td>[Yes] (No)</td>
<td>8</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>Indrawing</td>
<td>[Yes] (No)</td>
<td>10</td>
<td>0.72</td>
<td>0.67</td>
</tr>
<tr>
<td>Head Nodding</td>
<td>[Yes] (No)</td>
<td>9</td>
<td>0.77</td>
<td>0.74</td>
</tr>
<tr>
<td>Pallor</td>
<td>[0][1][+][++]</td>
<td>9</td>
<td>0.71</td>
<td>0.66</td>
</tr>
<tr>
<td>Skin pinch</td>
<td>[&lt;1s] [1–2s][&gt;2s]</td>
<td>14</td>
<td>0.67</td>
<td>0.63</td>
</tr>
<tr>
<td>AVPU at Alert</td>
<td>[Yes] (No)</td>
<td>7</td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>AVPU at Voice</td>
<td>[Yes] (No)</td>
<td>6</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>AVPU at Pain</td>
<td>[Yes] (No)</td>
<td>10</td>
<td>0.84</td>
<td>0.83</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0004626.t002
and without any conferring amongst participants. Health worker panels in district hospitals comprised hospital staff available at the time of the study team's visit who were responsible for administering routine medical care to children. In the national hospital all university consultants and consultant trainees within the university department of paediatrics were invited to participate. In each hospital only one viewing session was offered, taking from 60 to 80 minutes.

Data processing and Analysis

Responses to clinical sign questions were either dichotomous (present/absent) or one of three ordinal grades (absent or normal/mild-moderate abnormality/severe abnormality). Data were analysed using Stata® version 9.2 (StataCorp, Texas, USA). For all 104 video clips viewed by experts' proportionate agreement, Fleiss's kappa and the AC1 statistic were calculated. (The AC1 statistic is designed to correct the overall agreement probability for chance agreement). [14] No weighting was applied. Both the Fleiss's kappa and AC1 statistic generate a number representing the proportionate agreement adjusted for chance where the value 0 means no agreement beyond chance and the value 1 means perfect agreement. We chose to calculate both these statistics as the more commonly used kappa may be misleading as trait prevalence varies [15,16]. We calculated the AC1 statistic and kappa for individual videos and groups of videos showing the same clinical feature and to examine the difference between AC1 and kappa we plotted the values of both measures for each of the 104 clinical videos reviewed by experts against the underlying proportionate agreement.

To identify the 20 consensus based ‘standard’ video clips we selected those with very high proportionate agreement amongst experts (Po≥0.94) on the presence, absence or grade of the sign in question. We also selected 11 video clips of high quality with an absence of consensus (proportionate agreement amongst experts Po< = 0.68) to form the 31 video test set. For each local health worker the ability to identify clear-cut clinical signs was assessed by calculating their sensitivity and specificity with respect to the 20 gold-standard video clips. In this analysis sensitivity is interpreted as health workers ability to detect a truly positive clinical sign and specificity as their ability to detect a truly negative clinical sign. Mean sensitivities and specificities were derived for groups aggregated by hospital, clinical background or other characteristics and compared in exploratory analyses. There were no clinically meaningful differences in sensitivities or specificities between groups (data not shown). Agreement statistics, kappa and AC1, were calculated for each hospital group for the 11 videos without a clinical consensus.

Ethical approval

Taking the videos and the subsequent study were approved by the KEMRI/National Ethical Review Committee.

Results

Twelve of the 20 international panellists were female, the median years of clinical experience was 14 (10th–90th percentile: 7.5–22 years). A total of 99 health workers from the district hospitals participated including clinical officer interns, registered clinical officers, medical officer interns, medical officers, nurses, and consultant paediatricians. (A clinical officer is a form of substitute doctor with a 3 year diploma in medicine). The national referral hospital panel included 55 paediatric consultants and registrars (consultant trainees). Details of participants' characteristics are presented in table 1.

The AC1 measure of agreement amongst the international experts was generally high for the 104 videos individually ranging from 0.62 to 0.92. For the 20 consensus videos, the AC1 measure of agreement was very high, ranging from 0.95 to 1.00. Agreement as assessed by kappa values was considerably lower than the AC1 in most cases (Table 2) with 29 of the 104 videos associated with poor or fair agreement (kappa<0.4) on a commonly used scale [14]. The relationship between AC1, kappa and the proportionate agreement is demonstrated in Figure 2. In common with the experts agreement scores for the videos in the high consensus set was very good within the 9 Kenyan hospital sites or when analysed according to health worker cadre (data not shown).

Within the high consensus set of videos there was an equal number of videos with either presence or absence of a sign. The
local panels achieved high sensitivity across all health worker cadres in identifying the presence of positive clinical signs with sensitivities ranging between 0.927 and 0.975. The specificity across health worker categories was marginally lower at between 0.813 and 0.886 (Table 3). For the 11 videos with no consensus among experts, the crude proportionate agreement (0.48–0.70), AC1 (0.30–0.64) and kappa values (0–0.02) calculated for the different health worker groups or hospitals were low, indicative of absence of agreement in common with the experts.

**Discussion**

This study tested a novel method of conducting research on agreement when interpreting clinical signs between expert clinicians who were widely dispersed geographically. The successful use of the internet to host these videos and use of a version contained on a memory stick where internet access is still poor suggests that this approach can be further developed to include clinicians even from remote areas with access to a computer. Such methodologies have obvious extensions to teaching new skills to students and health workers. We then extended the approach, using a group presentation, to explore the ability of health workers in routine practice to identify consensus defined clinical signs.

It is possible that a different set of experts would have classified the signs presented on videos differently. However, we included experts from a wide variety of settings internationally. It is also possible that agreement within local hospital panels was high because we used an open presentation despite our attempts to limit contamination between observers. Despite these potential limitations we believe the study demonstrated that very clear consensus can be reached over the presence (or absence/grade) of specific clinical signs amongst experts. Furthermore it also demonstrated that where experts have a clear view on a clinical sign then health workers of a wide variety of cadres and with widely different levels of clinical experience in routine practice, at least in Kenya, are also able to identify the clinical sign. This provides some reassurance that teaching or guidelines based on these clinical signs have the potential to be understood and implemented widely. However, the study also demonstrated that for many clinical videos experts showed only moderate or even poor agreement. Where experts found it hard to agree health workers in routine settings also found it hard to agree. This finding has several implications.

Firstly, clinical signs may be depicted better as a spectrum from obviously present to obviously not present with the position on the spectrum for any one child or video being best represented by the proportionate agreement amongst multiple, expert observers. The
and the AC1 statistic (Figure 2). At the extremes of crude agreement the AC1 and Fleiss’ kappa scores approached each other. For the other values of crude agreement, kappa scores were usually lower than AC1 scores and were not linearly correlated with crude agreement.

In conclusion, we have shown that there can be widespread agreement in identifying obvious examples of clinical signs amongst all types of clinicians. However, greater attention should be paid to establishing where possible standardised thresholds for decisions on when a sign is or is not present, as appropriate, to delineate a particular condition. Video records provide one possible means to achieve this. Clinicians should also be more aware of the development of statistical theory underpinning measures of agreement to avoid well-described pitfalls. This study adds to the wider body of evidence on work done to understand workers abilities in recognising signs recommended by IMCI [3,4,5,6,7,8,9].

Acknowledgments

The authors are grateful to the following experts (listed alphabetically) for their participation in this process: Steve Allen, James Berkley, Kalifa Bojang, Abdalla Brooks, Thomas Cherian, Jane Crawley, Trevor Duke, Anthony Emenul, Steve Graham, Tabish Hazir, Richard Idro, Prakash Jeena, Kathryn Maitland, Shabir Mahdi, Kim Mulholland, Assumpta Muriithi, Hadija Mwatemi, Dennis Nansera, Charles Newton, Shamim Qazi, Michael Van Henegbroek, Martin Weber.

We would also like to thank Dr. Joln Wachira for help in collecting the video recordings, the panel participants in Kenya and the parents of children who consented to the video recordings whom we cannot name individually. This work is published with the permission of the Director of KEMRI.

Author Contributions

Conceived and designed the experiments: JW JS ME. Performed the experiments: JW ME. Analyzed the data: JNW GF ME. Contributed reagents/materials/analysis tools: JW JS. Wrote the paper: JW GF ME.

References