Expert Reliability for the World Health Organization Standardized Ultrasound Classification of Cystic Echinococcosis

Nadia Solomon,[1,2] Paul J. Fields,[1] Francesca Tamarozzi,[4] Enrico Brunetti,[4,5] and Calum N. L. Macpherson[1,2,3]*

[1]*Windward Islands Research and Education Foundation (WINDREF), St. George's, Grenada;* [2]*School of Medicine, St. George's University, St. George's, Grenada;* [3]*School of Veterinary Medicine, St. George's University, St. George's, Grenada;* [4]*WHO Collaborating Centre for Clinical Management of Cystic Echinococcosis, University of Pavia, Pavia, Italy;* [5]*Department of Infectious Diseases, San Matteo Hospital Foundation, Pavia, Italy*

Abstract. Cystic echinococcosis (CE), a parasitic zoonosis, results in cyst formation in the viscera. Cyst morphology depends on developmental stage. In 2003, the World Health Organization (WHO) published a standardized ultrasound (US) classification for CE, for use among experts as a standard of comparison. This study examined the reliability of this classification. Eleven international CE and US experts completed an assessment of eight WHO classification images and 88 test images representing cyst stages. Inter- and intraobserver reliability and observer performance were assessed using Fleiss' and Cohen's kappa. Interobserver reliability was moderate for WHO images ($\kappa = 0.600$, $P < 0.0001$) and substantial for test images ($\kappa = 0.644$, $P < 0.0001$), with substantial to almost perfect interobserver reliability for stages with pathognomonic signs (CE1, CE2, and CE3) for WHO ($0.618 < \kappa < 0.904$) and test images ($0.642 < \kappa < 0.768$). Comparisons of expert performances against the majority classification for each image were significant for WHO ($0.413 < \kappa < 1.000$, $P < 0.005$) and test images ($0.718 < \kappa < 0.905$, $P < 0.0001$); and intraobserver reliability was significant for WHO ($0.520 < \kappa < 1.000$, $P < 0.005$) and test images ($0.690 < \kappa < 0.896$, $P < 0.0001$). Findings demonstrate moderate to substantial interobserver and substantial to almost perfect intraobserver reliability for the WHO classification, with substantial to almost perfect interobserver reliability for pathognomonic stages. This confirms experts' abilities to reliably identify WHO-defined pathognomonic signs of CE, demonstrating that the WHO classification provides a reproducible way of staging CE.

INTRODUCTION

Cystic echinococcosis (CE) is a chronic and highly complex zoonotic infection of economic and public health importance on all the inhabited continents. The ubiquitous distribution is due to the presence of the predominant definitive host, the domestic dog, which harbors the cestode *Echinococcus granulosus* sensu lato. Human infection results in the development of fluid- and parasite-filled cysts in the viscera, particularly the liver and lungs.[1,2] Research indicates that some cysts can remain asymptomatic for over a decade, and that some spontaneously degenerate over time.[3–8] Degeneration of cysts is accompanied by physical changes in cyst composition and morphology, which can be easily visualized via various imaging techniques. Studies have established that parasite viability is markedly reduced in cysts in degenerated states.[3,9,10]

In 2003, the World Health Organization (WHO) published a standardized classification of ultrasound (US) images of CE (Figure 1).[11] US is easy to use in field settings, noninvasive, and sensitive and specific for CE.[12–14] Cyst stages as described by the classification are defined by their observable physical characteristics and account for changes in parasite activity/viability as cyst morphology of the different cyst stages. The WHO classification consists of six stages.[11] CE1, CE2, and CE3 cysts display pathognomonic signs, whereas CL, CE4, and CE5 cysts are undifferentiated early cystic lesions, caseated, and calcified, respectively, and have suggestive features but not pathognomonic signs.[11] CE3A and CE3B substages of the CE3 stage were officially incorporated in the classification, but, while described

in the publication, substages of CE2 were not officially adopted.[11] Substages were not labeled in the original official classification image.

The WHO Informal Working Group on Echinococcosis (IWGE) intended for the standardized CE classification "... to be easy to use, whatever the setting" and was agreed on by radiologists, surgeons, physicians, and researchers most involved in the field at that time.[11] Demonstrating the reliability of a standardized classification of CE 1) allows for more accurate comparison of cysts and disease stages, 2) facilitates communication between experts in the field around the world, 3) maximizes comprehension among those experts, 4) encourages the standardization of treatment (as current treatment guidelines based on WHO classification),[15] and 5) promotes confidence in discussions of disease stages and treatment choices. In this study, the inter- and intraobserver reliability of the WHO CE US classification among experts was investigated.

MATERIALS AND METHODS

Eight digitized, black and white, two-dimensional (2-D) images representing all six cyst stages and substages from the WHO standardized US classification of CE were randomly interspersed within 88 additional images of the same quality (including at least eight images of each cyst stage). Images were obtained from patients in Turkana, Kenya. The eight images from the WHO classification represented "standard" images, whereas the 88 selected images represented a "test" group. These images were used to create a 96-question online assessment.

Eleven individuals who were very familiar with the WHO classification and over many years had frequently used it were recruited as experts and instructed to use the WHO classification to classify these images, but were not told whether the lesions were caused by CE: instead, they were

*Address correspondence to Calum N. L. Macpherson, School of Medicine, St. George's University, True Blue Campus, True Blue, P.O. Box 7, St. George's, Grenada. E-mail: cmacpherson@sgu.edu
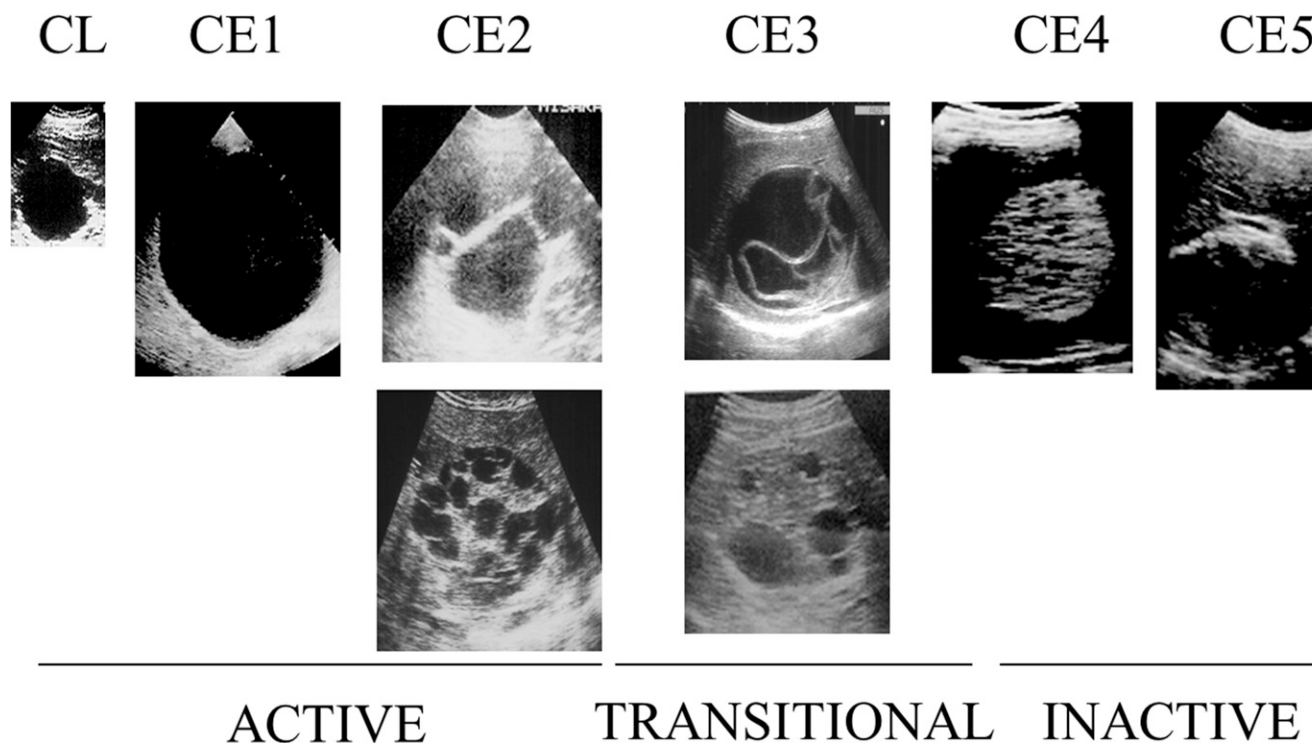
FIGURE 1. Original World Health Organization Informal Working Group on Echinococcosis standardized ultrasound classification of cystic echinococcosis (CE) published in 2003,[11] depicting CL, CE1, CE2 (A and B), CE3 (A and B), CE4, and CE5 cyst stages.

asked to use the WHO classification to classify US images. Although expected to use their expertise of CE staging via WHO classification, the 2003 WHO classification scheme (Figure 1) was also provided along with the link to the assessment for the experts' reference. Consent was taken as completion of the assessment, and this was specified on distribution of the assessment link to the experts.

Upon opening the assessment, a US image was displayed at the top of the page, below which experts were instructed to "classify based on the WHO classification system." They were asked to select among the following options: "CL," "CE1," "CE2A," "CE2B," "CE3A," "CE3B," "CE4," "CE5," and "cannot determine."

This assessment was programmed such that items were viewed serially, each item required a response to progress to the next item, and there was no option for forward or backward navigation. The presentation of all images was standardized: the only difference between questions was the US image itself. The assessment was administered on two separate occasions to a group of international experts.

Some experts were recruited directly via e-mail contact, whereas others received the assessment from directly recruited experts. Experts were recruited based on their activity in the field of tropical medicine and their familiarity with CE and US, and most had been members of the WHO-IWGE, the team that had formulated the WHO classification. Assessment administrations were separated by a minimum time interval of 1 week to ensure reliability was being tested as opposed to memory. The original, official 2003 WHO CE US classification decided on by all members of the WHO-IWGE was provided to the experts for their reference with each administration of the assessment. To address interobserver reliability and observer performance,

ratings from the first assessments completed by experts were collated and compared. To address intraobserver reliability, each individual expert's first and second assessments were compared. Interobserver reliability was assessed using Fleiss's kappa, and observer performance and intraobserver reliability were assessed using Cohen's kappa. Microsoft Excel was used to perform the analyses.[16] Significance was defined at the $P < 0.05$ level. The strength of kappa values was evaluated using the following scale: < 0.00 as poor, $0.00 < \kappa < 0.20$ as slight, $0.21 < \kappa < 0.40$ as fair, $0.41 < \kappa < 0.60$ as moderate, $0.61 < \kappa < 0.80$ as substantial, and $0.81 < \kappa < 1.00$ as almost perfect.[17]

RESULTS

Interobserver reliability and observer performance. Eleven experts participated in the interobserver reliability portion of the study. Participating experts have specializations including tropical medicine, parasitology, infectious disease, gastroenterology, hepatology, and radiology and work out of Italy, Turkey, Grenada, Bulgaria, Argentina, Romania, Kenya, and the United Kingdom. Interobserver Fleiss' kappa for WHO images was calculated and found to be a moderate 0.600 ($P < 0.0001$). For test items, Fleiss' kappa was found to be substantial, at 0.644 ($P < 0.0001$).

Interobserver Fleiss' kappa for each stage as defined by the WHO classification (Table 1) was also calculated for the WHO image group and the test image group. Kappa values for all stages in both groups were significant ($P < 0.0001$).

A cyst stage was then assigned to each image in the assessment. The US CE image stages were retained for the standard group, and the highest frequency cyst stage chosen by the expert participants for each image was assigned to

TABLE 1

Interobserver kappa values for cyst stages as defined by the WHO system are charted for the WHO image group and test image group

WHO CE US cyst stages	Kappa for WHO images	Kappa for test images
CL	0.364	0.496
CE1	0.796	0.672
CE2	0.636	0.719
CE3A	0.904	0.768
CE3B	0.618	0.642
CE4	0.565	0.624
CE5	0.478	0.586

CE = cystic echinococcosis; US = ultrasound; WHO = World Health Organization.
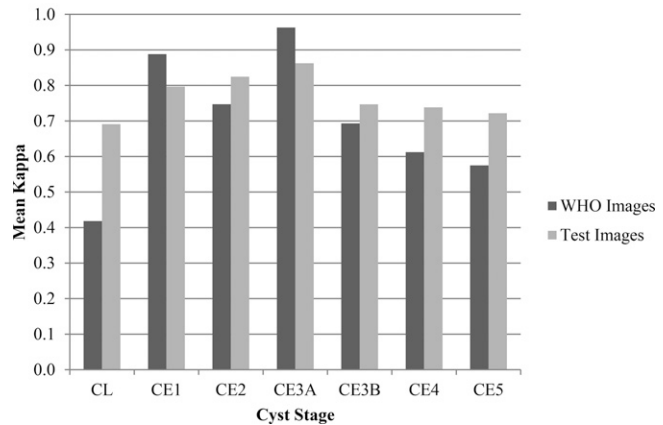

FIGURE 3. The average of observers' Cohen's kappa values for each cyst stage for the World Health Organization (WHO) classification and test image groups.

that image for the test group. This was done to compare observers' individual selections with these majority stage assignments, providing a kappa value for each observer. These kappa values would act as "scores" for each observer, allowing individual performances to be compared. These kappa scores ranged from moderate to almost perfect (0.413–1.000, $P < 0.005$) for the WHO US CE standard images and from substantial to almost perfect (0.718–0.905, $P < 0.0001$) for the test images. Kappa values with confidence intervals for the WHO images and test images (Figure 2) were then graphed to compare expert performances.

Individual observer's kappa values for each cyst stage were then averaged and evaluated to assess strengths and weaknesses in the observers' abilities to use the classification and of the WHO CE US classification system itself for WHO and test images (Figure 3).

CL (WHO κ = 0.364, test κ = 0.496), CE4 (WHO κ = 0.565, test κ = 0.624), and CE5 (WHO κ = 0.478, test κ = 0.586) groups are observed to have lower measures of interobserver reliability compared with groups with pathognomonic signs, CE1 (WHO κ = 0.796, test κ = 0.672), CE2 (WHO κ = 0.636, test κ = 0.719), CE3A (WHO κ = 0.904, test κ = 0.768), and CE3B (WHO κ = 0.618, test κ = 0.642). To more directly evaluate reliability of pathognomonic stages, the nonpathognomonic stages (CL, CE4, and CE5) were combined with the "cannot determine" category into an overarching nonpathognomonic category. When interobserver reliability was reevaluated after the nonpathognomonic stages were grouped together, the interobserver kappa increased to a substantial 0.754 ($P < 0.0001$) for the WHO images, and a substantial 0.701 ($P < 0.0001$) for the test images.

When the CE2 stage WHO images were analyzed as separate CE2A and CE2B substages, the kappa value for the CE2A substage remained substantial (κ = 0.684), whereas the kappa value for the CE2B substage was reduced to only fair (κ = 0.224). Reliability for test images was found to be moderate for both the CE2A (κ = 0.520) and CE2B substages (κ = 0.460).

Intraobserver reliability. Of the 11 experts who completed the assessment, 10 participated in the intraobserver part of the study by completing the assessment a second time. A kappa value of agreement reflecting intraobserver
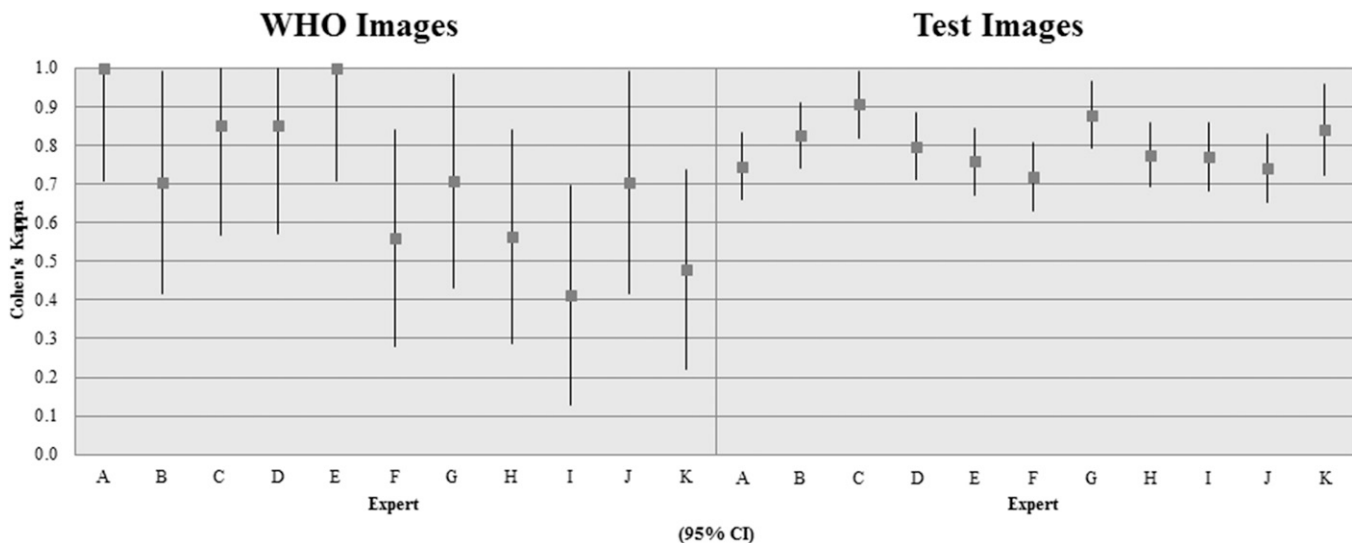

FIGURE 2. Expert observers' kappa values comparing each expert's classifications to majority classifications for World Health Organization (WHO) classification and test images (95% confidence interval).
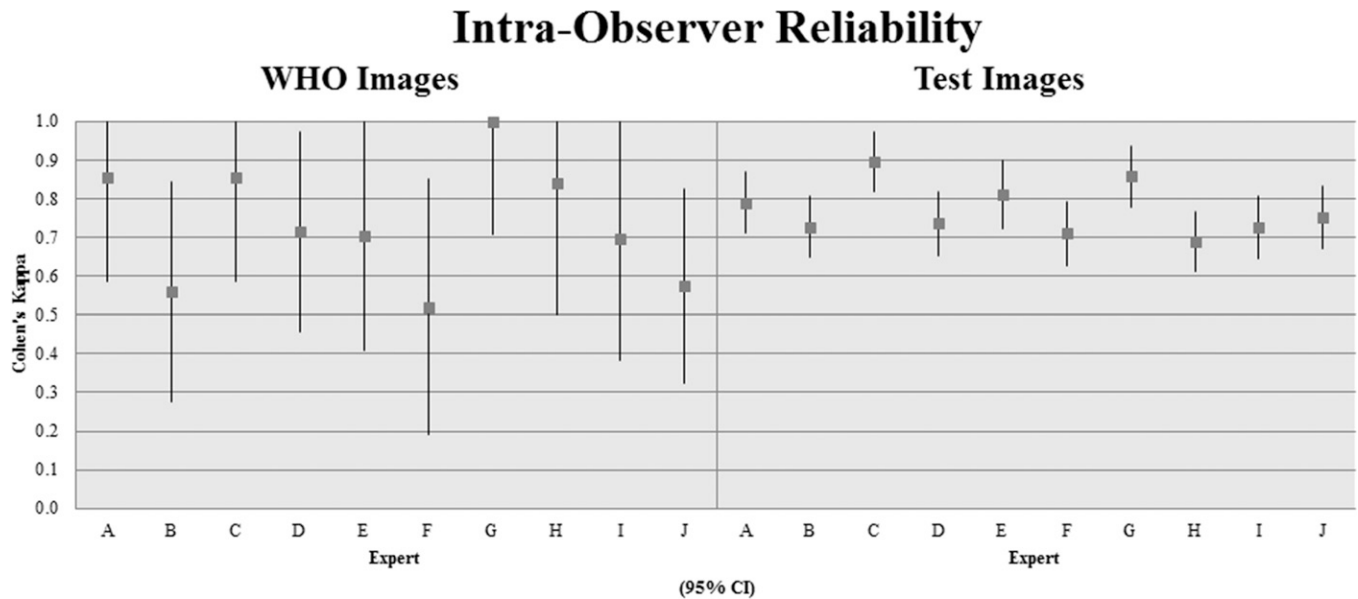
Intra-Observer Reliability

FIGURE 4. Expert observers' intraobserver kappa values for World Health Organization (WHO) classification and test images (95% confidence interval).

reliability was calculated by comparing each observer's two assessments. Kappa values for intraobserver reliability among experts ranged from moderate to almost perfect for the WHO image group (0.520–1.000, $P < 0.005$), and from substantial to almost perfect for the test image group (0.690–0.896, $P < 0.0001$) (Figure 4).

DISCUSSION

The CL, CE4, and CE5 stages lack pathognomonic signs on US, and therefore have uncertain etiology. Although some features of these stages are highly suggestive of CE, CE cannot be diagnosed with absolute certainty. This is reflected by groups without pathognomonic signs (CL, CE4, and CE5) having the lower measures of interobserver reliability compared with groups with pathognomonic signs (CE1, CE2, and CE3). This supports the expectation that experts are reliable in diagnosing and staging CE when pathognomonic signs are present, and more reliable in these cases than when pathognomonic signs are absent. The increase in interobserver reliability observed when data was analyzed after nonpathognomonic stages were combined and considered as a single group also supports the assertion that experts recognize pathognomonic signs of CE as defined by the WHO standardized US classification and are able to stage accordingly and with consistency.

The WHO image group contains only eight images with a single image representing each stage/substage. With such a small sample size, difficulty or disagreement regarding a single image can have a strong effect on statistical outcomes. Before commencing the study, the CE2B image was noted to have certain characteristics (i.e., thickened regions of echogenicity between daughter cysts), which could lead some to classify it as the CE3B stage. When viewed along with the rest of the WHO images, and especially when directly compared with the CE3B image, the CE2B image is observed to contain far less degenerated material between

its daughter cysts. This ability to compare makes classification more obvious. If presented with both the WHO CE2B and CE3B images and directed to decide which belongs to which stage, it is clear that the CE3B image has fewer daughter cysts suspended in a great deal more echogenic material, and this image would quickly be classified accordingly as CE3B; this would influence the other image, with far less echogenic material between the daughter cysts, to be classified as CE2B. When presented on its own, however, without a point of comparison, the echogenic regions surrounding the daughter cysts appear more distinct. This highlights the importance in appraising each image individually to classify it (i.e., "cyst X appears as a unilocular cyst with a laminated membrane, and so should be classified as CE1"), rather than appraising images together and classifying them comparatively ("cyst X and cyst Y both have daughter cysts, but cyst X has less degeneration than cyst Y, so is more representative of the CE2 stage in comparison with cyst Y"). In fact, this image was classified as a CE3B cyst by five of the 11 expert observers, whereas only five of the 11 classified the image as a CE2B cyst. Maintenance of substantial reliability for the CE2A image ($\kappa = 0.684$) and reduction to fair reliability for the CE2B image ($\kappa = 0.224$) when the CE2 stage WHO images were analyzed using their separate substages support the theory that this CE2B image may not be the most reliably recognized representation of the CE2 stage.

Regarding the CE2A/2B subclassification, findings of moderate reliability for both substages in the test image group suggest that, although they are not yet strictly defined, definitions could be developed and used. The concept underlying these substages is based on identifying the extent of cyst development. Differentiation into distinct substages (CE2A, or "early" CE2; and CE2B, or "late" CE2) could have implications for treatment: if a CE2 cyst contains only two or three daughter cysts (an example of what could be considered as CE2A), it might be more susceptible to treatments typically reserved for earlier cyst stages as opposed to more

complex CE2 cysts with many daughter cysts (CE2B). Future studies would be necessary to provide more insight into the reliability and usefulness of this subcategorization.

Since the initial publication of this classification, a paper on CE treatment guidelines by three members of the WHO-IWGE was published and included the WHO classification scheme with some of the images replaced, including the aforementioned CE2B image.[15] The findings of this study support the replacement of the CE2B image and propose an additional amendment to the newer version of this classification to highlight the variability that can be observed in the CE2 stage, the inclusion of an image which may represent CE2A, a cyst only partially filled with daughter cysts (Figure 5).[15]

After the assignment of a stage to each image—based on the WHO classification scheme for the WHO images and majority observations for the test images—all experts had moderate to almost perfect agreement for the WHO image group (0.413 > κ > 1.000), and substantial to almost perfect agreement for the test image group (0.718 > κ > 0.905). The most common decisions negatively affecting performance on the WHO images among the four experts with moderate kappa values (F, H, I, and K) were the classification of the CE2B image as CE3B, and the classification of stages without pathognomonic signs (CL, CE4, and CE5) as "cannot determine."

Experts' intraobserver reliability kappa values were more varied for the set of WHO images than for the set of test images. The intraobserver reliability for the eight WHO images ranged from moderate to almost perfect (0.520 < κ < 1.000) for each of the 10 experts who participated. Variations in intraobserver values for the WHO image group are difficult to interpret, however, there being only two observations per expert (before and after) and due to the small sample size of images (eight).

The intraobserver reliability for the test images ranged from substantial to almost perfect (0.690 < κ < 0.896) for every observer. This indicates that observers are consistent in their individual classification schemas, and experts are able to recognize features and differentiate cyst stages based upon them. If individual observers recognize the differences between cysts in a reliable manner, this strengthens the assertion that the stages are distinguishable from one another.

Variations in experts' classifications of individual images may also be attributable to confusion over the task itself. As experts were not explicitly told whether the images were of CE, they may have classified images in either part of the assessment 1) under the supposition that the images were not all CE and therefore may have more frequently classified images perceived as uncertain as "cannot determine," or 2) under the expectation that all images were CE and should be classified accordingly in all cases but those eliciting high uncertainty.

A major limitation of this study lies in the use of 2-D images to classify cysts. The WHO classification is presented as a series of 2-D images, but a 2-D image of any one cyst may not completely represent a cyst's contents. Observation in situ, or presentation with video clips, would provide a more comprehensive view of cyst contents and allow for more accurate classification. Additionally, US is only as accurate as the person operating it is experienced: if asked to provide a 2-D image, a less experienced operator may not capture a cysts' structure and contents as effectively. As such, appropriate recognition of pathognomonic signs and classification of CE stage rely on the quality of the image produced, and on the person producing it.

CONCLUSIONS

Evaluation of the data reveals that interobserver reliability for the WHO standardized US classification is substantial to almost perfect for stages with pathognomonic signs, demonstrating that experts are reliable in identifying pathognomonic signs of CE even in 2-D images. Findings of intraobserver reliability values in the substantial to almost
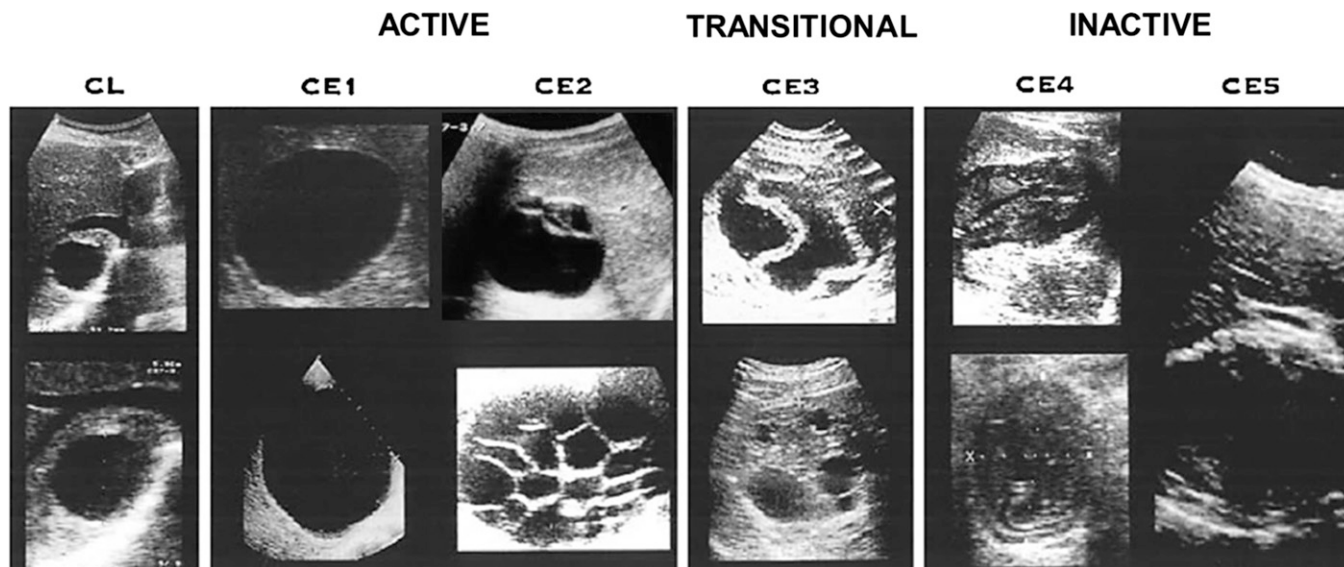

FIGURE 5. World Health Organization Informal Working Group on Echinococcosis ultrasound classification of cystic echinococcosis, with suggested amendments.[11,15]

perfect range for the test image group for all experts support this conclusion, and indicate that individual observers can reliably recognize distinguishing features of cysts and classify based on those features.

The conclusions which may be drawn from this study are limited, primarily due to the use of 2-D images for the assessment. The WHO classification is intended for use in the field and in clinical settings, where patients' cysts can be viewed from an infinite number of positions and angles, allowing for better visualization of cysts and their features. In 2-D images, certain pathognomonic signs, such as the presence of a double membrane surrounding a unilocular CE1 cyst, may be difficult to discern. This makes classification challenging, and increases the likelihood of potential observer disagreement due to different observer criterions for categorical assignment. The finding of substantial to almost perfect intraobserver reliability for the majority of experts, however, inspires confidence that experts are very consistent in recognizing the structures being assessed, and that these structures are present as described by the WHO US classification of CE. The use of cyst images taken from multiple angles, or even video clips of cysts, may have a profound effect on standardizing observer criterions, thereby improving intraobserver reliability.

Discussion of the original WHO CE2B image highlights the importance of appraising each image individually when choosing stage-representative images, particularly when the images are intended to be 2-D representations of what otherwise would be observed in vivo from multiple perspectives. For the purpose of choosing representative images for a classification, an assessment such as the one developed for this study could feasibly be used to facilitate the selection of images: this would objectively maximize agreement among those most involved in the field and quickly rule out difficult to classify images simply based on poor agreement.

Experts are able to reliably identify pathognomonic signs of CE as defined by the WHO standardized US classification, and are consistent in using their own individual criteria for identifying pathognomonic signs of disease and classifying cysts based on those characteristics. Nonetheless, care must be taken in evaluating any US classification intended for use in field epidemiological and clinical settings using 2-D images. Observation in situ, or with video clips, will provide a more effective and more accurate method of assessing CE and its classification.

Experts' close agreement on CE staging based on pathognomonic signs of CE as described by the WHO classification affords confidence that the classification provides a reliable and readily interpretable way of staging CE in field and clinical settings: experts from all different parts of the world were able to agree on what they were shown and classify accordingly. Moving toward using a standardized classification has inherent advantages, including facilitating the use of the classification for standardized WHO treatment protocols, and readily comparing the public health importance of this disease from different parts of the world.

Authors' addresses: Nadia Solomon and Calum N. L. Macpherson, Windward Islands Research and Educational Foundation (WINDREF), St. George's, Grenada, and St. George's University School of Medicine, St. George's, Grenada, E-mails: nsolomon12@gmail.com and cmacpherson@sgu.edu. Paul J. Fields, Windward Islands Research and Educational Foundation (WINDREF), St. George's, Grenada, E-mail: pjfphd@comcast.net. Francesca Tamarozzi, WHO Collaborating Centre for Clinical Management of Cystic Echinococcosis, University of Pavia, Pavia, Italy, E-mail: f_tamarozzi@yahoo.com. Enrico Brunetti, WHO Collaborating Centre for Clinical Management of Cystic Echinococcosis, University of Pavia, Pavia, Italy, and Department of Infectious Diseases, San Matteo Hospital Foundation, Pavia, Italy, E-mails: enrico.brunetti@unipv.it.

REFERENCES

1. Thompson RCA, McManus DP, 2001. Aetiology: parasites and life-cycles. Eckert J, Gemmell MA, Meslin F-X, Pawlowski ZS, eds. *WHO/OIE Manual on Echinococcosis in Humans and Animals: A Public Health Problem of Global Concern*. Paris, France: World Organisation for Animal Health (Office International des Epizooties), 1–19.
2. Brunetti E, Garcia HH, Junghanss T, 2011. Cystic echinococcosis: chronic, complex, and still neglected. *PLoS Negl Trop Dis 5:* e1146.
3. Kern P, 2003. *Echinococcus granulosus* infection: clinical presentation, medical treatment and outcome. *Langenbecks Arch Surg 388:* 413–420.
4. Romig T, Zeyhle E, Macpherson CN, Rees PH, Were JB, 1986. Cyst growth and spontaneous cure in hydatid disease. *Lancet 1:* 861.
5. Keshmiri M, Baharvahdat H, Fattahi SH, Davachi B, Dabiri RH, Baradaran H, Ghiasi T, Rajabimashhadi MT, Rajabzadeh F, 1999. A placebo controlled study of albendazole in the treatment of pulmonary echinococcosis. *Eur Respir J 14:* 503–507.
6. Rogan MT, Hai WY, Richardson R, Zeyhle E, Craig PS, 2006. Hydatid cysts: does every picture tell a story? *Trends Parasitol 22:* 431–438.
7. Larrieu E, Del Carpio M, Mercapide CH, Salvitti JC, Sustercic J, Moguilensky J, Panomarenko H, Uchiumi L, Herrero E, Talmon G, Volpe M, Araya D, Mujica G, Mancini S, Labanchi JL, Odriozola M, 2011. Programme for ultrasound diagnoses and treatment with albendazole of cystic echinococcosis in asymptomatic carriers: 10 years of follow-up of cases. *Acta Trop 117:* 1–5.
8. Frider B, Larrieu E, Odriozola M, 1999. Long-term outcome of asymptomatic liver hydatidosis. *J Hepatol 30:* 228–231.
9. Gil-Grande LA, Rodriguez-Caabeiro F, Prieto JG, Sánchez-Ruano JJ, Brasa C, Aguilar L, García-Hoz F, Casado N, Bárcena R, Alvarez AI, 1993. Randomised controlled trial of efficacy of albendazole in intra-abdominal hydatid disease. *Lancet 342:* 1269–1272.
10. Hosch W, Junghanss T, Stojkovic M, Brunetti E, Heye T, Kauffmann GW, Hull WE, 2008. Metabolic viability assessment of cystic echinococcosis using high-field 1H MRS of cyst contents. *NMR Biomed 21:* 734–754.
11. WHO Informal Working Group, 2003. International classification of ultrasound images in cystic echinococcosis for application in clinical and field epidemiological settings. *Acta Trop 85:* 253–261.
12. Macpherson CN, Bartholomot B, Frider B, 2003. Application of ultrasound in diagnosis, treatment, epidemiology, public health and control of *Echinococcus granulosus* and *E. multilocularis*. *Parasitology 127 (Suppl):* S21–S35.
13. Macpherson CN, Milner R, 2003. Performance characteristics and quality control of community based ultrasound surveys for cystic and alveolar echinococcosis. *Acta Trop 85:* 203–209.
14. Macpherson CN, Romig T, Zeyhle E, Rees PH, Were JB, 1987. Portable ultrasound scanner versus serology in screening for hydatid cysts in a nomadic population. *Lancet 2:* 259–261.
15. Brunetti E, Kern P, Vuitton DA, 2010. Expert consensus for the diagnosis and treatment of cystic and alveolar echinococcosis in humans. *Acta Trop 114:* 1–16.
16. Microsoft, 2010. *Microsoft Excel*. Redmond, WA: Microsoft.
17. Landis JR, Koch GG, 1977. The measurement of observer agreement for categorical data. *Biometrics 33:* 159–174.