ULTRA SONO GRAPHY

Intra- and interobserver reliability of gray scale/dynamic range evaluation of ultrasonography using a standardized phantom

Song Lee, Joon-Il Choi, Michael Yong Park, Dong Myung Yeo, Jae Young Byun, Seung Eun Jung, Sung Eun Rha, Soon Nam Oh, Young Joon Lee

Department of Radiology, Seoul St. Mary's Hospital, The Catholic University of Korea College of Medicine, Seoul, Korea

ORIGINAL ARTICLE

http://dx.doi.org/10.14366/usg.13021
pISSN: 2288-5919 • eISSN: 2288-5943
Ultrasonography 2014;33:91-97

Purpose: To evaluate intra- and interobserver reliability of the gray scale/dynamic range of the phantom image evaluation of ultrasonography using a standardized phantom, and to assess the effect of interactive education on the reliability.

Methods: Three radiologists (a resident, and two board-certified radiologists with 2 and 7 years of experience in evaluating ultrasound phantom images) performed the gray scale/dynamic range test for an ultrasound machine using a standardized phantom. They scored the number of visible cylindrical structures of varying degrees of brightness and made a 'pass or fail' decision. First, they scored 49 phantom images twice from a 2010 survey with limited knowledge of phantom images. After this, the radiologists underwent two hours of interactive education for the phantom images and scored another 91 phantom images from a 2011 survey twice. Intra- and interobserver reliability before and after the interactive education session were analyzed using κ analyses.

Results: Before education, the κ-value for intraobserver reliability for the radiologist with 7 years of experience, 2 years of experience, and the resident was 0.386, 0.469, and 0.465, respectively. After education, the κ-values were improved (0.823, 0.611, and 0.711, respectively). For interobserver reliability, the κ-value was also better after the education for the 3 participants (0.067, 0.002, and 0.547 before education; 0.635, 0.667, and 0.616 after education, respectively).

Conclusion: The intra- and interobserver reliability of the gray scale/dynamic range was fair to substantial. Interactive education can improve reliability. For more reliable results, double-checking of phantom images by multiple reviewers is recommended.

Keywords: Ultrasonography; Imaging phantoms; Health care quality assurance; Reproducibility of results

Received: December 1, 2013
Revised: December 28, 2013
Accepted: December 30, 2013

Correspondence to:
Joon-Il Choi, MD, Department of Radiology, Seoul St. Mary's Hospital, The Catholic University of Korea College of Medicine, 222 Banpo-daero, Seocho-gu, Seoul 137-701, Korea
Tel. +82-2-2258-1431
Fax. +82-2-599-6771
E-mail: dumkycji@gmail.com

How to cite this article:
Lee S, Choi JI, Park MY, Yeo DM, Byun JY, Jung SE, et al. Intra- and interobserver reliability of gray scale/dynamic range evaluation of ultrasonography using a standardized phantom. Ultrasonography. 2014 Apr;33(2):91-97.

Introduction

In the modern era of medicine, diagnostic imaging has become a crucial tool in correct diagnosis, which underpins appropriate treatment. Quality assurance (QA) of medical imaging is of paramount importance.

In Korea, QA for computed tomography (CT), magnetic resonance imaging (MRI), and mammography has been required by law since 2004. Accreditation programs for these imaging modalities are run by the Korean government, and QA testing is performed by the Korean Institute for Accreditation of Medical Imaging under the direction of the Ministry of Health and Welfare [1–3]. However, QA for ultrasonography (US) examinations has not yet been legislated, reflecting the diversity of roles and performance of US devices used in clinical practice and the lack of iodizing radiation. However, in the United States, some scientific bodies have formulated recommendations for US QA [4–6]. The Korean government is currently formulating additional regulations of medical imaging modalities including US.

In Korea, US screening of the liver for hepatocellular carcinoma (HCC) is a part of the National Cancer Screening Program run by the National Cancer Control Institute, which is a part of the National Cancer Center. This program has included surveys regarding QA of US [7–9]. In these surveys, six test items were assessed for phantom image evaluation (dead zone, vertical and horizontal measurement, axial and lateral resolution, sensitivity, gray scale/dynamic range) [7,8]. Among them, gray scale/dynamic range was the most common cause of the failure of phantom image evaluation [7–9]. However, the assessment of gray scale/dynamic range is subjective and might be influenced by the experience and inclination of reviewers. Therefore, for legal regulation, the reliability of subjective items should be validated.

We designed a study to evaluate the intra- and interobserver reliability of a gray scale/dynamic range test using a standardized US phantom. The aims of this study were to verify the intra- and interobserver reliability of the gray scale/dynamic range test in reviewers with different experience levels, and to determine the influence of education sessions on the reliability.

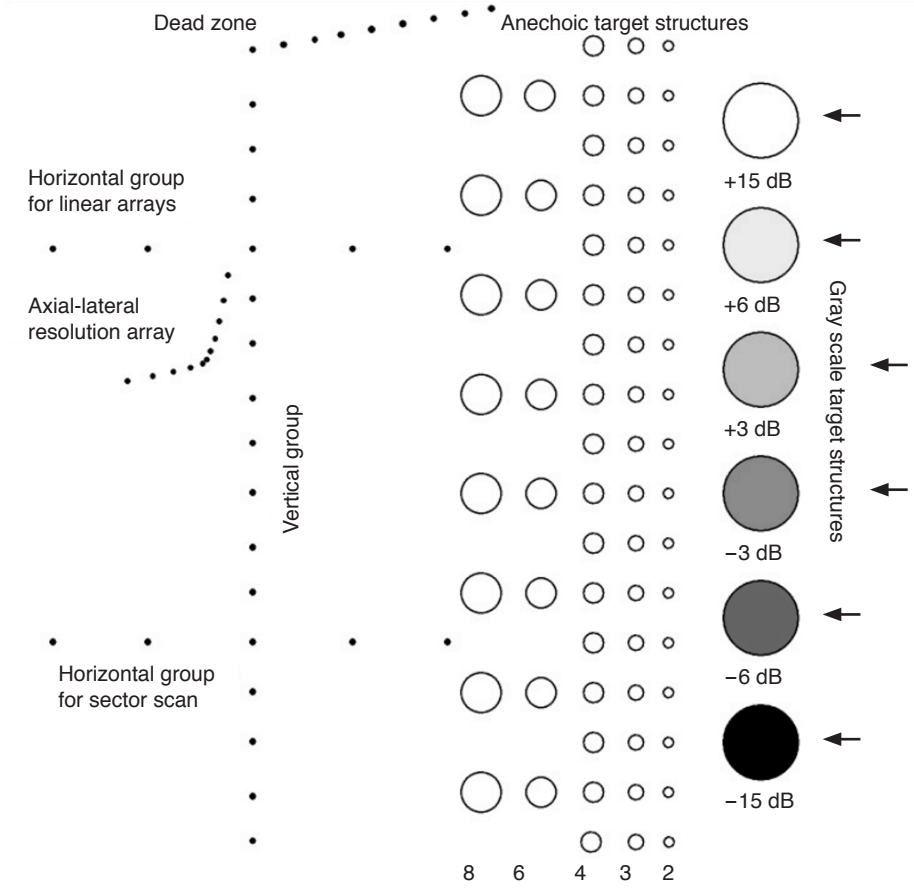

Fig. 1. Target diagram of an ATS 539 multipurpose phantom. The gray scale/dynamic range can be evaluated using six structures with different contrast values (arrows) [12].

Materials and Methods

Institutional Review Board and Institutional Animal Care and Use Committee approvals were not required because this study did not use any human or animal data.

Acquisition of Phantom Images

Phantom images were recruited as part of a nationwide survey in Korea for the investigation of the quality of US scanners for the screening of HCC in high-risk patients. Forty-nine phantom images from the 2010 survey and 91 phantom images from the 2011 survey were obtained. All of the phantom images were digital images, including Digital Imaging and Communications in Medicine (DICOM) and JPEG file types; no film or thermal paper images were evaluated. An ATS 539 multipurpose phantom (ATS Laboratories, Bridgeport, CT, USA) was used because this phantom is recommended as the standard phantom for QA of US for abdominal imaging by the Korean Society of Radiology and the Korean Society of Ultrasound in Medicine [7–9]. This phantom has also been adopted in several studies as a test phantom [10,11]. The phantom is made of rubber-based tissue-mimicking material that matches the acoustic properties of human tissue and provides test structures (Fig. 1). Research assistants, who were researchers of the Korean Institute for Accreditation of Medical Imaging, transported a standard phantom to medical institutions and obtained the phantom images. All of the phantom images were obtained with a 3.0–5.0 MHz curved-array probe and software settings for abdominal ultrasound, using the test methods recommended by the manufacturer's manual and the American Association of the Physicist in Medicine (AAPM) guideline [5,12]. The scanning of the phantom was done by the research assistants in the presence of the physician on site.

The six test items evaluated were the dead zone, vertical and horizontal measurement, axial and lateral resolution, sensitivity, and gray scale/dynamic ranges. Among them, we assessed the reliability of the gray scale/dynamic range, which is the test item for evaluating the contrast of the images, and which uses the amplitude of the received echoes to vary the degree of brightness in US images. Six cylindrical targets with varying degrees of brightness were visible in the US images. The contrast values of six targets compared to the background material were +15, +6, +3, −3, −6, and −15 dB.

First Review Round

US phantom images were independently reviewed by three radiologists. Two were board-certified radiologists with 7 years of experience (reviewer 1) and 2 years of experience (reviewer 2) evaluating US phantom images. The third member was a junior resident who had 6 months experience in abdominal US (reviewer 3). The process used to evaluate the phantom images is summarized in Fig. 2. Each reviewer initially evaluated 49 phantom images from the 2010 survey with brief knowledge for judgment of the gray scale/dynamic range test. All of the images were reviewed on an M-view picture archiving and communications system (PACS) workstation monitor (Infinitt, Seoul, Korea). The number of cylindrical targets that appeared as discrete round structures through more than 180° were counted. The 'pass or fail' cutoff value was more than four cylindrical structures visible as round structures [8] (Figs. 3, 4). The first review consisted of two review sessions to calculate interobserver reliability, with two reviews of the phantom images. To avoid recall bias, the second review session was performed 2 weeks after the first session.

Interactive Education

After the first review round, the three reviewers received two hours

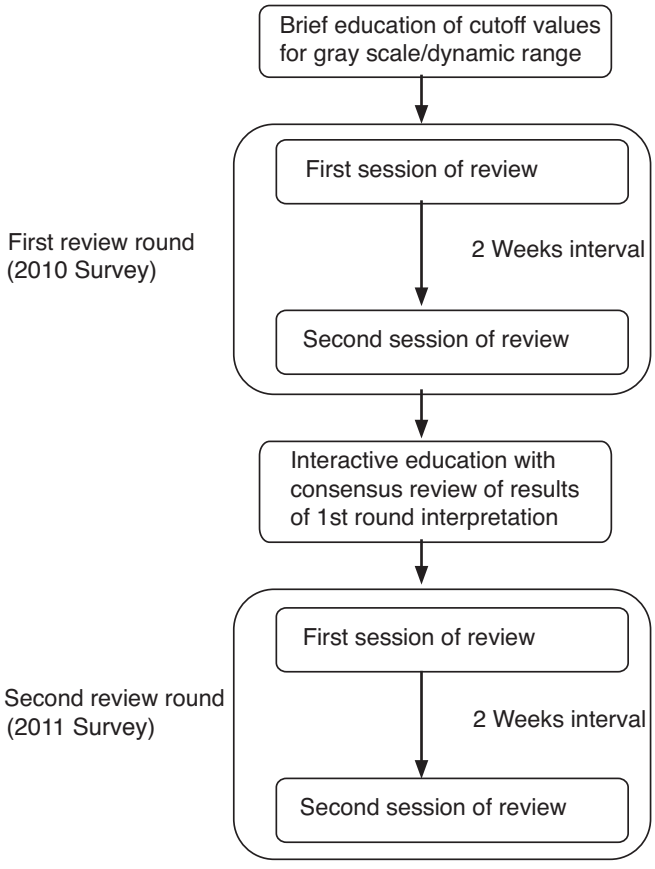

Fig. 2. **Flow chart of the review process.** The review process consists of two review rounds and one interactive education session between rounds. Each round consists of two review sessions separated from each other by more than two weeks. The first review round was performed with 49 phantom images from the 2010 survey and the second round with 91 phantom images from the 2011 survey.

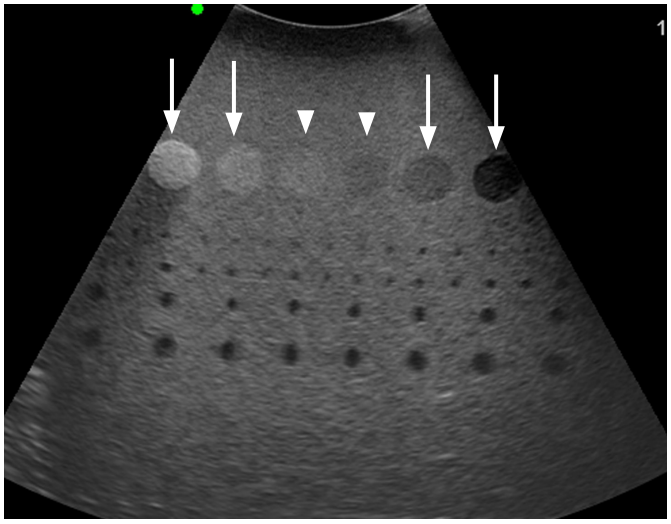

Fig. 3. An example of a "passed" phantom image of the gray scale/dynamic range. Four round structures are clearly visible over 180°. The contrast values of these targets are +15, +6, +3, −3, −6, and −15 compared to the background. In this case, four targets are clearly visible as round structures (arrows) and two structures are not visible as round over 180° (arrowheads).

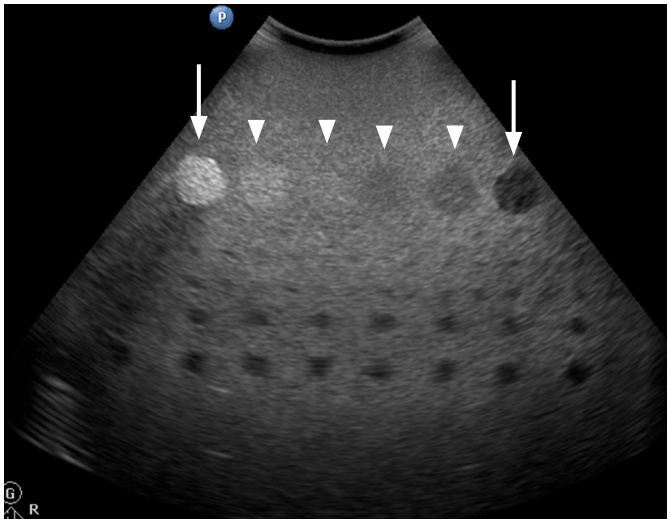

Fig. 4. An example of a "failed" phantom image of the gray scale/dynamic range. Only three round structures are clearly visible over 180°. The contrast values of these targets are +15, +6, +3, −3, −6, and −15 compared to the background. In this case, only three targets are clearly visible as round structures (arrows) and three structures are not visible as round over 180° (arrowheads).

of interactive education concerning phantom image evaluation. In this education session, they evaluated 49 phantom images from the 2010 survey together and reached consensus concerning the number of round structures. This education session was supervised by the most experienced reviewer (reviewer 1).

Second Review Round

After the interactive education, the three reviewers independently scored another set of 91 phantom images from the 2011 survey. The review also consisted of two review sessions with an intervening 2-week interval. The reviewers again recorded the number of visible round structures and passed or failed results according to the above cutoff value.

Statistical Analyses

Statistical analyses were performed using MedCac ver. 9.2 (MedCalc Software, Mariakerke, Belgium). Inter- and intraobserver reliability before and after the interactive education were analyzed using κ statistics with a weighted κ-value for the number of visible round structures (0–6 cylindrical structures) and κ-value for pass/fail. Interobserver reliability was calculated from the second set of data of each review round. Strengths of the intra- and interobserver reliability were determined using criteria detailed previously [13]. The following classification was used for the level of agreement by κ-value: <0.20, poor agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, good agreement; and 0.81–1.00, very good agreement. In addition, intra- and interobserver agreement rates (%) were also calculated in terms of passing or failing. Comparison of pass rates (%) of the two sets of interpretation was performed using Fisher exact test. Agreement rates were calculated because κ-values can be distorted due to the prevalence effect [14,15]. A $P<0.05$ indicated statistical significance.

Table 1. Intraobserver reliability (weighted κ-values) for the number of round structures

Reviewer	First round of review before education	Second round of review after education
1	0.403 (moderate)	0.847 (almost perfect)
2	0.536 (moderate)	0.450 (moderate)
3	0.439 (moderate)	0.675 (substantial)

Table 2. Intraobserver reliability (κ-values) for pass/fail

Reviewer	First round of review before education	Second round of review after education
1	0.386 (fair)	0.823 (almost perfect)
2	0.469 (moderate)	0.611 (substantial)
3	0.465 (moderate)	0.711 (substantial)

Results

Intraobserver Reliability

After interactive education, intraobserver reliability was improved for reviewers 1 and 3, but not for reviewer 2 (Table 1). On the other

Table 3. Intraobserver agreement rate (%) for pass/fail

Reviewer	First round of review before education	Second round of review after education	P-value[a)]
1	79.6 (39/49)	94.5 (86/91)	0.010
2	98.0 (48/49)	90.1 (82/91)	0.165
3	75.5 (37/49)	86.8 (79/91)	0.104

[a)]Calculated using Fisher exact test.

Table 4. Interobserver reliability (weighted κ-values) for round structures

Reviewer	First round of review before education	Second round of review after education
1–2	0.018 (slight)	0.639 (substantial)
2–3	0.002 (slight)	0.246 (fair)
1–3	0.275 (fair)	0.410 (moderate)

Table 5. Interobserver reliability (κ-values) for pass/fail

Reviewer	First round of review before education	Second round of review after education
1–2	0.067 (slight)	0.635 (substantial)
2–3	0.002 (slight)	0.667 (substantial)
1–3	0.547 (moderate)	0.616 (substantial)

Table 6. Interobserver agreement rate (%) for pass/fail

Reviewer	First round of review before education	Second round of review after education	P-value[a)]
1–2	73.5 (36/49)	90.1 (82/91)	0.014
2–3	51.0 (25/49)	67.0 (61/91)	0.071
1–3	79.6 (39/49)	70.3 (64/91)	0.315

[a)]Calculated using Fisher exact test.

hand, intraobserver reliability (κ-values) for passing/failing improved in all of the reviewers (Table 2). In addition, Fisher exact test of the intraobserver agreement rate (%) for passing/failing showed an improved reliability rate in reviewer 1, but no significant change in reviewers 2 or 3 (Table 3).

Interobserver Reliability

After interactive education, the interobserver reliability for scores improved all reviewer-pairs (Table 4). The interobserver reliability for passing/failing was also improved in all of the reviewer-pairs (Table 5). In addition, Fisher exact test of the interobserver reliability rate (%) for passing/failing showed an improved reliability rate between the reviewer 1–2 pair, but no significant change in the reviewer 2–3 and 1–3 pairs (Table 6).

Discussion

US screening of HCC is recommended by many scientific bodies worldwide [16–19]. The American Association for the Study of Liver Disease has stressed the importance of QA of screening US examinations for HCC [17]. In Korea, legal regulations concerning QA of medical imaging are limited to CT, MRI, and mammography. However, the need for QA of other imaging modalities is real and becoming more important; US, fluoroscopy, and positron emission tomography are expected to be included in the Korean regulatory framework soon. Surveys of QA of US have been done for several years in Korea as a prelude to legislation [7–9,20].

The QA of medical imaging consists of three components: personnel evaluation (assessment of personnel performing imaging studies), phantom image evaluation (assessment of the performance of the hardware and software of imaging devices), and clinical image evaluation (testing of imaging protocols). Among these, we concentrated on the phantom image evaluation because the failure rate was relatively higher compared to clinical image evaluation. In the analyses of the 3-year survey from 2008 to 2010, the failure rate of phantom image evaluation increased from 20.9% to 24.5%, and that of clinical image evaluation increased from 5.5% to 9.5% [7]. The failure rate of clinical imaging evaluation can also be reduced by the education of physicians. However, improving the performance of phantom image evaluation is challenging because it is related to the hardware itself and often requires a hardware upgrade. The most common cause of failure in phantom image evaluation was the gray scale/dynamic range, which represented 42.6% of failures overall [7]. However, assessment of the gray scale/dynamic range by visual inspection can be very subjective, and intraobserver and interobserver reliability should be validated for test items used for legal regulation.

Due to the subjective nature of visual inspection, computerized automated evaluation of parameters of US images has been considered [10,11,21–24]. A relatively high subjectivity of visual inspection has been reported [25]. However, in the real world, many US units use thermal paper or film as the output method, negating automatic computerized evaluation. The reality is that visual inspection needs to be more reliable. The present study is useful in this situation.

In this study, intraobserver reliability after interactive education was moderate to almost perfect, particularly the pass/fail reliability. Intraobserver reliability was also improved for both the number of round structures and the pass/fail decision after the interactive education, and intraobserver agreement rates were quite high after interactive education (>85% in all reviewers), although statistically significant improvement was evident only in reviewer 1. Therefore,

especially after proper education, intraobserver reliability was quite good.

Interobserver reliability for the number of round structures after education was fair to substantial. This result is somewhat disappointing, especially given the declined performance in reviewer 3 after education. Interobserver reliability for passing/failing was good and improved in all of the reviewers after education. However, the interobserver agreement rate was 67.0% to 90.1%, and significant improvement was observed only in the pair of reviewer 1–2. This result was poorer than the level of interobserver reliability of a previous study [8] that reported interobserver reliability of two reviewers who were experienced abdominal radiologists with more than 5 years of experience in evaluating US phantom images for gray scale/dynamic range (κ-value of 0.652 for the number of round structures and 0.969 for pass/fail, interobserver agreement rate 98.6%).

The cause of poorer interobserver agreement of our study might have been the relative inexperience of the reviewers, especially reviewer 3, who was a senior resident. The outcomes between the reviewer 2–3 and 1–3 pairs were inferior in both interobserver reliability (weighted κ-values) for scores after the education sessions and the interobserver agreement rate (%) for the pass/fail decision. The level of experience may influence interobserver reliability even after interactive education. In addition, before education, reviewer 2 had standards that were too generous and interobserver agreement was very low in the reviewer pairs including reviewer 2. However, this tendency was corrected after interactive education, and this provides evidence for the necessity of proper education for reviewers to attain reliable results.

Our results are not sufficient for testing concerning legal regulation, and therefore, reviewers for legal regulation should be experienced radiologists and appropriate education must be performed. Furthermore, a double-check system is mandatory to reduce personal errors. For CT, MRI, and mammography, two to five reviewers need to be involved in QA testing for legal regulation. This system should be adopted for the QA of US when legal regulation is legislated.

Our study has some limitations. First, only three reviewers participated. For the accreditation system and legal regulations, more robust results are needed. Studies involving multiple, experienced reviewers must be performed before legislation is formulated and implemented. Second, we analyzed various US scanners and probes with a single, standard phantom. The optimal phantoms for individual US units can vary. However, a previous study reported that the various combinations of scanners and probes do not significantly alter the results of phantom images [26]. Third, we only evaluated digital images. US units with analogue outputs, such as films or thermal papers, still comprise the majority of units in use. As the image quality of these analogue images is generally poorer than those of digital images, reliability could be inferior to the present results. For legal regulation, a study with a larger number of cases of analogue images with multiple, experienced reviewers will be necessary. Fourth, the data from the 2010 survey and 2011 survey might be of varying quality and this could make a difference in the reliability. However, the failure rates of the 2010 and 2011 surveys for the phantom image evaluation were similar according to a government report. Fifth, this study was performed in two months, and reviewer 3 (a junior resident) experienced many US cases during that period. Therefore, the performance of reviewer 3 could have improved for the second round of review compared to the first round, and this could have influenced the results of this study. However, the tendency of the results from other reviewers was robust, and the impact of the resident's increasing experience was probably not substantial.

In conclusion, the intraobserver reliability of the results of the gray scale/dynamic range was good, especially after the interactive education session. The interobserver reliability could be improved by education. Whether this approach will yield results necessary for legal regulation is unclear. Therefore, the involvement of experienced reviewers, proper education, and a double check system for accreditation are mandatory.

ORCID: Song Lee: http://orcid.org/0000-0002-2367-3588; Joon-Il Choi: http://orcid.org/0000-0003-0018-8712; Michael Yong Park: http://orcid.org/0000-0002-5247-1475; Dong Myung Yeo: http://orcid.org/0000-0002-1362-1362; Jae Young Byun: http://orcid.org/0000-0002-0038-3860; Seung Eun Jung: http://orcid.org/0000-0003-0674-5444; Sung Eun Rha: http://orcid.org/0000-0003-1514-929X; Soon Nam Oh: http://orcid.org/0000-0003-2373-7024; Young Joon Lee: http://orcid.org/0000-0001-8309-0272

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This study was supported in part by the Phillips Research Fund of the Korean Society of Ultrasound in Medicine. This study was also partly supported by a grant from the National Cancer Control Institute of the National Cancer Center of Korea and by the Ministry of Health, Welfare and Family Affairs of Korea. This study was performed with the help of the Korean Society of Radiology and the Korean Institute for Accreditation of Medical Imaging.

References

1. Kim YS, Jung SE, Choi BG, Shin YR, Hwang SS, Ku YM, et al. Image

quality improvement after implementation of a CT accreditation program. Korean J Radiol 2010;11:553-559.
2. Park HJ, Jung SE, Lee YJ, Cho WI, Do KH, Kim SH, et al. The relationship between subjective and objective parameters in CT phantom image evaluation. Korean J Radiol 2009;10:490-495.
3. Park HJ, Jung SE, Lee YJ, Cho WI, Do KH, Kim SH, et al. Review of failed CT phantom image evaluations in 2005 and 2006 by the CT accreditation program of the Korean Institute for Accreditation of Medical Image. Korean J Radiol 2008;9:354-363.
4. American College of Radiology. ACR technical standard for diagnostic medical physics performance monitoring of real time ultrasound equipment [Internet]. Reston, VA: American College of Radiology, 2011 [cited 2013 Dec 18]. Available from: http://www.acr.org/~/media/152588501B3648BA803B38C8172936F9.pdf.
5. Goodsitt MM, Carson PL, Witt S, Hykes DL, Kofler JM Jr. Real-time B-mode ultrasound quality control test procedures: report of AAPM Ultrasound Task Group No. 1. Med Phys 1998;25:1385-1406.
6. AIUM Technical Standards Committee. Quality assurance manual for gray scale ultrasound scanners: stage 2. Laural, MD: American Institute of Ultrasound in Medicine, 1995.
7. Choi JI, Jung SE, Kim PN, Cha SH, Jun JK, Lee HY, et al. Quality assurance in ultrasound screening for hepatocellular carcinoma using a standardized phantom and standard images: a 3-year national investigation in Korea. J Ultrasound Med 2014 [In press].
8. Choi JI, Kim PN, Jeong WK, Kim HC, Yang DM, Cha SH, et al. Establishing cutoff values for a quality assurance test using an ultrasound phantom in screening ultrasound examinations for hepatocellular carcinoma: an initial report of a nationwide survey in Korea. J Ultrasound Med 2011;30:1221-1229.
9. Kim PN, Lim JW, Kim HC, Yoon YC, Sung DJ, Moon MH, et al. Quality Assessment of Ultrasonographic Equipment Using an ATS-539 Multipurpose Phantom. J Korean Radiol Soc 2008;58:533-541.
10. Gibson NM, Dudley NJ, Griffith K. A computerised quality control testing system for B-mode ultrasound. Ultrasound Med Biol 2001;27:1697-1711.
11. Thijssen JM, Weijers G, de Korte CL. Objective performance testing and quality assurance of medical ultrasound equipment. Ultrasound Med Biol 2007;33:460-471.
12. ATS Laboratories. Clinical quality assurance phantoms: multipurpose phantom model 539. Bridgeport, CT: ATS Laboratories, 2000;2-21.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
14. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-558.
15. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543-549.
16. Park JW. Practice Guideline for Diagnosis and Treatment of Hepatocellular Carcinoma. Korean J Hepatol 2004;10:88-98.
17. Bruix J, Sherman M; American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma: an update. Hepatology 2011;53:1020-1022.
18. Kudo M, Izumi N, Kokudo N, Matsui O, Sakamoto M, Nakashima O, et al. Management of hepatocellular carcinoma in Japan: Consensus-Based Clinical Practice Guidelines proposed by the Japan Society of Hepatology (JSH) 2010 updated version. Dig Dis 2011;29:339-364.
19. European Association For The Study Of The Liver; European Organisation For Research And Treatment Of Cancer. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2012;56:908-943.
20. Kim PN, Kim KW, Byun JH. Quality assessment of hepatic ultrasound images examined after a medical check-up. J Korean Soc Ultrasound Med 2009;28:31-37.
21. Browne JE, Watson AJ, Gibson NM, Dudley NJ, Elliott AT. Objective measurements of image quality. Ultrasound Med Biol 2004;30:229-237.
22. Dudley NJ, Griffith K, Houldsworth G, Holloway M, Dunn MA. A review of two alternative ultrasound quality assurance programmes. Eur J Ultrasound 2001;12:233-245.
23. Sipila O, Mannila V, Vartiainen E. Quality assurance in diagnostic ultrasound. Eur J Radiol 2011;80:519-525.
24. Rowland DE, Newey VR, Turner DP, Nassiri DK. The automated assessment of ultrasound scanner lateral and slice thickness resolution: use of the step response. Ultrasound Med Biol 2009;35:1525-1534.
25. Sipila O, Blomqvist P, Jauhiainen M, Kilpelainen T, Malaska P, Mannila V, et al. Reproducibility of phantom-based quality assurance parameters in real-time ultrasound imaging. Acta Radiol 2011;52:665-669.
26. Tradup DJ, Hangiandreou NJ, Taubel JP. Comparison of ultrasound quality assurance phantom measurements from matched and mixed scanner-transducer combinations. J Appl Clin Med Phys 2003;4:239-247.