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

Joon-Il Choi, MD, Pyo Nyun Kim, MD, Woo Kyoung Jeong, MD, Hyun Cheol Kim, MD, Dal Mo Yang, MD, Sang Hoon Cha, MD, Jae-Joon Chung, MD

Article includes CME test

Received February 16, 2011, from the Department of Radiology, Seoul St Mary's Hospital, the Catholic University of Korea, Seoul, Korea (J.-I.C.); Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Seoul, Korea (P.N.K.); Department of Radiology, Hanyang University Guri Hospital, Hanyang University College of Medicine, Seoul, Korea (W.K.J.); Department of Radiology, Kyung Hee University Hospital at Gangdong, Seoul, Korea (H.C.K., D.M.Y.); Department of Radiology, Ansan Hospital College of Medicine, Korea University, Ansan, Korea (S.H.C.); and Department of Radiology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea (J.-J.C.). Revision requested March 7, 2011. Revised manuscript accepted for publication March 30, 2011.

This study was supported by a grant from the National Cancer Control Institute of the National Cancer Center of Korea and by the Ministry of Health and Welfare of Korea.

Address correspondence to Pyo Nyun Kim, MD, Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, 388-1 Poongnap-dong, Songpa-gu, Seoul 138-736, Korea.

E-mail: pnkim@amc.seoul.kr

Abbreviations

AAPM, American Association of Physicists in Medicine; ACR, American College of Radiology; AIUM, American Institute of Ultrasound in Medicine; QA, quality assurance; US, ultrasound

Objectives—The purpose of this study was to evaluate the results of ultrasound (US) device testing using a US phantom and to determine cutoff values for phantom quality assurance tests in US examinations for the screening of hepatocellular carcinoma in Korea.

Methods—Ultrasound phantom images were acquired from the general hospitals in Korea that participated in the National Cancer Screening Program for hepatocellular carcinoma. Ultrasound images of the phantom were acquired with a 3.0- to 5.0-MHz convex transducer and evaluated in terms of the dead zone, vertical and horizontal measurement, axial and lateral resolution, sensitivity, and gray scale/dynamic range. Appropriate cutoff values were determined to guarantee minimal qualifications for the performance of the US scanners.

Results—Three hundred fifty-seven US scanners were tested using the following cutoff values: less than 2 mm for the dead zone, 5% discrepancy in the vertical measurement, 7.5% discrepancy in the horizontal measurement, all 11 identifiable line targets for axial and lateral resolution, more than 14 cm for sensitivity, and more than 4 cylindrical structures for gray scale/dynamic range. With these criteria, 283 US scanners (79.3%) passed the tests. The most common cause of disqualification was the dynamic range/gray scale. No statistical difference was observed in the disqualification rate between 3 groups based on different years of manufacture.

Conclusions—Through this study, we have defined cutoff values for phantom images acquired with US scanners. These will be used in performing screening US examinations for hepatocellular carcinoma in Korea.

Key Words—hepatocellular carcinoma; phantom; quality assurance; ultrasound

The importance of ultrasound (US) image quality assurance (QA) is widely recognized, and recommendations for performing US QA have been made by the major international scientific bodies, including the American Institute of Ultrasound in Medicine (AIUM), American Association of Physicists in Medicine (AAPM), and American College of Radiology (ACR).[1–8] However, a standardized QA test has not yet been solidly established for US imaging, primarily because US examinations are conducted by

highly diverse professional groups for their own purposes, and in most cases, there is no legal regulation system for US such as for ionizing diagnostic imaging modalities. Furthermore, the application of uniform standards is not easy because the technical development of US equipment has been rapid. For example, the ACR standard for monitoring the performance of real-time US equipment[7] relegates the determination of the standards and methods of QA and the analysis of the results to the users. Therefore, it is reasonable to establish separate QA standards for each professional group that performs its own specific examinations.

In Korea, the QA of computed tomography, magnetic resonance imaging, and mammography has been regulated since 2004 by the Korean Institute for Accreditation of Medical Image under the Ministry of Health and Welfare.[9] The goal of this program is to evaluate the image quality in medical examinations for improvement of national health. However, an accreditation program for US has not yet been established.

In Korea, US examinations of the liver for the group at risk for hepatocellular carcinoma, ie, carriers of hepatitis B and hepatitis C viruses and patients with liver cirrhosis, are included in the National Cancer Screening Program.[10] This program is run by the National Cancer Control Institute, which is a part of the National Cancer Center of Korea under the Ministry of Health and Welfare and is funded through taxes. The government of Korea decided to evaluate the quality of US examinations through this tax-funded program. A 3-year survey was planned for the period between 2008 and 2010 for all medical institutes participating in the program. The plan included the evaluation of all general hospitals in 2008, hospitals other than general hospitals in 2009, and private clinics in 2010.

The evaluations for QA of imaging examinations can be divided into 3 categories: personnel evaluation, phantom image evaluation, and clinical image evaluation. However, in the case of US screening examinations for hepatocellular carcinoma, the clinical imaging standards are evaluated according to standard images established by the Korean Society of Radiology and Korean Society of Ultrasound in Medicine,[11] but there was no standard for QA of phantom images. Although an experimental US study was performed in the past using an ATS-539 phantom (ATS Laboratories, Inc, Bridgeport, CT) in Korea, that study cannot be implemented with the wide variety of US scanners that are currently used in different medical institutions. Furthermore, that study was performed before 2004, and it does not contain proper analytic methods that consider international standards such the AIUM, AAPM, and ACR standards.[12] Therefore, applying the standards of that study is not thought to be appropriate. To the best of our knowledge, no similar national project has been undertaken in another country since then, indicating the need to define a new standard. Therefore, we analyzed phantom image data acquired from general hospitals in 2008 and created an evaluation standard for a survey that was conducted over the next 3 years.

The purpose of this study, as a part of a larger project to develop standards for US QA in Korea, was to evaluate the results of the testing of US devices using a standardized phantom and to determine cutoff values of a standard QA test in US examinations of the liver for hepatocellular carcinoma screening.

Materials and Methods

Approval was not required from the Institutional Review Board or Institutional Animal Care and Use Committee because this study did not use any human or animal data. The study only used data from a manufactured phantom.

Ultrasound Phantoms

For our examination, we used an ATS-539 multipurpose phantom, which was specified as the standardized phantom for US images by the Korean Society of Radiology and Korean Society of Ultrasound in Medicine in 2003. This phantom is constructed of rubber-based tissue-mimicking material and is used to evaluate the accuracy and performance of US scanners. The phantom mimics the acoustic properties of human tissue and provides test structures within a simulated environment (Figure 1). The tests performed using this phantom focused on the dead zone, vertical and horizontal measurements, focal zone, sensitivity, axial and lateral resolution, functional resolution, and gray scale/dynamic range.[13]

Acquiring Images From US Scanners

Between April and August 2008, research assistants transported the standard phantom to the general hospitals involved in hepatocellular carcinoma screening in the National Cancer Screening Program and obtained US images using the scanners that were used in the screening program. A 3.0- to 5.0-MHz convex probe and software settings for abdominal US were used in the acquisition. Because we intend to survey real situations, we decided to perform the tests with the same settings used for patients undergoing US scanning. Therefore, we asked the physician on-site to set up the scanner for optimization. The power output, brightness, contrast levels, and time-gain control were controlled and optimized by the research as-

sistants and physicians on-site, We obtained phantom images using the measurement methods described in the manufacturer's manual and AAPM guideline.[1,13] Scanning of the phantom was performed by research assistants in the presence of the physician on-site because we considered the scanning conducted by the research assistants to be superior because of a lack of sufficient understanding of the US phantom by the hospital staff.

Measurement Parameters of the Multipurpose Phantom

The dead zone, vertical and horizontal measurement, axial and lateral resolution, sensitivity, and gray scale/dynamic range were evaluated using the ATS-539 phantom. The focal zone and functional resolution were also measured, but they were excluded from the evaluation because of the difficulty in defining objective standards for these parameters.

Dead Zone

The dead zone is the distance from the front face of the transducer to the first identifiable echo at the phantom or patient interface. No clinical data can be collected in the dead zone. The target group was composed of 9 line targets with the first line target positioned 2 mm below the scan surface. Subsequent targets were spaced 1 mm apart to a depth of 10 mm. We measured the distance from the scan surface to the first identifiable line target. If the first line target of the 9 targets was identifiable, the dead zone was less than 2 mm (Figure 2).

Vertical and Horizontal Measurements

The vertical and horizontal distance measurements were obtained both parallel and perpendicular to the axis of the sound beam. Accurate measurement of the size, depth, and volume of a structure is one of the critical factors in making a proper diagnosis. We measured 10.0 cm along the axis of the sound beam for the vertical measurement (from a 1.0-cm-deep line target to an 11.0-cm-deep line target) and 8.0 cm perpendicular to the sound beam for the horizontal measurement, and the resulting measurements were compared with the actual distance between the line targets in the phantom using the US scanner's calipers (Figure 3). The focus was at the depth of the horizontal targets, and we made sure to use as little pressure as possible when applying the transducer to the scanning membrane to avoid displacement of the line targets in the phantom. For vertical measurement, the caliper markers were placed at the top of the echo from line target, and for horizontal measurement, we placed the caliper markers above the centers of the echoes from the line targets

Axial and Lateral Resolution

Resolution is defined as the minimum reflector separation between two closely spaced objects that can be imaged separately. If a system has poor resolution, small structures lying close to each other will appear as a single structure. The axial resolution is dependent on the pulsing system of the imaging device and the condition of the transducer, whereas the lateral resolution is affected by the beam width.

Figure 1. Target diagram of the standardized phantom. The ATS-539 multipurpose phantom (ATS Laboratories, Inc, Bridgeport, CT) has 4 scanning surfaces and many internal structures with which various measurements can be performed.

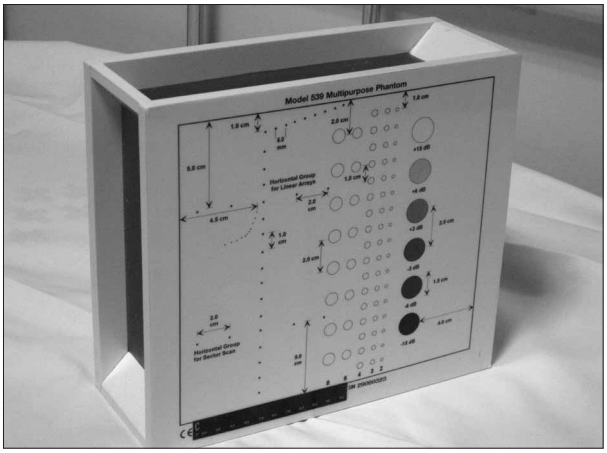


Figure 2. Dead zone. Nine line targets are positioned between 2 and 10 mm below the scan surface. In this image, all 9 line targets are clearly visualized (arrowheads). The distance between the scan surface and first line target is the dead zone. In this case, the dead zone is 2 mm (arrow). The focus is located as near as possible.

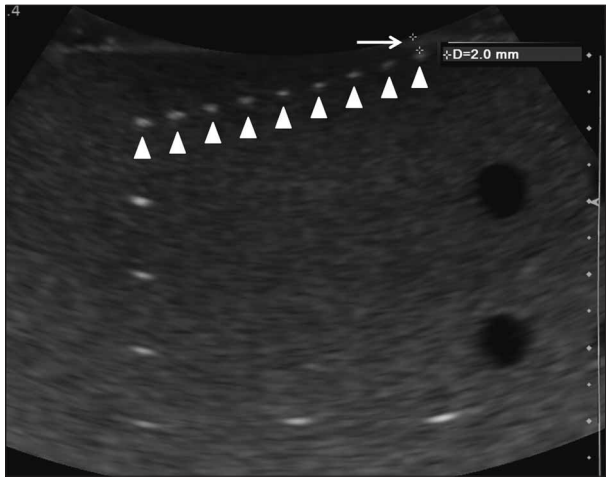
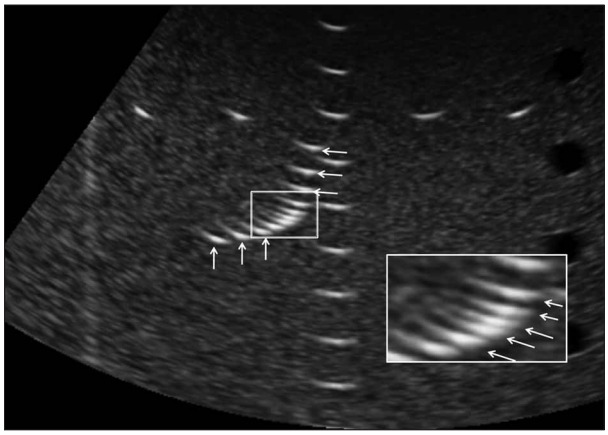


Figure 3. Vertical and horizontal measurements. A 10-cm distance along the ultrasound beam axis (arrows) and an 8-cm distance perpendicular to the ultrasound beam axis (arrowheads) are measured. Measurement should be done at the center and top of the each line target. In this ultrasound scanner, the vertical measurement is 10.04 cm, and the horizontal measurement is 8.28 cm. The discrepancies of the vertical and horizontal measurements are 0.4% and 3.5%, respectively.

The line targets in the phantom were spaced at 5.0-, 4.0-, 3.0-, 2.0-, and 1.0-mm intervals, both axially and laterally. Eleven line targets were present in the phantom, and we counted the number of line targets that were identifiable separately (Figure 4). The focus was located at the target group, and image zooming was applied.

Figure 4. Axial and lateral resolution. Eleven line targets with a curved array are clearly visible separately. The distances between the line targets are from 1 mm in the central area to 5 mm in the peripheral area. The curved array of line targets is used for the test of the axial and lateral resolution. In this image, all 11 line targets are visualized separately and clearly (arrows). The central part of the line targets is zoomed in the inset.



Sensitivity

Sensitivity, which is a test of the penetration depth of the US beam, refers to the ability to image small objects located at specified depths. Anechoic 8-mm round structures were located in the phantom along the direction of the US beam. The distance between the structures was 2.0 cm. We recorded the deepest target structure that was displayed on the US images; ie, if the eighth structure was visible and appeared round, the sensitivity was more than 16 cm, and if the sixth structure was the deepest visible structure, then the sensitivity was 12 cm (Figure 5). The focus was located as deep as possible.

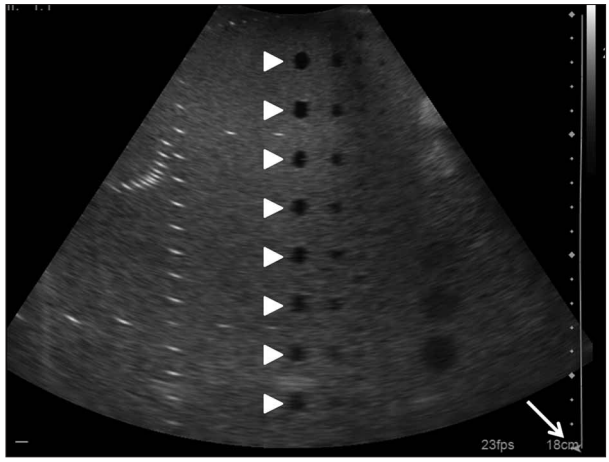
Gray Scale/Dynamic Range

The gray scale/dynamic range, which is a test of the contrast on US images, uses the amplitude of the received echoes to vary the degree of brightness in the displayed image. Six cylindrical targets with varying degrees of brightness were visible on the US images. These targets appeared circular in the US image plane. The contrast values of these targets relative to the background material were +15, +6, +3, –3, –6, and –15 dB. We counted the number of cylindrical targets that appeared as discrete round structures through more than 180° (Figure 6).

Overall Image Analyses and Establishment of Cutoff Values

The US phantom images were reviewed by 2 (of 7) experienced abdominal radiologists, each with more than 5 years of US experience. Before evaluation, the readers

Figure 5. Sensitivity. Eight-millimeter anechoic round structures are well visualized. Eight structures are clearly visible as round structures, and the sensitivity is greater than 16 cm in this case (arrowheads). This test is for the penetration depth of the ultrasound beam. The focus is located as deeply as possible (arrow), and the depth is set to 18 cm.

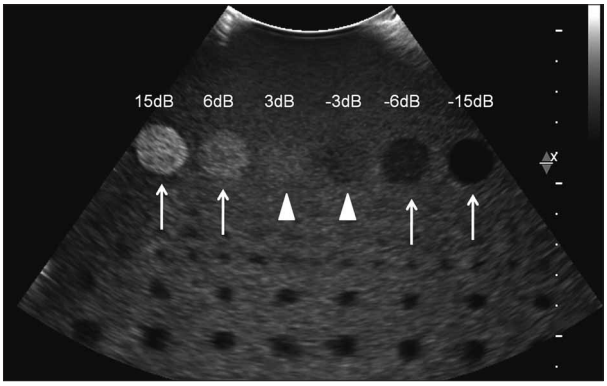


received 2 hours of training on US phantom image interpretation. The objective of the QA testing was not to identify the best equipment but to identify defective equipment. Therefore, setting cutoff values too high was inappropriate, and we selected cutoff values for each test according to the following criteria: (1) the criteria recommended by the phantom manufacturer's manual and the major international scientific bodies, such as the AIUM, AAPM, and ACR; and (2) the highest cutoff values that allowed at least 90% of the scanners to pass the QA testing.

Using the above criteria, we checked the overall test results for all US scanners and analyzed the causes of disqualification. We also examined whether the years of manufacture of the US scanners affected the disqualification rates and whether we could simplify the test items without having to use all 6 of the aforementioned items.

Interobserver agreement between the 2 observers was assessed by weighted κ statistics. The κ values were interpreted as follows: 0.80 to 1.00, excellent agreement; 0.61 to 0.80, good agreement; 0.41 to 0.60, moderate agreement; 0.21 to 0.40, fair agreement; and 0.00 to 0.20, poor agreement. In addition, we calculated the agreement rate, which means the rate of agreement of "pass or fail" between the observers with the decided cutoff values. To compare the mean discrepancies of the vertical and horizontal measurements, an unpaired *t* test was performed, and to compare the disqualifying rates of different groups by years of manufacture, a Fisher exact test was performed. All statistical analyses were performed using commercial statistical software (MedCalc version 9.2 for Windows; MedCalc Software, Mariakerke, Belgium), and all of the charts were created using Excel 2007 software (Microsoft Corporation, Redmond, WA). Differences were considered significant at $P < .05$.

Figure 6. Gray scale/dynamic range. Four or more of 6 cylindrical structures should be clearly visible over 180° for passing the gray scale/dynamic range test. The contrast values of these targets relative to background material are +15, +6, +3, –3, –6, and –15 dB. In this case, 4 cylindrical structures (+15, +6, –6, and +15 dB) are clearly visible as circular structures (arrows). However, 2 cylindrical structures with contrast values +3 and –3 dB are not clearly visible as circular structures (arrowheads).



Results

A total of 357 US scanners were surveyed from 271 general hospitals. A total of 70 different models of US scanners were included in the study. The years of manufacture ranged between 2006 and 2008 for 145 scanners (40.6%), between 2003 and 2005 for 112 (31.4%), between 2000 and 2002 for 47 (13.2%), before 2000 for 22 (6.2%), and unknown for 31 (8.7%). Digital Imaging Communications in Medicine data were acquired for 339 scanners (95.0%), thermal papers for 14 (3.9%), and films for 4 (1.1%).

Dead Zone

The allowable dead zone range recommended by the AAPM US task group is less than 5 mm in the case of a 3.0- to 7.0-MHz probe.[1] The dead zone was less than 2 mm in 351 scanners (98.3%), 3 mm in 4 (1.1%), 4 mm in 1 (0.003%), and 5 mm in 1 (0.003%). Therefore, we set the dead zone cutoff value as less than 2 mm. The weighted κ value of the dead zone for the observers was 0.855, indicating excellent agreement. In addition, when we evaluated the interobserver agreement of qualification or disqualification with the cutoff value of less than 2 mm, the κ value was 0.747, indicating good agreement, and the agreement rate was 98.9% (353 of 357).

Vertical and Horizontal Measurements

According to the AAPM US task group, the allowed discrepancies in the vertical and horizontal measurements are 1.5% and 2.0%, respectively.[1] However, with these cutoff values, 77.6% and 95.0% of US scanners would fail to qualify in our survey. Therefore, we adopted more generous cutoff values of 5.0% (ie, 5 mm) for vertical measurements and 7.5% (ie, 6 mm) for horizontal measurements. With these cutoff values, 354 scanners (99.2%) could be qualified for vertical measurement, and 344 (96.4%) could be qualified for horizontal measurement. When converted to percentages, the mean discrepancy values ± SD were 1.02% ± 1.00% for the vertical measurements and 5.16% ± 1.81% for the horizontal measurements. The mean discrepancy in the horizontal measurements was significantly larger than the discrepancy in the vertical measurements ($P < .001$). The distribution of the discrepancies in the vertical and horizontal measurements are illustrated in Figure 7.

Axial and Lateral Resolution

In 346 scanners (97.0%), all 11 line targets were identifiable separately. Therefore, we selected all 11 identifiable line targets as the cutoff value for our test. In 8 scanners (2.2%), only 10 line targets were identifiable separately, and in 3 (0.08%), only 9 targets were identifiable. The weighted κ value of the axial and lateral resolution for the observers was 0.629, indicating good agreement. In addition, when we evaluated the interobserver agreement of qualification or disqualification with the cutoff value of all 11 targets, the κ value was 0.659, indicating good agreement, and the agreement rate was 98.0% (350 of 357).

Sensitivity

With a 14-cm cutoff value, 340 scanners (95.2%) could be qualified in this test, whereas with more than 16 cm, only 285 (79.3%) passed the examination. Therefore, we selected 14 cm as the cutoff value. The weighted κ value of the sensitivity for the observers was 0.704, indicating good agreement. Also, when we evaluate the interobserver agreement of qualification or disqualification with the cutoff value of 14 cm, the κ value was 0.660, indicating good agreement, and the agreement rate was 98.6% (352 of 357).

Gray Scale/Dynamic Range

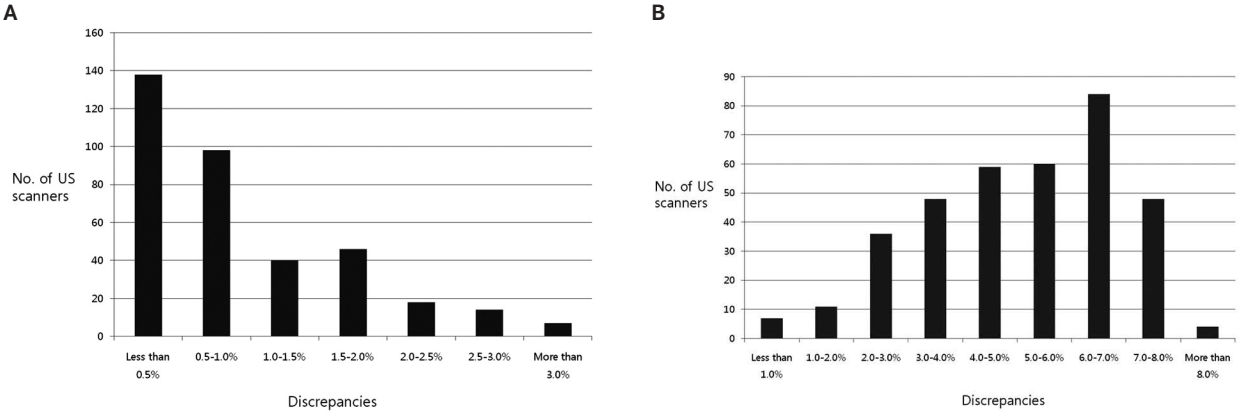
With a cutoff value of more than 4 cylindrical structures, 325 scanners (91.0%) could be qualified, whereas only 118 (33.1%) could be qualified with a cutoff value of more than 5 cylindrical structures, and 352 (98.6%) could be qualified with a cutoff value of more than 3 cylindrical structures. Given our initial condition that more than 90% of the scanners must pass the test, we selected a cutoff value of more than 4 cylindrical structures. The distribution of the results of the gray scale/dynamic range test is illustrated in Figure 8. The weighted κ value of the gray scale/dynamic range for the observers was 0.652, indicating good agreement. Also, when we evaluated the interobserver agreement of qualification or disqualification with the cutoff value of more than 4 cylindrical structures, the κ value was 0.969, indicating excellent agreement, and the agreement rate was 98.6% (352 of 357).

Overall Evaluation With the Decided Cutoff Values

In this study, 283 US scanners (79.3%) passed the QA test with the following cutoff values: less than 2 mm for the dead zone, 5% discrepancy for vertical measurements, 7.5% discrepancy for horizontal measurements, all 11 identifiable line targets for axial and lateral resolution, more than 14 cm for sensitivity, and more than 4 cylindrical structures for the gray scale/dynamic range. The causes of disqualification are summarized in Table 1. The most common cause of disqualification was the gray scale/dynamic range, which was responsible for 44.6% of the disqualifications. The disqualification rates with respect to the years of manufacture are summarized in Table 2, and no significant differences were observed in the disqualification rates between the 3 groups based on the years of manufacture.

We attempted to simplify these results to implement the cutoff values as references standards. We included only 3 important factors: resolution (axial and lateral resolution), penetration depth (sensitivity), and contrast (dynamic range/gray scale). The sensitivity and specificity of the combination of only these 3 factors for equipment qualification were 75.7% and 93.6%, respectively. With these simplified criteria, 301 US scanners passed the QA test, for a qualification rate of 84.3%.

Figure 7. Distribution of the discrepancies in the vertical and horizontal measurements. **A**, Distribution of the discrepancies in the vertical measurements. Most ultrasound (US) scanners have discrepancies of less than 1 mm (1%). **B**, Distribution of the discrepancies in the horizontal measurements. Compared with that of the vertical measurements, the distribution is more widely dispersed.



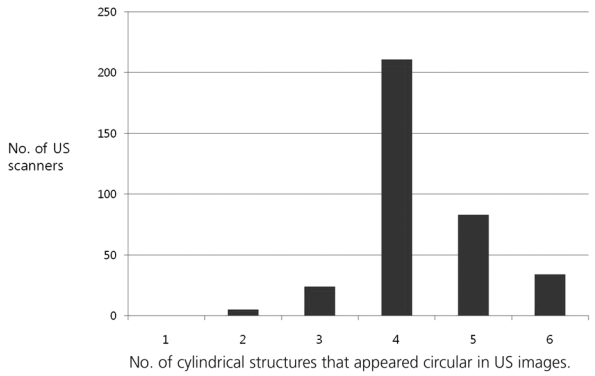


Figure 8. Distribution of the gray scale/dynamic range. On most ultrasound (US) scanners, 4 cylindrical structures are visualized as circular structures on phantom images.

Discussion

Although US examinations are conducted widely, the overall interest in US QA is considerably lower compared with computed tomography and magnetic resonance imaging, perhaps because various types of US examinations are conducted by a wide variety of professionals, and US usage varies widely, ranging from simple bedside procedures to complex examinations using high-tech equipment. Thus, the application of general standards for the overall QA of US examinations is difficult. Therefore, implementation of a national QA program with standards related to specific examinations might be relatively advantageous over QA standards related to US examinations overall. We have designed an intensive QA test for the US examination used in the cancer-screening program administered by the Korean government. To the best of our knowledge, a study focusing on such work has not been reported previously. The findings of our research are important and should be shared with other countries.

Ultrasound examinations are commonly recommended in the United States, Europe, and Japan as screening tests for patients at risk for hepatocellular carcinoma.[14–16] In Korea, US examinations are also included in the National Cancer Screening Program as screening tests for patients at risk for hepatocellular carcinoma, and the need for QA of US examinations has been emphasized.

A QA program for imaging examinations is generally divided into 3 parts: personnel evaluation, device evaluation, and imaging protocol evaluation. The evaluation of the imaging device is performed primarily through phantom imaging. In US examinations, image-based performance measurements must be made with a US phantom. Both subjective visual methods and objective computer-based approaches may be used to make these measurements.[7] Other approaches for performance measurement that do not require US images of phantoms have been reported; the FirstCall aPerio system (Sonora Medical Systems, Inc, Longmont, CO) can test the electrical and acoustic characteristics of each individual transducer array element without the phantom.[8]

Because of subjectivity, manual measurement and visual assessment of phantom images are known to be less accurate than computerized automated measurements.[3–5] However, considering the large number of scanners to be included in our future survey (>3000 US scanners in numerous hospitals and clinics) and the different formats for storing data (many private clinics use thermal paper or film), subjective visual assessment had to be accepted in this study. To overcome the subjectivity stemming from manual measurement and visual assessment, the results were accepted only if both readers reached a consensus. The cutoff values used in this study were generous for two reasons: (1) this QA test was designed to filter out severely defective equipment; and (2) the values were chosen to account for the subjectivity stemming from visual assessments. The interobserver agreement of our study was ro-

Table 1. Causes of Disqualification

Failed Tests	Scanners, n (%)
Failed 1 test	
Dead zone	6 (8.1)
Vertical measurement	3 (4.1)
Horizontal measurement	9 (12.2)
Axial and lateral resolution	9 (12.2)
Sensitivity	13 (17.6)
Gray scale/dynamic range	25 (33.8)
Failed 2 tests	
Horizontal measurement and gray scale/dynamic range	4 (5.4)
Sensitivity and axial and lateral resolution	1 (1.4)
Sensitivity and gray scale/dynamic range	3 (4.1)
Gray scale/dynamic range and axial and lateral resolution	1 (1.4)
Overall	74 (100)

Failed 1 test indicates that the ultrasound scanners failed only 1 of 6 quality assurance tests; and Failed 2 tests indicates that the scanners failed 2 of 6 tests.

Table 2. Disqualification Rates by Years of Manufacture

Years of Manufacture	Disqualification Rate, n (%)	*P*
2006–2008	30/145 (20.7)	
2003–2005	24/112 (21.4)	.8788
Before 2002	12/69 (17.4)	.7130

P values indicate comparisons with the disqualification rate in 2006 to 2008 (Fisher exact test).

bust. However, the κ values were not as high considering the nearly perfect agreement rates for qualification and disqualification. These results are caused by the paradoxes of κ statistics due to prevalence effects.[17,18] High agreement but low κ values can be observed when marginal totals are highly symmetrically unbalanced; that is, the number of qualified scanners was much larger than that of the disqualified scanners in our study. Therefore, our κ values were substantially underestimated, and the degree of agreement in our study was much greater than the κ values themselves. In our study, intraobserver agreement was not assessed; however, considering robust results for interobserver agreement, we believe that intraobserver agreement might not have caused a problem.

The dead zone occurs because the imaging system cannot simultaneously send and receive data. Artifacts in this zone are caused by fluctuations in the received data. In the case of the dead zone, the allowable range recommended by the AAPM US task group is less than 5 mm for a 3.0- to 7.0-MHz probe.[1] In fact, for US scans of the liver, a dead zone of approximately 5 mm does not considerably affect the quality of the test. This study did not find any scanner with a dead zone exceeding 5 mm. Therefore, for scanners that are dedicated to liver imaging, the dead zone measurement will not have a notable impact.

The AAPM US task group allows an error of 1.5% for vertical measurements and 2% for horizontal measurements.[1] However, these values were too strict for the manual measurements in our survey, and the manufacturer's manual suggests a 10% discrepancy.[13] Vertical and horizontal measurements of the distance based on manual placement of electrical caliper markers can be very subjective, and the accuracy can be affected by factors such as the pixel resolution of the image. In many systems, the accuracy of the measurement can be improved by zooming in and increasing the pixel resolution; however, that capability was not the case in our study because the field of view for a 10-cm depth and 8-cm width was too large to zoom in. Also, we believe that our well-trained research assistants did their best to measure the exact distance. Distance measurement should be done by the electrical calipers of US machine itself, and postscanning analysis of the distance is not permitted.[1,13] Therefore, to evaluate interobserver and intraobserver agreement, multiple measurements should be done when images are acquired, but that procedure was not the case in our study. Furthermore, more than 80% of US scanners use thermal paper in Korean private clinics, and in those cases, postscanning analysts of the distance is actually impossible because of the possibility of geometric distortion associated with the hard copy device.

Well-known measurement errors include a temperature-dependent velocity change in the sound beam within the phantom, a distortion in the phantom geometry due to excessive pressure on the phantom during the measurement, and measuring obliquely instead of perpendicularly on the beam axis during horizontal measurement.[1] The accuracy of the vertical distance measurement is dependent on the integrity of the internal timing circuitry of the imaging system, and the accuracy of the horizontal measurement is determined by the integrity of the transducer, the output intensity, and the resolution of the imaging system.[1,13] An error is more likely in the horizontal measurement than in the vertical measurement, consistent with our results.

In our study, no significant differences were observed for the disqualification rates based on the years in which the US scanners were manufactured. This result was somewhat surprising and may have been attributable to a selection bias that occurred when old equipment with poor performance was discarded. Our result might also indicate that US scanners last for long durations if maintained properly.

As stated earlier, we considered simplified criteria involving only the most important elements of the US examination for hepatocellular carcinoma screening, ie, axial and lateral resolution, sensitivity, and gray scale/dynamic range. Given that the importance of the dead zone in US examinations for hepatocellular carcinoma screening is unclear and that the probability of error in manual measurements of vertical and horizontal distances due to subjectivity is high, these parameters were not considered in the simplified criteria, which led to a 5% improvement in the disqualification rate.

This study had several limitations. First, the criteria for cutoff value selection were subjective. We selected the highest cutoff values that allowed at least 90% of the scanners to pass the QA testing, and this condition was quite subjective. However, there were no standardized cutoff values for US phantom evaluation by major international scientific bodies, only guidelines for maintenance, which were not applicable to QA tests deciding qualification or disqualification. Furthermore, because we wanted to consider real practice situations, we could not disqualify too many scanners. Second, there is insufficient evidence that phantom image quality is directly related to patient image quality for hepatocellular carcinoma screening. It would have been more meaningful to see whether there was any correlation between US scanners that were disqualified and the ability to detect liver lesions. However, this study was a survey of US scanners only, and we did not have a "standardized patient" with hepatocellular carcinoma; there-

fore, that kind of analysis could not be performed. However, if phantom images are poor, then the resolution and contrast of patient images are expected to deteriorate, and we believe that the QA tests for measurements, penetration depth, resolution, and contrast are essential factors affecting US image quality. Third, although the US scanning was optimized and conducted by the physician on-site and qualified research assistants, it is uncertain whether the images obtained were of the best quality. However, because we would like to survey real situations, we decided to perform the QA tests with the same settings used for patients undergoing US scanning. Fourth, the use of a single type of phantom for all devices made the results susceptible to shortcomings of the phantom itself. However, this susceptibility would extend uniformly to all of the US scanners included in the study, which seems fair. Fifth, defects in transducer crystals were not included in the phantom image evaluation due to the difficulty of identifying fine crystal damage through visual inspection. However, no US scanner in this study showed a crystal defect that was identifiable by visual assessment. Finally, this study was based on data from general hospitals, which are expected to have satisfactory QA results. Hence, the data may differ substantially from those of private clinics, which is the subject of validation in the upcoming survey.

In conclusion, we have defined standards for US scanner phantom images that will be used in performing screening examinations for hepatocellular carcinoma in Korea. This is the first step toward establishing a standardized US QA test, and validation of these standards should be performed in an upcoming nationwide survey of all US scanners in Korea.

References

1. 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.
2. American Institute of Ultrasound in Medicine. *Quality Assurance Manual for Gray Scale Ultrasound Scanners: Stage 2*. Laurel, MD: American Institute of Ultrasound in Medicine; 1995.
3. 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.
4. Gibson NM, Dudley NJ, Griffith K. A computerised quality control testing system for B-mode ultrasound. *Ultrasound Med Biol* 2001; 27:1697–1711.
5. Browne JE, Watson AJ, Gibson NM, Dudley NJ, Elliott AT. Objective measurements of image quality. *Ultrasound Med Biol* 2004; 30:229–237.
6. Thijssen JM, Weijers, G, de Korte CL. Objective performance testing and quality assurance of medical ultrasound equipment. *Ultrasound Med Biol* 2007; 33:460–471.
7. American College of Radiology. ACR technical standard for diagnostic medical physics performance monitoring of real time ultrasound equipment. American College of Radiology website; 2009. http://www.acr.org/SecondaryMainMenuCategories/quality_safety/guidelines/med_phys/us_equipment.aspx
8. Sipilä O, Mannila V, Vartiainen E. Quality assurance in diagnostic ultrasound [published online ahead of print December 6, 2010]. *Eur J Radiol*. doi:10.1016/j.ejrad.2010.11.015 .
9. Kim YS, Jung SE, Choi BG, et al. Image quality improvement after implementation of a CT accreditation program. *Korean J Radiol* 2010; 11:553–559.
10. Yoo KY. Cancer control activities in the Republic of Korea. *Jpn J Clin Oncol* 2008; 38:327–333.
11. Kim PN, Kim KW, Byun JH. Quality assessment of hepatic ultrasound images examined after a medical checkup. *J Korean Soc Ultrasound Med* 2009; 28:31–37.
12. Kim PN, Lim JW, Kim HC, et al. Quality assessment of ultrasonographic equipment using an ATS-539 multipurpose phantom. *J Korean Radiol Soc* 2008; 58:533–541.
13. ATS Laboratories, Inc. *Tests Performed*. Bridgeport, CT: ATS Laboratories, Inc; 2000.
14. Bruix J, Sherman M. Management of hepatocellular carcinoma: an update. *Hepatology* 2011; 53:1020–1022.
15. Bruix J, Sherman M, Llovet J, et al; EASL Panel of Experts on HCC. Clinical management of hepatocellular carcinoma: conclusions of the Barcelona 2000 EASL conference. European Association for the Study of the Liver. *J Hepatol* 2001; 35:421–430.
16. Kudo M, Okanoue T; Japan Society of Hepatology. Management of hepatocellular carcinoma in Japan: consensus-based clinical practice manual proposed by the Japan society of hepatology. *Oncology* 2007; 72(suppl 1):2–15.
17. Feinstein AR, Cicchetti DV. High agreement but low kappa, I: the problems of two paradoxes. *J Clin Epidemiol* 1990; 43:543-549.
18. Cicchetti DV, Feinstein AR. High agreement but low kappa, II: resolving the paradoxes. *J Clin Epidemiol* 1990: 43:551–558.