Do Radiologists Agree on Findings in Radiographer-Acquired Sonographic Examinations?

Carina Stenman, LicMedSci, Shazia Jamil, MD, Lars Thorelius, MD, Anders Knutsson, MD, Örjan Smedby, DrMedSci

Objectives—Sonographic examinations are usually regarded as observer dependent, but a recently introduced method using documentation with cine loops acquired in a standardized way attempts to address this problem. The aim of this study was to evaluate the intraobserver and interobserver agreement of sonographic liver examinations using strictly standardized examination protocols with cine loop documentation.

Methods—Ninety-eight outpatients were examined by a radiographer using the standardized method. Three radiologists, each with 10 to 20 years of experience in sonography, reviewed the cine loops retrospectively. After 4 weeks, the review was repeated; the 3 radiologists were blinded to the initial reading. The κ coefficient was used to analyze intraobserver and interobserver agreement, and agreement in percent was also calculated.

Results—The intraobserver agreement was highest for concrements in the gallbladder (κ= 0.91–0.96) and lowest when assessing the need for further examination (κ = 0.38–0.64). For increased liver echogenicity, κ varied between 0.73 and 0.92 and for skip areas between 0.73 and 0.90. The interobserver agreement was also highest for concrements in the gallbladder (κ = 0.84–1.00) and lowest for the need for further examination (κ = –0.12–0.46). For most other findings, substantial intraobserver agreement was found.

Conclusions—For sonographic examinations performed according to a standardized examination protocol by a radiographer and viewed by an experienced radiologist, good interobserver agreement was found, except for judgments of the need for further examinations.

Key Words—agreement; liver examination; observer; radiologist; sonography

Sonography is often a first choice for radiologic examinations. It is a quick, inexpensive, noninvasive, and safe diagnostic method, which spares the patient unnecessary ionizing radiation that would be used in alternative imaging methods.[1,2] Due to recent technological advances in image quality, sonography is often used to assess liver disease, including diagnosing focal changes and steatosis, a condition that may progress to fibrosis and cirrhosis.[3] Sonographic evaluation of the echogenicity of the liver in suspected steatosis is usually based on a subjective impression.[3,4] Sonography is also used as the first examination of the gallbladder, and it has high accuracy for the diagnoses of, for example, gallstones and wall thickening.[5,6] In general, sonographic findings are presently regarded as subjective and highly dependent on the skill of the individual examiner.[7–9]

Received May 9, 2012, from the Division of Radiological Sciences, Department of Medical and Health Sciences (C.S., Ö.S.), Department of Radiology (C.S., S.J., A.K.), and Center for Medical Image Science and Visualization (Ö.S.), Linköping University, Linköping, Sweden; and Department of Radiology, Copenhagen University Hospital, Herlev, Denmark (L.T.). Revision requested June 4, 2012. Revised manuscript accepted for publication August 10, 2012.

Address correspondence to Carina Stenman, LicMedSci, Division of Radiological Sciences, Department of Medical and Health Sciences, Linköping University, SE-58183 Linköping, Sweden.

E-mail: carina.stenman@liu.se

Currently, the most common way to document sonographic examinations is to store still images. This method offers very limited possibilities for reevaluation of a sonographic examination. Especially when new clinical questions arise after the examination, reevaluation is often not helpful.[8,10]

A possible solution that we have presented in an earlier publication[11] is to use a method, with standardized examination protocols, in which documentation of the examination is made with cine loops, ie, films in which the scan covers 5 to 10 cm in 5 to 10 seconds depending on the target organ. The documentation should include both longitudinal and transverse views covering the whole organ or region of interest. After the examination, the cine loops are stored in the picture archiving and communication system, and evaluation is made later at dedicated workstations, where the examination can be viewed once, for writing the report, or many times.[11] Another advantage is that examinations performed by a radiographer can be evaluated later by a radiologist, thus increasing the availability of the radiologist for more complex examinations.[12] The aim of this study was to evaluate the intraobserver and interobserver agreement of sonographic liver examinations using strictly standardized examination protocols with cine loop documentation.

Materials and Methods

A retrospective review was conducted of 98 sonographic examinations of the liver and biliary system between January 2006 and 2008 from our center. The local Ethics Committee waived the need for a committee review in this type of retrospective study. Included in the study were patients with upper abdominal pain who were referred to the Department of Radiology for a sonographic examination of the abdomen and were considered by a radiologist as suitable for examination by a radiographer. Examinations suitable for a radiographer to perform are those for patients with a clear clinical question. Before accepting a referral for examination, one of the radiologists reviewed the history of the patient to determine whether an examination was medically indicated and whether the patient was expected to be suitable for the examination by a radiographer. All patients had symptoms such as pain or occasional cramps in the upper right abdomen, with or without elevated liver enzyme levels. The clinical questions included concrements (calculi) in the gallbladder, polyps in dilated bile ducts, focal changes, echogenicity, and liver size.

All sonograms were obtained with an Acuson Sequoia system (Siemens Medical Systems, Erlangen, Germany) using convex transducers (6C2 with 4–5 MHz in thin patients and 4C1 with 3–4 MHz, in normal-sized and thick patients). The examinations were optimized on a case-by-case basis, which included different technical parameters such as gain adjustment and focal zone locations. All data were acquired by the same radiographer, using the standardized examination protocol for the liver and biliary system, consisting of 11 sweeps. With the use of the standardized examination protocol, it is possible to cover the whole liver, ie, all 8 segments. The goal of the examination is coverage of the entire liver and definition of the segment where any abnormality is located. All scanning of transverse scan planes is done craniocaudally, and longitudinal planes are scanned from left to right. Together, the transverse scans should cover the entire liver in the craniocaudal direction with some margin. The radiographer had been trained specifically to acquire images according to the standardized examination protocol. All of the dynamic films were stored in the dedicated picture archiving and communication system (Syngo Dynamics; Siemens Medical Systems).

The examinations were reviewed by 3 radiologists, all of whom had 10 to 20 years of experience in abdominal sonography. Two of the radiologists (A and B) were very familiar with standardized examination protocols and documentation with cine loops. They had worked with the method since 2002, whereas the third (C) was introduced to the technique 2 weeks before the start of the study. At the review, the radiologists filled out an evaluation form including several possible pathologic findings in the liver parenchyma (Table 1): echogenicity with 4 levels (normal, slightly increased, moderately increased, or greatly increased), skip areas (areas unaffected by steatosis),[13] focal changes, increased gallbladder wall thickness, concrements in the gallbladder, polyps in the gallbladder, biliary dilatation, and whether there was a need for further examination with computed tomography or magnetic resonance imaging. There was also a possibility to indicate findings not covered by these categories by filling out a field for "other findings." The echogenicity of the liver was assessed by comparison to that of the right kidney, and beam penetration and the conspicuity of vessels were taken into consideration when characterizing fatty infiltration. When it was judged that there was a need for further examination, it could have been for a number of reasons, including that the patient had steatosis, which complicates the diagnosis. Further examination was also recommended by the radiologist when the patient was overweight, which could limit the patient's ability to take a deep breath, so that some subdiaphragmatic parts of the liver were difficult to visualize. The films were reviewed by all observers on the same monitor, and they were blinded to the clinical data including patient history and to each other's assessment. After an interval of 4 weeks, the cases were randomly rearranged in a new sequence. The review was repeated, using the same evaluation protocol, and the 3 radiologists were blinded to the initial reading.

The protocols were compared for intraobserver and interobserver agreement. The κ coefficient was used to assess the agreement, but since κ may be less reliable in cases when the prevalence of a response is very high or low, it was supplemented with agreement expressed as a percentage.[14,15] For the increased echogenicity in the liver, the weighted κ was used, regarding the 4 different levels of increased echogenicity. Values for the κ statistic were labeled as follows: less than 0.00, poor agreement; 0.00 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.00, almost perfect agreement.[16] In cases with fewer than 3 observations in either category, κ or percent agreement was not calculated.[17] The Friedman test was used to determine whether there was a significant difference between the 3 radiologists' observations. In cases in which the Friedman test indicated a significant difference, the Conover test was applied for pair-wise comparisons between reviewers and review occasions. Calculations were made with BrightStat.[18]

Results

Of the 98 patients, 60 were women and 38 men. The age range was 18 to 99 years (median, 56 years). The numbers of examinations with pathologic findings are summarized in Table 1. The most common finding was increased echogenicity, occurring in 38 to 42 of the 98 patients. Bile duct dilatation, on the other hand, was found only in 1 to 3 of the patients, depending on the observer and review occasion.

Intraobserver Agreement

Intraobserver agreement for the different types of liver pathologic findings is summarized in Table 2. Concrements in the gallbladder showed almost perfect agreement.

Table 1. Number of Examinations With Pathologic Findings (n = 98)

		Radiologist A		Radiologist B		Radiologist C		
Characteristic	Finding	1st Review	2nd Review	1st Review	2nd Review	1st Review	2nd Review	Friedman P
Echogenicity	0	56	59	58	58	55	55	.84
	1	16	12	10	11	14	13	
	2	12	15	14	9	13	15	
	3	14	12	15	18	14	13	
Skip areas	No	78	77	75	67	79	82	<.001
	Yes	20	21[a,b]	22	29[a]	17	14[a]	
Abnormal parenchymal appearance	No	97	98	94	93	96	96	.06
	Yes	1	0	3	3	0	0	
Focal changes	None	85	82	71	70	79	80	NA
	Cyst	10[c]	12[c]	18[c]	20[a]	12	11[c]	.006[d]
	Other	3	4	7	6	5	5	.33[d]
Liver size	Small	2	0	1	3	0	0	<.001
	Normal	96	98	95	93	88	87	
	Large	0[c]	0[c]	0[c]	0[c]	8[a]	9[a]	
Increased gallbladder wall thickness	No	83	82	80	79	83	84	<.001
	Yes	1[c]	2[c]	5[a]	7[a]	1[c]	0[c]	
Concrements in gallbladder	No	67	70	67	67	67	68	.16
	Yes	17	14	18	19	17	16	
Polyps in gallbladder	No	76	76	80	80	75	70	<.001
	Yes	8	8[c]	5	5[c]	9	14[a,b]	
Bile duct dilatation	No	95	95	97	96	95	97	.15
	Yes	1	3	1	2	1	1	
Need for further examination	No	89	93	80	78	90	91	<.01
	Yes	9	5[c]	18[c]	18[a]	8[c]	7[c]	

[a]$P < .05$ compared to both other reviewers on the same occasion.
[b]$P < .05$ compared to the first occasion for the same reviewer.
[c]$P < .05$ compared to one of the other reviewers on the same occasion.
[d]The Friedman test compared each of the cyst and other findings separately to the none category.

For polyps in the gallbladder, the intraobserver agreement varied between moderate and almost perfect. Lower agreement was found for the need for further examination, where the values ranged from fair to substantial. For skip areas and focal changes, the agreement was substantial or almost perfect.

Interobserver Agreement

The interobserver agreement within each pair of observers (A and B, A and C, and B and C) for different types of pathologic findings in the liver is given in Table 3. Almost perfect agreement was again seen for concrements in the gallbladder and mostly poor agreement for the need for further examination. For the other findings, the agreement was moderate to almost perfect, with a somewhat lower value for focal changes. No tendency toward higher agreement in the second reading was found. In general, interobserver agreement was slightly lower than intraobserver agreement.

Table 2. Intraobserver Agreement Between 2 Readings 4 Weeks Apart

	κ Value (% Agreement)		
Characteristic	**A1 vs A2**	**B1 vs B2**	**C1 vs C2**
Increased echogenicity[a]	0.92[b] (97)	0.89[b] (95)	0.73[b] (89)
Skip areas	0.90[b] (97)	0.81[b] (93)	0.73[b] (93)
Focal changes	0.76 (94)	0.67 (88)	0.89 (97)
Concrements in gallbladder	0.91[b] (97)	0.95[b] (99)	0.96[b] (99)
Polyps in gallbladder	0.86[b] (97)	0.57[b] (95)	0.65[b] (92)
Need for further examination	0.38 (92)	0.57[b] (87)	0.64[b] (95)

[a]For increased echogenicity, the weighted κ was used.
[b]$P < .05$

Discussion

The aim of this study was to evaluate the intraobserver and interobserver agreement of sonographic liver examinations using strictly standardized examination protocols with cine loop documentation. It turned out that for the probably easiest finding, concrements in the gallbladder, there was almost perfect agreement within and between observers. Most other findings (second, third, and fifth rows in Tables 2 and 3) showed (with 2 exceptions) moderate to almost perfect agreement. For echogenicity of the liver parenchyma, for which the weighted κ was used, the agreement ranged from substantial to almost perfect. For many of the findings, the intraobserver agreement, not surprisingly, was slightly higher than the interobserver agreement. The frequency figures for the assessed need for further examination (bottom row of Table 1) indicate that our observers judged the diagnostic quality as adequate in 82% to 95% of the examinations, but the disagreement between observers on this point is notable.

Figure 1 illustrates the variability in assessed echogenicity between the 3 observers. In one of the cases, radiologists A and B rated the echogenicity in the liver as greatly increased, whereas radiologist C assessed the same examination as moderately increased echogenicity. Radiologists A and B agreed in a few more cases when some form of increased echogenicity was assessed. This finding may have been related to the fact that observers A and B had been working closely together for several years in the same sonography unit. The situation at the workstation when the examinations were reviewed was artificial, with no clinical information or clinical history for the patients, which could certainly have influenced the various options in the predetermined protocol.

Table 3. Interobserver Agreement

	κ Value (% Agreement)					
	First Reading			Second Reading		
Characteristic	**A vs B**	**A vs C**	**B vs C**	**A vs B**	**A vs C**	**B vs C**
Increased echogenicity[a]	0.86[b] (95)	0.70[b] (89)	0.75[b] (90)	0.82[b] (93)	0.72[b] (89)	0.72[b] (89)
Skip areas	0.82[b] (94)	0.76[b] (93)	0.84[b] (95)	0.73[b] (89)	0.75[b] (93)	0.56[b] (84)
Focal changes	0.50 (82)	0.70 (92)	0.60 (85)	0.37 (78)	0.85 (96)	0.40 (78)
Concrements in gallbladder	1.00[b] (100)	1.00[b] (100)	1.00[b] (100)	0.84[b] (95)	0.84[b] (95)	0.93[b] (98)
Polyps in gallbladder	0.58[b] (94)	0.93[b] (99)	0.69[b] (95)	0.69[b] (95)	0.48[b] (88)	0.48[b] (89)
Need for further examination	–0.05 (74)	0.29 (89)	0.04 (76)	0.10 (80)	0.46[b] (94)	–0.12 (74)

[a]For increased echogenicity, the weighted κ was used.
[b]$P < .05$

In a retrospective study conducted by Strauss et al,[3] in which 3 radiologists reevaluated still images to assess steatosis of the liver, the agreement was in general slightly lower than in our study, with intraobserver κ values between 0.51 and 0.63 and interobserver κ values between 0.40 and 0.53. Whether the slightly higher values in this study were due to the use of standardized examination protocols with cine loops cannot be reliably answered by our study. It should also be noted that the selection criteria differed between the two studies.

Echogenicity in the liver is affected by the ultrasound machine settings. In our study, the radiographer who conducted the examinations had at least 6 years of experience in abdominal sonography, which should have increased the probability of selecting appropriate parameters. It might thus be possible, with the use of standardized examination protocols, documentation of cine loops stored in a picture archiving and communication system, and the possibility of reviewing the examinations, to give the radiologist a more complete picture of the examination, even if it has been performed by someone else.

Observer experience is an important issue in all medical imaging. In this study, observer C had 10 years of experience in sonographic diagnosis but had never worked with recorded examinations before. She had a 2-week introduction to learn how to work with the standardized method and to review an examination made by someone else. Interestingly, this radiologist had almost equal results as the other 2 radiologists in this study regarding the intraobserver and interobserver agreement. On the other hand, we observed no tendency toward higher agreement with the more experienced observers in the second reading, which would have constituted evidence of a learning curve.

The fact that 2 of the 3 radiologists had worked closely together in the same sonography unit for at least 15 years may have had a positive effect on the results. However, it was only for increased echogenicity that a higher agreement level was found between these observers. Evidently, in our study, it was possible for a radiologist who had never worked with this sonographic method and only had a short introduction to achieve almost the same results as radiologists who were experienced in the method. This finding was seen for concrements in the gallbladder, focal changes, increased echogenicity, and bile duct dilatation.

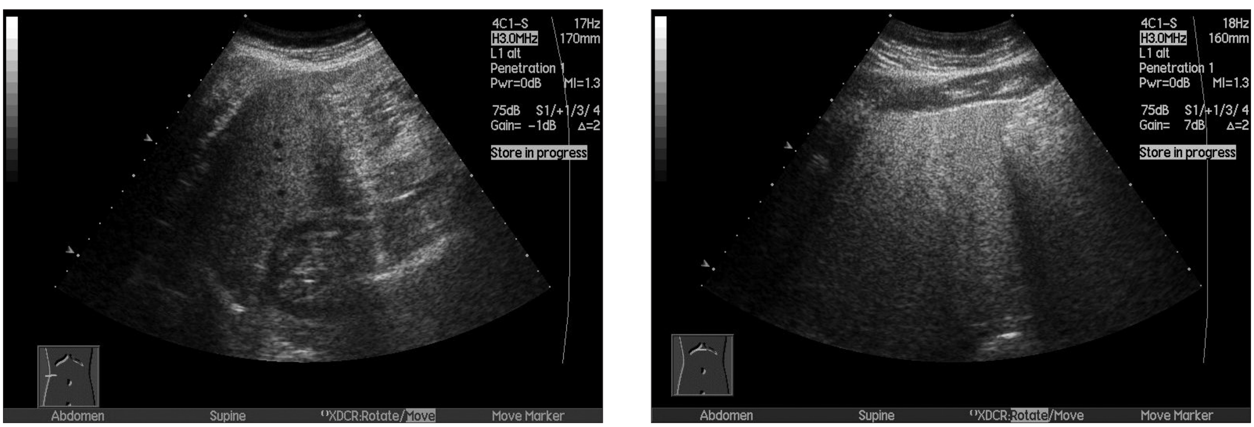
One limitation of this study was the small number of positive findings. A larger number of patients might have given more reliable results. In addition, a more targeted selection could have possibly provided higher κ values by bringing about a better balance between negative and positive findings.[17] It would also have been desirable to compare accuracy by correlating the findings to an independent reference method (reference standard), which, unfortunately, was not feasible in this setting.

An advantage of working with standardized examination protocols stored as cine loops is that it gives the opportunity to review the sonographic films and compare new and old examinations. It can offer the radiologist the opportunity to observe findings that were not noted at the time of examination, which may be useful, for example, if new clinical questions are brought up at a later time.[19] If the examination is performed in a technically correct manner, this method implies an increase in quality and safety, and technically uncomplicated normal sonographic scans can be performed by a radiographer.[19]

Figure 1. **A**, Liver with increased echogenicity for which the 3 observers did not agree. Their assessment of the echogenicity ranged from mildly to greatly increased. **B**, Liver with increased echogenicity for which there was complete agreement between the 3 observers. They all assessed the echogenicity as greatly increased.

A

B



If the radiographer performs uncomplicated examinations in which there is a clear clinical question, it will give the radiologist more time for acute and advanced examinations. The professional roles thus become more clear cut, with the radiographer concentrating on perfecting the examination technique and the radiologist on improving diagnostic skills, just as in other radiologic modalities. However, since professional roles differ between countries, it may be difficult to directly transfer this process to a different environment.

A relevant task for future studies would be to investigate the agreement between multiple radiographers who use the same standardized examination protocol in a larger study with more variation in pathologic findings. Since we have not investigated the relative importance of cine loop documentation versus a standardized protocol, another potential study could be to investigate how the assessment of static images agrees between and within observers. It would also be interesting to investigate the use of a standardized reporting system.

In conclusion, this study showed, in general, good interobserver agreement between the 3 radiologists when reviewing sonographic examinations of the liver and biliary system acquired by a trained radiographer using stored standardized cine loops, except for judgments of the need for further examination. The intraobserver agreement was higher than the interobserver agreement.

References

1. Harvey CJ, Albrecht T. Ultrasound of focal liver lesions. *Eur Radiol* 2001; 11:1578–1593.
2. Leslie A, Lockyer H, Virjee JP. Who should be performing routine abdominal ultrasound? A prospective double-blind study comparing the accuracy of radiologist and radiographer. *Clin Radiol* 2000; 55:606–609.
3. Strauss S, Ella G, Gottlieb P, Katsnelson L. Interobserver and intraobserver variability in the sonographic assessment of fatty liver. *AJR Am J Roentgenol* 2007; 189:W320–W323.
4. Zwiebel JW. Sonographic diagnosis of diffuse liver disease. *Semin Ultrasound CT MR* 1995; 16:8–15.
5. Golea A, Badea R, Suteu T. Role of ultrasonography for acute cholecystic conditions in the emergency room. *Med Ultrason* 2010; 12:271–279.
6. Hanbidge AE, Buckler PM, O'Malley ME, Wilson SR. From the RSNA refresher courses: imaging evaluation for acute pain in the right upper quadrant. *Radiographics* 2004; 24:1117–1135.
7. Choi SH, Kim EK, Kwak JY, Kim MJ, Son EJ. Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid* 2010; 20:167–172.
8. Rosen MP, Levine D, Carpenter JM, et al. Diagnostic accuracy with US: remote radiologists' versus on-site radiologists' interpretations. *Radiology* 1999; 210:733–736.
9. Yazbek J, Ameye L, Testa AC, et al. Confidence of expert ultrasound operators in making a diagnosis of adnexal tumor: effect on diagnostic accuracy and interobserver agreement. *Ultrasound Obstet Gynecol* 2010; 35:89–93.
10. Van Holsbeke C, Yazbek J, Holland TK, et al. Real-time ultrasound vs evaluation of static images in the preoperative assessment of adnexal masses. *Ultrasound Obstet Gynecol* 2008; 32:828–831.
11. Stenman C, Thorelius L, Knutsson A, Smedby Ö. Radiographer-acquired and radiologist-reviewed ultrasound examination: agreement with radiologist's bedside evaluation. *Acta Radiol* 2011; 52:70–74.
12. Crawley MT, Shine B, Booth A. Radiation dose and diagnosticity of barium enema examinations by radiographers and radiologists: a comparative study. *Br J Radiol* 1998; 71:399–405.
13. Marchal G, Tshibwabwa-Tumba E, Verbeken E, et al."Skip areas" in hepatic steatosis: a sonographic-angiographic study. *Gastrointest Radiol* 1986; 1:151–157.
14. Campbell MJ, Machin D. *Medical Statistics*. 3rd ed. Chichester, England: John Wiley & Sons; 1999.
15. Viera AJ, Garett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005; 37:360–363.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33:159–174.
17. Feinstein AR, Cicchetti DV. High agreement but low kappa, II: resolving the paradoxes. *J Clin Epidemiol* 1990; 43:551–558.
18. Stricker D. BrightStat.com: free statistics online. *Comput Methods Programs Biomed* 2008; 92:135–143.
19. Thorelius L. Can ultrasound benefit from the rest of radiology? AuntMinnie.com website. http://www.auntminnie.com/redirect/redirect.asp?itemid=74965. Accessed January 12, 2011.