Original Contribution

REAL-TIME QUALITATIVE ULTRASOUND ELASTOGRAPHY OF CERVICAL LYMPH NODES IN ROUTINE CLINICAL PRACTICE: INTEROBSERVER AGREEMENT AND CORRELATION WITH MALIGNANCY

KUNWAR S. S. BHATIA, CARMEN C. M. CHO, YUEN-HOK YUEN, DARSHANA D. RASALKAR, ANN D. KING, and ANIL T. AHUJA

Department of Diagnostic Radiology and Organ Imaging, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin N.T., Hong Kong

(Received 11 April 2010; revised 27 July 2010; in final form 26 August 2010)

Abstract—To evaluate real-time qualitative ultrasound (US) elastography for cervical lymphadenopathy in routine clinical practice, 74 nodes (37 malignant, 37 benign) in 74 patients undergoing sonography underwent US elastography prior to fine needle aspiration for cytology. Dynamic cine loops of elasticity imaging displayed using a chromatic-scale were qualitatively scored by three independent observers for the proportion of stiff areas from ES1-4 (soft to stiff). There was fair to good interobserver agreement as indicated by weighted kappa ($\kappa$) statistic from 0.374 to 0.738. Median ES for benign and malignant nodes were 2 and 3 respectively. ES was higher in malignant nodes ($p = 0.0003–0.0049$, Mann Whitney U tests) although areas under receiver operating characteristic curves (0.68–0.74) indicated suboptimal discrimination. The optimal discriminatory cut-off, ES $> 2$, achieved only 62.2% sensitivity, 83.8% specificity and 73% accuracy for malignancy. Improvements in reliability and accuracy of real-time qualitative ultrasound elastography are required for it to be adopted into routine clinical practice. (E-mail: aniltahuja@cuhk.edu.hk)

Key Words: Ultrasound, Elastography, Lymph nodes, Neck, Diagnosis.

INTRODUCTION

Qualitative real-time ultrasound (US) elastography is a comparatively novel technique now available on conventional US systems. In US elastography, tissue elasticity is estimated by comparing local tissue displacements from ultrasonic signals before and after application of a compressive force. Under compression, stiff tissues show less deformation, or strain, than soft tissues. Reflecting the fact that malignant tissues are stiffer than their benign counterparts at many sites, increasing numbers of recent reports document the utility of US elastography to differentiate malignant from benign lesions in the breast, prostate, liver, gastrointestinal tract, cervix and thyroid (Lyshchik et al. 2005; Taylor et al. 2005; Itoh et al. 2006; Bae et al. 2007; Garra 2007; Janssen et al. 2007; Rago et al. 2007; Thomas et al. 2007; Zhi et al. 2007; Asteria et al. 2008; Hong et al. 2009). In addition, a few preliminary reports evaluating US elastography for lymph nodes suggest that this technique is useful for detection of malignant nodes with reported accuracies ranging between 89% to 93% (Saftoiu et al. 2006; Lyshchik et al. 2007; Alam et al. 2008; Aoyagi et al. 2009). This is promising as the most widely used imaging techniques, conventional US including gray-scale US and power Doppler, computed tomography (CT) and magnetic resonance imaging (MRI) are reliant on anatomic criteria, which can overlap between benign and malignant nodal disease. When US is combined with fine needle aspiration for cytology (FNAC), excellent overall accuracies for malignancy can be achieved (93%–97%) (Baatenburg de Jong et al. 1991; Knappe et al. 2000). However, US guided FNAC is subject to inadequate or nonrepresentative sampling issues and, importantly, is dependent on prior selection of appropriate nodes using anatomic imaging criteria. This study was performed to evaluate real-time qualitative US elastography for sonographically abnormal cervical lymph nodes undergoing FNAC in a routine clinical setting. The main objective was to assess its utility for predicting malignancy within the spectrum of pathologies...
encountered in routine practice. In addition, as qualitative elastography is based on subjective evaluation of elastograms, a second objective was to assess the interobserver agreement for this technique.

METHODS

Patient selection

Between August 2008 and February 2010, 74 cervical lymph nodes in 74 patients undergoing conventional gray-scale and power Doppler US in a major head and neck cancer tertiary referral centre in Hong Kong were also examined by real-time freehand qualitative elastography. Patients had been referred from a variety of sources including otorhinolaryngology, clinical oncology as well as other hospital and primary care specialties. All nodes included in this study had at least one conventional sonographic feature of abnormality that warranted US guided FNAC. Sonographic features of abnormality are well documented in the literature and include round shape, irregular margin, hypoechogeticity, intra-nodal necrosis or calcifications, disordered vascularity (Tohno et al. 1989; van den Brekel et al. 1990; Vassallo et al. 1992, 1993; Ahuja et al. 2001; Chan et al. 2007). These criteria included nodes considered to be benign (e.g., reactive or infective) or malignant on conventional US. Of note, completely necrotic nodes without an evaluable solid component were not included. The study group comprised of 31 males and 43 females (mean age 50.0 years, range 13–86 years). Approval for this study had been granted by the local ethics committee and informed consent was obtained from all patients per the WORLD Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects, 2008.

US technique

Patients initially underwent conventional neck US comprising of gray-scale and colour Doppler US using a 5–12 MHz linear array transducer (Philips iU22; Philips Healthcare Nederland, Eindhoven, The Netherlands), performed by radiologists experienced in neck ultrasound. Subsequently, elastography using a linear array 13.5 MHz transducer (Siemens Acuson Premium Edition with eSie TouchTM; Siemens Healthcare, Erlangen, Germany) was performed by one of three radiologists (A–C) who also had at least 1 year of practical experience of performing elastography in the neck region. For the elastographic technique, patient and transducer positioning were identical to conventional US and elastograms were displayed in dual-mode superimposed and alongside gray-scale sonograms in real-time. An appropriate region-of-interest (ROI) was selected that included the lymph node and a moderate amount of surrounding tissues. Intermittent freehand transducer compressions were applied along the beam axis while paying attention to avoid out-of-plane or lateral transducer motions, which are causes of mistracking artifacts. Real-time elastograms were a colour-coded graphic representation of the relative stiffness of structures within the selected ROI such that purple indicated soft, green and yellow indicated intermediate stiffness and red indicated stiff. The correct amount of compression was determined by manual adjustments such that surrounding tissues predominantly displayed an intermediate strain pattern, yellow or green, as opposed appearing over- or under-compressed, i.e., red or gray. Three to five dynamic cine loops of elastography were acquired for analysis. Elastography took approximately 5 min and patients did not experience any additional discomfort from the technique.

Final diagnosis

Following elastography, all lymph nodes underwent FNAC using conventional sonography for needle guidance. Final diagnosis was based on cytologically adequate FNAC result. In addition, the electronic clinical records of patients with cytologic diagnoses of reactive nodes were also reviewed for a minimum follow-up period of 6 months to confirm that no malignancy had developed in these lymph nodes subsequently.

Image interpretation

Elastograms were reviewed independently by three radiologists (A–C) who had performed elastography. Observer A had 1.5 years practical experience with this technique whereas observer B and C had approximately 1 year experience each. All observers were blinded to the final diagnosis. To reduce recall bias, elastograms were anonymized and reviewed with a mean interval of 279 days (range, 14–550 days) from the time of US examination. To familiarize themselves with the scoring system, observers initially evaluated several cine loops using consensus. These cases were not included in the final analysis. Due to lack of universally accepted criteria for scoring elastograms of lymph nodes and other tissues in the published literature, elastograms were graded on a simplified 4-point scale (ES 1-4) evaluating the proportion of stiff areas within the node compared with surrounding fascia and soft tissues excluding muscle. (Table 1, Figs. 1–5). This scale was adapted from previous studies of thyroid US elastography (Lyshchik et al. 2005; Itoh et al. 2006; Rago et al. 2007; Asteria et al. 2008; Hong et al. 2009; Rubaltelli et al. 2009). Of importance, preliminary evaluation showed wide temporal fluctuation of elastogram cine loops at similar stages along compression-decompression cycles. Hence ES scores were based on the predominant elastographic appearance on dynamic segments as opposed to
evaluating a static image. In some cine loop segments, a thin red line or two parallel red lines were identified corresponding to the superficial margin of lymph nodes that were homogeneous on gray-scale US (Fig. 1). This was regarded as an artifact, related to tissue interfaces and with a predilection for the superficial margin as this was closest to the compression source. Similarly, the presence of second deeper red line paralleling the lymph node margin was regarded as an elastographic equivalent of a reverberation artifact. These red lines were excluded from elastogram analysis. Finally, areas of intra-nodal cystic change were excluded as this may lower the ES depending on their relative proportions within lymph nodes and these would not be indicative of the ES of the solid component. After elastogram review, nodal long axis and short axis dimensions were recorded from conventional US images for subsequent analysis.

**Statistical analysis**

Statistical analyses were performed using MedCalc for Windows, version 11.2.1.0 (MedCalc Software, Mariakerke, Belgium). Interobserver agreement for ES was evaluated using the kappa statistic ($\kappa$) with linear weighting to allow for the fact that the scale was ordered, hence, a lymph node scored as 1 and 2 by two observers was considered as a minor disagreement compared with scores of 1 and 4. The level of agreement based on the $\kappa$ statistic was interpreted as follows: $\kappa < 0.2$, poor; $\kappa = 0.21–0.4$ fair, $\kappa = 0.41–0.6$ moderate, $\kappa = 0.61–0.8$, good; $0.81–1.00$ very good (Altman 1991). Receiver operating characteristic (ROC) analysis was performed using the area under the ROC curves (AUC) to compare the performance of observers in terms of predicting malignancy. The optimum ES cut-off achieving the highest accuracy for malignancy was determined and corresponding precision data were obtained. To assess for a significant difference of ES between benign and malignant nodes, Mann Whitney U tests were performed for ES categories 1–4 and Fisher’s exact test were performed for nodes dichotomized using the optimal ES cut-off. Student’s $t$-tests were performed to evaluate for a significant difference in short and long axis diameters between malignant and benign nodes as this may be a potential confounding factor for elastography. All statistical tests were two-sided, using $p < 0.05$ to indicate a statistically significant difference.

**RESULTS**

Final diagnosis of lymph nodes based on FNAC with clinical follow-up was as follows: reactive, 21; necrotizing

<table>
<thead>
<tr>
<th>Elastographic score</th>
<th>Overall impression</th>
<th>Elastographic appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES 1</td>
<td>soft</td>
<td>Predominantly purple, green or yellow with less than 10% displaying red. The node is indistinguishable from surrounding tissues (Fig. 1a and b).</td>
</tr>
<tr>
<td>ES 2</td>
<td>moderately soft</td>
<td>Predominantly yellow or green and with red areas comprising between 10% and 50%. The node is partially delineated from surrounding tissues. (Fig. 2).</td>
</tr>
<tr>
<td>ES 3</td>
<td>moderately stiff</td>
<td>Predominantly red and with yellow or green areas comprising between 10% and 50%. The node is partially delineated from surrounding tissues. (Fig. 3).</td>
</tr>
<tr>
<td>ES 4</td>
<td>stiff</td>
<td>Predominantly red and with less than 10% appearing yellow or green. The node is distinguishable from surrounding tissues. (Figs. 4 and 5).</td>
</tr>
</tbody>
</table>

Fig 1. Gray-scale ultrasound (US) and elastogram of a reactive lymph node, which displays as mostly purple and green and was scored as soft (ES1). The superficial margin of the node (white arrows) displays as a red line (long arrows) beneath which another red line is identified (short arrows). These lines were considered to be artifacts and were excluded from the scoring of elastograms.
granulomatous lymphadenitis, 7; tuberculous lymphadenitis, 9; lymphomatous infiltration, 2; nodal metastases, 35. Nodal metastases comprised as follows; lung carcinoma 4 (1 adenocarcinoma, 1 small cell carcinoma, 2 squamous cell carcinoma); breast adenocarcinoma, 1; poorly differentiated nasopharyngeal carcinoma, 12; thyroid papillary carcinoma, 5; head and neck squamous carcinoma, 11; uterine mixed mullerian tumour, 1; ovarian adenocarcinoma, 1. Thirty-seven nodes were malignant (50%) and 37 were benign (50%). Mean long axis and short axis diameter of malignant nodes (2.0 ± 0.9 cm, 1.3 ± 0.7 cm) were not significantly different from benign nodes (2.0 ± 0.9 cm, 1.1 ± 0.5 cm) (long axis diameter, \( p = 0.77 \); short axis diameter, \( p = 0.23 \), Students t-tests).

The ES of nodes according to diagnosis including ES scores pooled for benign and malignant pathologic conditions are shown for each observer in Table 2. There was good agreement between observers A and B (weighted \( \kappa = 0.738 \)), fair agreement between observers A and C (weighted \( \kappa = 0.374 \)) and moderate agreement between observers B and C (weighted \( \kappa = 0.527 \)). The median score for benign nodes was ES2, and for malignant nodes was ES3. A statistically significant difference in ES score was found between benign and malignant lymph nodes for all three observers (\( p = 0.0003, p = 0.0006, p = 0.0049 \) for observers A, B and C, respectively, Mann Whitney U tests). ROC analysis indicated AUCs of 0.738, 0.720 and 0.678 for observers A, B and C, respectively (Fig. 6, Table 3). The optimal cut-off to identify malignant nodes was ES > 2 for each observer. The sensitivity, specificity and accuracy for the observer with highest global accuracy, observer A, using ES > 2 to predict malignancy was 62.2%, 83.8% and 73%, respectively. On a case by case basis, using ES > 2 to indicate malignancy missed between 10 and 15 malignant nodes (27%–40%). Lowering the cut-off to ES > 1 missed between two and six malignant nodes (5%–16%). Conversely, using a cut-off to ES > 3 missed between 27 and 33 malignant nodes (73%–89%) and produced between 0 and 2 false positives for malignancy (0%–8%).

**DISCUSSION**

To date there are only a few published systematic studies evaluating US elastography of lymph nodes. Lyshchik et al. evaluated 141 cervical lymph nodes using...
semiquantitative relative strain indices (ratios) from elastograms generated after relatively laborious off-line post-processing. Their results indicated 98% sensitivity, 85% specificity and 92% accuracy for elastography to predict malignancy, compared with 79% accuracy for conventional sonography (Lyshchik et al. 2007). While the results are encouraging, off-line elastography has two important drawbacks; first, the time delay incurred by producing elastograms off-line means that it cannot be used to guide clinical decisions during ultrasound examinations; second, the results cannot be extrapolated to real-time elastography because of differences in software postprocessing used to generate elastograms. In this respect, off-line processing permits the use of more sophisticated albeit time consuming algorithms to track tissue displacements (cf. computationally efficient hence faster algorithms are required for real-time elastography) with resultant increases in image signal and fewer mistracking artifacts. Nevertheless, two other studies have evaluated real-time qualitative elastography on commercially available systems; Saftoiu et al. documented 92% sensitivity, 94% specificity and 93% accuracy for real-time elastography in a relatively small study (n = 31) that included five cervical nodes (Saftoiu et al. 2006); Alam et al. evaluated 85 cervical lymph nodes and documented 83% sensitivity, specificity 100% and 89% accuracy and documented that elastography was more accurate than gray-scale sonography (Alam et al. 2008).

The present study results indicate significantly higher elastographic scores of malignant compared with benign lymph nodes, which is in general agreement with the aforementioned other studies. However, the precision data for an optimised cut-off in the present study (62% sensitivity, 84% specificity and 73% accuracy) are appreciably worse than that documented in the other studies (Saftoiu et al. 2006; Lyshchik et al. 2007;
Alam et al. 2008). We are uncertain as to the cause for discrepancy although it may reflect differences in elastographic software and settings, operator technique or inclusion criteria. With respect to the latter, we postulate that lower accuracy results in the present study may be partly due to greater variety of primary tumour sites, as well as inclusion of lymphoma and tuberculous lymphadenitis (both of these conditions produced false negatives and positives for malignancy, respectively). Our findings highlight the need for further research to evaluate elastography for a wide variety of different benign and malignant nodal pathologies if this technique is to be adopted in routine clinical practice. In this regard, the utility of elastography will depend partly on the emerging precision data since if a highly sensitive test can be found it will be useful for screening, whereas a highly specific test may be useful for the work-up of equivocal nodes on conventional sonography by influencing the decision whether or not to perform FNAC. Unfortunately, our findings suggest that real-time qualitative elastography, using the specific system in this study, is neither sufficiently nor sensitive or specific to accomplish either of the roles alluded to above. Consequently, while it may have an ancillary role to conventional B-mode US and colour Doppler, it is unlikely to obviate the requirement for FNAC.

Manual compression technique, selection of representative elastograms from a dynamic series and elastogram interpretation are all aspects of elastography that can vary between operators. Published evidence for reproducibility of this technique is extremely limited as most studies have evaluated elastograms using consensus opinion. Saftiou et al. reported no statistical difference between two observers scoring elastograms qualitatively but did not provide further details of their analysis (Saftiou et al. 2006). The present study findings indicate a fair to good interobserver agreement of scoring of real-time qualitative elastograms of cervical lymph nodes for the range of conditions undergoing FNAC in routine clinical practice. This is encouraging although not ideal because an excellent level of agreement was not achieved. The highest global accuracy as indicated by the AUC was achieved by observer A, who also had most experience of performing elastography, which suggests that there is a prolonged learning curve for this technique. It should be highlighted that, for practical reasons, elastography was performed by one radiologist per patient and the resultant archived elastograms were scored independently by three observers, as opposed to elastography being performed and scored independently. Consequently, our study has not accounted for any potential variation as result of differing compression technique. Despite this limitation, our data pertaining to review of identical elastograms indicate that there is a definite requirement for improvements to increase the robustness of this technology.

Several other aspects of the study design should be addressed.

- Because tissue elasticity changes non-linearly depending on the strain applied, lack of quantitative feedback regarding the appropriate amount of compression to

### Table 2. Elastographic scores of 74 lymph nodes from three independent observers listed according to final diagnosis

<table>
<thead>
<tr>
<th>Diagnosis (number)</th>
<th>Observer A</th>
<th>Observer B</th>
<th>Observer C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES1 ES2 ES3 ES4</td>
<td>ES1 ES2 ES3 ES4</td>
<td>ES1 ES2 ES3 ES4</td>
</tr>
<tr>
<td>Benign Reactive (21)</td>
<td>7 9 3 2</td>
<td>8 5 6 2</td>
<td>4 10 7 0</td>
</tr>
<tr>
<td>Necrotizing granulomatous</td>
<td>3 4 0 0</td>
<td>2 5 0 0</td>
<td>2 4 1 0</td>
</tr>
<tr>
<td>lymphadenitis (7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuberculous lymphadenitis (9)</td>
<td>3 5 1 0</td>
<td>4 3 2 0</td>
<td>2 3 4 0</td>
</tr>
<tr>
<td>Pooled benign diagnoses (37)</td>
<td>13 18 4 2</td>
<td>14 13 8 2</td>
<td>8 17 12 0</td>
</tr>
<tr>
<td>Malignant Lymphoma (2)</td>
<td>1 1 0 0</td>
<td>0 1 1 0</td>
<td>0 1 1 0</td>
</tr>
<tr>
<td>Metastatic (35)</td>
<td>6 7 10 12</td>
<td>6 3 21 5</td>
<td>2 12 17 4</td>
</tr>
<tr>
<td>Pooled malignant diagnoses (37)</td>
<td>6 8 11 12</td>
<td>6 4 22 5</td>
<td>2 13 18 4</td>
</tr>
<tr>
<td>Totals</td>
<td>19 26 15 14</td>
<td>20 17 30 7</td>
<td>10 30 30 4</td>
</tr>
</tbody>
</table>

Fig. 6. Receiver operating characteristic (ROC) curves of data from three observers using ES thresholds to identify malignant nodes. (Dotted line at 45 degrees indicates a theoretical test with no discrimination).
apply was a limitation. Newer elastography software packages now offer quantitative feedback regarding the quality or degree of compression, however, this was not available on the software used at the time of this study. In an attempt to overcome this limitation, over-compression was avoided by subjective evaluation of surrounding connective tissues to ensure that they did not appear predominantly red, i.e., stiff.

- The scoring system used in the present study was simplified to four categories due to our preliminary experience that revealed that more elaborate scoring systems were too unreliable for lymph nodes. In this respect, evaluation of nodal margins on elastograms was particularly unreliable due to the presence of artifacts on the superficial margin of nodes (Fig. 1). Furthermore, elastographic criteria comparing the size of lesions on elastograms with that on gray-scale US, termed area ratios, may be valid for breast lesions where microscopic tumour infiltration into the surrounding parenchyma may be present (Hall et al. 2003; Regner et al. 2006), however, this is unlikely to be as useful for lymph nodes where a discrete tissue boundary, namely the nodal capsule, is present. With future studies, a universally acceptable and accurate elastographic scoring system for lymph nodes may emerge.

- Lack of quantitative output for elastography is a major limitation of many currently available US elastography software including in the present study. Furthermore, qualitative elastographic scoring systems have so far assigned a score based on the entire node, which will be suboptimal for nodes that are only partially infiltred, e.g., by micro-metastases. Off-line processing can be used for strain quantification including for assessment of selected areas within nodes, however, as stated earlier, this cannot guide decision making during busy US clinics. More recent advances in tissue elasticity imaging may translate into a more accurate and clinically useful technique, including real-time quantitative elasticity imaging utilising strain ratios, as well as acoustic radiation force impulse (ARFI) elasticity imaging utilising shear wave velocities (Garra 2007; Melodelima et al. 2007; Tanter et al. 2008).

- Other preliminary studies suggest that elastography is capable of achieving accuracies comparable or superior to conventional US for detection of malignancy in lymph nodes (Saftoiu et al. 2006; Lyshchik et al. 2007). The present study did not analyse conventional sonography partly because this is already documented in the literature and partly because of the study design, namely that all lymph nodes had least one sonographic abnormality that had warranted FNAC. In this respect, the false negative rate and, thus, the sensitivity and accuracy of conventional sonography could not be calculated. Similarly, the accuracy data for elastography is also based on a preselected cohort of lymph nodes, although we envisage that elastography will be adopted within busy US clinics in a similar fashion, i.e., conventional sonography will be used as the initial screening tool followed by elastographic assessment of any nodes that are sonographically abnormal. Nevertheless, this limitation can be overcome in future studies of cohorts of unselected nodes with sufficient follow-up for nodes that are not biopsied.

Table 3. Discriminatory performances of ES scores for differentiating malignant from benign nodes using ES > 2 to indicate malignancy

<table>
<thead>
<tr>
<th>Observer</th>
<th>AUC (95% CI)</th>
<th>Sensitivity %</th>
<th>Specificity %</th>
<th>Accuracy %</th>
<th>+LR</th>
<th>-LR</th>
<th>Fisher’s exact test</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.738 (0.623–0.833)</td>
<td>62.2</td>
<td>83.8</td>
<td>73.0</td>
<td>3.83</td>
<td>0.45</td>
<td>p = 0.0001</td>
</tr>
<tr>
<td>B</td>
<td>0.720 (0.604–0.818)</td>
<td>73.0</td>
<td>73.0</td>
<td>73.0</td>
<td>2.70</td>
<td>0.37</td>
<td>p = 0.0001</td>
</tr>
<tr>
<td>C</td>
<td>0.678 (0.560–0.782)</td>
<td>59.5</td>
<td>67.6</td>
<td>47.0%</td>
<td>1.83</td>
<td>0.60</td>
<td>p = 0.035</td>
</tr>
</tbody>
</table>

AUC = area under receiver operating characteristic (ROC) curve; +LR = positive likelihood ratio; -LR = negative likelihood ratio; CI = confidence interval.

In conclusion, this preliminary study documents fair to good interobserver agreement for the review of real-time qualitative elastograms of lymph nodes undergoing FNAC in routine clinical practice. This is encouraging but indicates the need for further refinements to improve the robustness of the technique. Although the present data indicates a significant correlation between increasing tissue stiffness and malignancy in lymph nodes, from the present data this technique appears to be insufficiently accurate to influence decision making in terms of selecting or excluding lymph nodes for FNAC. Furthermore, our findings are less promising compared to other preliminary studies, suggesting that the technique is highly operator or software dependent and needs to be standardized for it to be adopted into routine clinical practice.
REFERENCES


