THE INTRA-OBSERVER VARIABILITY OF ACOUSTIC RADIATION FORCE IMPULSE (ARFI) – IMPLICATIONS FOR THE USE OF THE METHOD IN MONITORING CHRONIC LIVER DISEASES

George Sebastian Gherlan¹, Petre Iacob Calistru¹, Cristina Voinea², Cristian Szabo²
¹Infectious Diseases Department, “Dr. Victor Babes” Center for Diagnostic and Treatment, Bucharest, Romania
²Laboratory Department, “Dr. Victor Babes” Center for Diagnostic and Treatment, Bucharest, Romania

ABSTRACT

Introduction and aims. Liver fibrosis is reversible if the cause is removed or it may progress otherwise. Noninvasive methods tend to replace liver biopsy for the assessment of liver fibrosis. There is an intra-observer variability for ARFI that has to be taken into account when this method is used for monitoring the evolution of a liver disease. We aimed to identify the level of variation between two ARFI results that reflects a real change in liver histology rather than this “normal” variability.

Patients and methods. Two sets of 10 ARFI measurements were performed in 147 patients by the same observer and we computed the median and the interquartile range for each set. We assessed the variation between the two results as percentage of each result. Data regarding weight, height, age, etiology, biochemical and hematological test were collected.

Results. The mean variation was of 9.84% ± 9.7%. The variability is reduced if we use IQR for validation (valid if IQR < 1/3 of the result), with a variation of 7.66% ± 7.09% in the group with both measurements valid versus 16.33% ± 13.03% in the group with at least one set invalid. ARFI was feasible in all patients and we found an intra-observer intraclass correlation coefficient of 0.976.

Conclusions. Only results with over 19.5% variation in plus or minus compared to a previous result should be considered to reflect the progression respectively regression of liver fibrosis. If IQR is used as a validation parameter, the variability is less than 15%.

Keywords: ARFI, intra-observer, hepatitis, liver fibrosis, monitoring, variability

INTRODUCTION

Currently more and more data is accumulating in the support of the information that liver fibrosis (and even cirrhosis) is reversible (1,2). Studies based on seriated liver biopsies have been conducted in alcoholic liver disease (3,4,5), chronic hepatitis B (6,7,8,9,10,11), chronic hepatitis C (11,12,13,14,15), autoimmune hepatitis (16,17,18,19), biliary obstruction (20) and they showed that fibrosis is reversible after removal of the injury stimulus. If the chronic liver disease is not treated – on the other hand – fibrosis may progress.

Liver biopsy is still considered the golden standard for the evaluation of liver fibrosis, but is subject to possible sampling errors and intra-observer variability (21,22) and the available histological scoring systems are semi quantitative and observer-dependent (22,23). It has a risk of severe complications of 0.57%, a risk of mortality of 0.009-0.12% (22,24,25) and is hardly accepted by the patients. Noninvasive tests that tend to replace liver biopsy...
for the purpose of staging the disease have been developed in the last 10 years. Some of these noninvasive tests, serologic markers or imaging techniques have been extensively validated in large and numerous studies and proved their accuracy against liver biopsy for staging liver fibrosis (26,27,28,29,30,31,32).

There are studies showing that if we monitor patients treated for hepatitis B or C we can observe a decrease in liver stiffness measured by either transient elastography (TE/Fibroscan®) or acoustic radiation force impulse imaging (ARFI) (33,34,35,36,37). The above-mentioned studies show a significant decrease of liver stiffness in patients that have good virologic response to treatment compared to those who do not respond. However, when it comes to an individual patient, in whom we perform seriated liver stiffness measurements in time, in the monitoring process, there are no clear data on what difference we should consider as significant in terms of improving or worsening of the liver disease as the liver stiffness decreases or respectively increases. The amount of the decrease or increase reported to a previous measurement that actually reflects histological changes is not well defined in the literature yet. For transient elastography we found in a previous study that only a variation of over 25% in plus or in minus should be considered as a real modification in the liver fibrosis stage, below this limit the variation being attributable to the intra-observer normal variability of the method (38). Therefore is useful to know that some of the variations between two liver stiffness measurements performed on the same patient are the result of the variability of the technique and do not reflect an actual change in liver stiffness or liver fibrosis stage.

The main objective of this study was to identify the level of variation (the difference) between two ARFI measurements performed over time in the same patient that should be considered significant in terms of improving or worsening of a liver disease and reflects a real change in the liver stiffness rather than the intra-observer variability of the method. We also tried to identify factors that influence the variability of the ARFI measurements.

PATIENTS AND METHODS

All patients referred to our ultrasonography department for abdominal ultrasound examination, with known liver diseases but also those that had no known liver disease at the time of referral were asked if they wanted to participate in the study after they were informed regarding the procedures that were involved in this research. Over a period of three months, 147 patients agreed to participate in the study and provided written consent before they were subjected to any procedure, in accordance with the principles of the Declaration of Helsinki (revision of Fortaleza 2013). The local ethics committee approved the study protocol.

For every patient, the height and weight were measured and the body mass index (BMI) was calculated. We collected the information regarding the age and sex of every patient as well as data concerning their underlying liver pathology. We also took blood samples for the following parameters: alanine aminotransferase (ALT), aspartate aminotransferase (AST), gama-glutamyl transpeptidase (GGT) and complete blood count (for platelets). ARFI was performed on an ACUSON S2000 ultrasound device equipped with the 4C1 1-4.5 MHz transducer and with Virtual Touch Tissue QuantificationTM software installed. All the examinations were done by the same operator, with an experience of over 500 procedures at the time of the study. Every patient was examined in left lateral decubitus, with the right arm in maximum abduction above the head. The transducer was placed in the intercostal spaces without applying any pressure. The region of interest (ROI) was chosen and set by the operator at a depth of at least 2 cm from the liver capsule and away from any visible blood vessel, in the right hepatic lobe. For each patient 20 consecutive valid measurements were done (two sets of 10 measurements) at the same spot. We calculated the median value for every set of 10 measurements as well as the interquartile range (IQR) and also the median and IQR for all the 20 measurements. The interquartile range, also called the midspread or middle fifty, is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles. We used ARFI cut-offs of 1.31 m/s for significant fibrosis (F ≥ 2) and 1.8 m/s for cirrhosis (F4) (30).

RESULTS

A total number of 147 patients out of the 362 examined in the mentioned period in our department were included in the study. Their characteristics are summarized in Table 1. The etiologies of the underlying diseases of the patients were: non-alcoholic fatty liver disease (NAFLD) in 28 patients, chronic hepatitis B (CHB) in 22 patients and chronic hepatitis C (CHC) in 55 patients. 42 patients had no known liver disease at the time of en-
rollment and their ultrasound examination did not revealed any pathological findings regarding the liver.

**TABLE 1. Patient characteristics**

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Number/Mean ± Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients number</td>
<td>-</td>
<td>-</td>
<td>147</td>
</tr>
<tr>
<td>Male</td>
<td>-</td>
<td>-</td>
<td>71 (48.29%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>17</td>
<td>82</td>
<td>52.88 ± 14.7</td>
</tr>
<tr>
<td>ALT</td>
<td>11</td>
<td>123</td>
<td>38.46 ± 25.45</td>
</tr>
<tr>
<td>AST</td>
<td>12</td>
<td>169</td>
<td>31.73 ± 24.46</td>
</tr>
<tr>
<td>GGT</td>
<td>7</td>
<td>455</td>
<td>46.78 ± 58.74</td>
</tr>
<tr>
<td>Platelets (x103/mm)</td>
<td>60</td>
<td>392</td>
<td>225 ± 58</td>
</tr>
<tr>
<td>BMI</td>
<td>15.2</td>
<td>38.6</td>
<td>25.92 ± 4.36</td>
</tr>
</tbody>
</table>

ARFI was feasible in all 147 patients (20 measurements). If we consider valid only the measurements that have an IQR of less than 1/3 of the median value, the total number of reliable results varies from 119 (80.95%) to 128 (87%) for 20 respectively 10 hits/procedure (Table 2). The influence of the number of measurements (10 versus 20) on the rate of reliable results is not statistically significant (p = 0.373 and 0.6 for the first and the second set of 10 measurements versus 20 measurements).

Based on the mentioned cut-offs and on all 20 measurements, 56 patients had significant fibrosis and 33 of them had cirrhosis. 3 patients out of the 147 (2.04%) were differently classified as significant/non-significant by the two sets of 10 measurements and 3 patients (2.04%) were differently classified as cirrhotic/non-cirrhotic. All the misclassified patients had IQR less than 1/3 of the median values of the results for both sets of 10 measurements.

**TABLE 2. Main characteristics of the ARFI measurements**

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Number/Mean ± Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st set median (m/s)</td>
<td>0.67</td>
<td>3.39</td>
<td>1.46 ± 0.67</td>
</tr>
<tr>
<td>1st set IQR</td>
<td>0.03</td>
<td>1.29</td>
<td>0.32 ± 0.25</td>
</tr>
<tr>
<td>1st set valid (IQR &lt; 30%)</td>
<td>–</td>
<td>–</td>
<td>123 (83.6%)</td>
</tr>
<tr>
<td>2nd set median (m/s)</td>
<td>0.67</td>
<td>3.82</td>
<td>1.51 ± 0.72</td>
</tr>
<tr>
<td>2nd set IQR</td>
<td>0.01</td>
<td>1.66</td>
<td>0.30 ± 0.26</td>
</tr>
<tr>
<td>2nd set valid (IQR &lt; 30%)</td>
<td>–</td>
<td>–</td>
<td>128 (87%)</td>
</tr>
<tr>
<td>20 hits median</td>
<td>0.68</td>
<td>3.64</td>
<td>1.48 ± 0.68</td>
</tr>
<tr>
<td>20 hits IQR</td>
<td>0.04</td>
<td>1.44</td>
<td>0.34 ± 0.27</td>
</tr>
<tr>
<td>20 hits valid</td>
<td>–</td>
<td>–</td>
<td>119 (80.95%)</td>
</tr>
</tbody>
</table>

The correlation between the two sets of 10 measurements was high, with a correlation coefficient of 0.955, p < 0.001.

The intraclass correlation coefficient (ICC) was 0.976 (CI95% = 0.966-0.982, p < 0.001), showing a good intra-observer reproducibility.

We calculated the absolute value of the difference between the two results for every patient and then expressed it as a percent of each of the two results. The minimum variation was 0% and the maximum variation was of 74.28%. The frequency

![FIGURE 1. Frequencies of the percentage of variation between two measurements – results for all group](image-url)
of each percentage we obtained can be seen in Fig 1. The mean variation was of 9.84%, with a standard deviation of 9.70% (Fig. 1).

When we analyzed whether the IQR used as a quality factor has any influence on the variability of the results, we found a significant difference be-

**FIGURE 2.** Frequencies of the percentage of variation between two measurements in patients with both results valid (IQR less than one third of the mean of the 10 measurements)

**FIGURE 3.** Frequencies of the percentage of variation between two measurements in patients with at least one set invalid (IQR higher than one third of the mean of the 10 measurements)
between the group that had an IQR of less than one third of the result on both sets of measurements and the group in which IQR was higher than one third of the result in at least one set of 10 measurements.

In 110 patients both results had an IQR of less than 1/3 of the result, while in 37 patients, at least one set had an IQR of over one third of the final result.

In the group with both results validated by the mentioned rule, the mean difference between the results is of 7.66% with a standard deviation of 7.09% (Fig. 2). The group with at least one invalid set had a mean variability of 16.33% and a standard deviation of 13.03% (Fig. 3). The mean difference between the two groups was of 8.67% and the standard deviation of 1.57% (CI95% 5.46-12.81, p < 0.001).

Age and BMI showed no correlation with the variation between two measurements expressed as percent from the final results. The correlation coefficients and the statistical significance were respectively \( r = -0.042, \) \( r = 0.136 \) and \( p = 0.613, \) \( p = 0.108. \)

We also concluded that the analyzed biochemical and hematological parameters (ALT, AST, GGT and platelets number) do not influence the variability between two consecutive ARFI results (Table 3).

The statistical analysis showed that the etiology of the liver disease has no influence on the level of variation of the method. The median variation between two consecutive results by etiology was: for NAFLD 9.12% ± 13.66%, for CHB 11.68% ± 6.99%, for CHC 10.5% ± 10.7% and for the patients without a known liver disease 9.86% ± 11.10%.

**DISCUSSION**

Currently there is a tendency to replace liver biopsy for the initial evaluation of the liver fibrosis with noninvasive methods. When it comes to monitoring the evolution of a liver disease, this trend seems even more justified, mainly because the possible adverse effects of the biopsy.

Acoustic Radiation Force Impulse Imaging (ARFI) is a relatively new technique, based on the measurement of the speed of propagation of a shear wave (generated by an acoustic wave) through the liver. It has the advantage that is implemented on a normal ultrasound machine, so the exam can be done during a routine ultrasonography evaluation. The observer can select the exact region of interest where the measurements are to be done. Studies and meta-analyses have shown a good correlation of ARFI results with the liver biopsy (39,40,41,42,43). Studies regarding the usefulness of ARFI in monitoring patients with chronic liver diseases are rare. We found two studies in patients with chronic C and B hepatitis showing that there is a significant decrease in liver stiffness in patients who have a favorable outcome to the antiviral treatment versus those who do not respond to treatment (36,37). However, in the particular case of a patient that has two ARFI measurements over time, at two different moments, we have to decide if the difference between the two measurements is significant and reflects an actual change in liver histology, or is just an intrinsic variation due to the method (intra-observer variability) or the inter-observer variability.

In this study we have found that the overall mean intra-observer variability is 9.84% with a standard deviation of 9.7%.

We have also found that if we use a validation rule similar with the one used for transient elastography – to consider valid only those results that have an IQR of less than one third of the final result – the intra-observer variability of ARFI is of 7.66% ± 7.09% for the group that has both set of 10 measurements valid. By comparison, the group in which at least one result is invalid had a variability of 16.33% ± 13.03% and the difference between the two groups is statistically significant. This proves that the use of IQR as a validation parameter improves the accuracy of ARFI.

No other parameter analyzed in our study had any statistically significant influence on the intraobserver variability of ARFI.

Only a small number of patients – 3 (2.04%) have been classified in two different categories by the two sets of measurements for significant/no significant fibrosis and also 3 patients for cirrhosis/non cirrhosis with the cut-offs of 1.31 m/s for significant fibrosis (F ≥ 2) and 1.8 m/s for cirrhosis (F4) (30). It is important to note that all the misclassified patients had valid results at both sets of 10 ARFI measurements. Therefore, as expected, even if we use IQR for validation, we can have patients

**TABLE 3. Correlation of the variation between two results with age, BMI, ALT, AST and GGT**

<table>
<thead>
<tr>
<th>Difference between two measurements expressed as percent from the final results</th>
<th>AGE</th>
<th>BMI</th>
<th>ALT</th>
<th>AST</th>
<th>GGT</th>
<th>Platelets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation (r)</td>
<td>-0.042</td>
<td>0.136</td>
<td>0.020</td>
<td>0.027</td>
<td>-0.087</td>
<td>0.012</td>
</tr>
<tr>
<td>Sig. (2-tailed) (p)</td>
<td>0.613</td>
<td>0.108</td>
<td>0.852</td>
<td>0.799</td>
<td>0.420</td>
<td>0.917</td>
</tr>
</tbody>
</table>

BMI = Body mass index; ALT = alanine aminotransferase; AST = aspartat aminotransferase; GGT = gamma-glutamyl transpeptidase
that are classified in two different fibrosis stages by two different measurements.

The feasibility of the method was of 100% in our group, but if we used the mentioned validation rule, only 251 (85.37%) of the 294 sets of 10 measurements were considered reliable. 110 (74.82%) patients out of 147 had both results valid.

The intraclass correlation coefficient, a statistical parameter that describes how strongly two sets of results resemble each other, applied to the two sets of 10 measurements has shown a good intra-

observer reproducibility, with a value of 0.976.

In conclusion, our analysis showed that using IQR as a validation factor (IQR < 1/3 of the final result) lowers the variability of the method. With this validation criteria fulfilled, any variation of over 15% between two ARFI measurements made by the same operator at two different time points can be considered indicative (or at least suggestive) of a change in liver histology.

Disclosures: Nothing to disclose.


