Quantitative Doppler in musculoskeletal ultrasonography – suboptimal performance of both experienced and in-training sonographers in selection of the highest Doppler signal image from cine-loops

Florian Berghea¹, Violeta Vlad¹, Lavinia Palanciuc², Mihaela Micu³, Violeta Bojinca¹, Florentin Vreju⁴, Luminma Enache⁵, Monica Copotoiu⁶, Alexandra Kosevoi⁷, Teodora Serban⁷, Denisa Stanciu⁵, Mihai Abobului¹, Andreea Borangiu¹, Camelia Berghea⁸, Andra Rodica Balanescu¹, Ruxandra Ionescu¹

¹Department of Internal Medicine and Rheumatology “Sf. Maria” Hospital, Bucharest, Romania, ²Medical Technology Research Center, Bucharest, Romania, ³Rheumatology Division, 2nd Rehabilitation Department, Clinical Rehabilitation Hospital Cluj-Napoca, ⁴Rheumatology Department, University of Medicine and Pharmacy Craiova, Craiova, Romania, ⁵Rheumatology, “Dr. Ion Stoia” Clinical Center of Rheumatic Diseases, Bucharest, Romania, ⁶Emergency Hospital, Tg. Mures, Romania, ⁷Internal Medicine and Rheumatology, “Dr. I Cantacuzino” Hospital, Bucharest, Romania, ⁸Emergency Hospital MS Curie, Bucharest, Romania

Abstract

Aims: Doppler ultrasonography assessment is mandatory nowadays for the complete description of rheumatic disease activity. Initially it was performed in semi quantitative way but recently the (fully) quantitative assessment is gaining more interest. In quantitative assessment, the ratio between total colorized and total pixels (CTR) is computed for the whole image or just for the region of interest (ROI). The frame with the highest amount of Doppler signal (also called worst case scenario image – WCSI) is usually the only one analyzed. The technique requires a very precise identification of WCSI from a certain number of consecutive frames, captured from the same position of the US probe, (and in most cases this is done manually). Our study examined the ability of both experienced and in-training sonographers to identify WCSI using a computerized analytical system as the gold standard.

Materials and methods: The study analyzed 480 frame selections done in two distinct exercises. The WCSI and other 3 images with a 5%, 10% and respectively 20% lower level of CTR compared with WCSI were packed in one selection. All frames emerging from the same video clip were randomly presented to six experienced and six in training sonographers; the request was to select the frame with the highest CTR (WCSI) from each package (twenty packages in total). A similar exercise was performed with CTRs decreasing in steps of 2%. Results: In the first exercise the WCSI was correctly identified in 79.1% cases and in 67% of cases in the 2nd exercise. The interobserver agreement between experienced and in-trainer evaluators for the 1st exercise was 0.78 and 0.4 in the 2nd exercise. Conclusion: Using computerized analysis as the gold standard, we demonstrated a large heterogeneity across sonographers regarding their ability to identify the best Doppler image even from a small group of frames.

Keywords: ultrasonography; Doppler; quantitative assessment

Introduction

Many rheumatic diseases develop localized inflammation during their evolution; this inflammation correlates well with the severity of disease [1] and can be very informative in the assessment of treatment’s success [2]. Musculoskeletal ultrasound is widely used in the evaluation of these situations – power Doppler ultrasound (PD) offers a visual picture of the extension and intensity of inflammation, very useful for monitoring purposes [3-5]. PD (semi) quantitative scales are used in practice to describe the level of inflammation, mostly for practical reasons. Unfortunately, semi quantitative evaluation has several limits: it is subjective (related to the expertise of the evaluator), has a low sensitivity to change and thus can jeopardize the objective of a rapid adjustment of the
therapy [6], and is dependent on the quality of the machine [7]. Most of the semi quantitative scales are developed in 4 steps: 0–3 (0-no Doppler signals, 1- single vessel PD signals, 2- confluent vessel signals in less than half of the area of the synovium, 3- vessel signals in more than half of the area of synovium) [8]. On the other hand, the quantitative assessment is time consuming (see below) and highly dependent on the ability to select the worst case scenario image (WCSI) [9]. The fully quantitative PD evaluation is a three step process: first the images are captured as a cine-loop (that is a video clip composed from up to several hundred distinct frames), then the frame with the higher amount of Doppler signal (WCSI) is selected from the cine-loop, and finally the computerized counting of the colored pixels (positive for PD signal is performed) [10]. The process uses one or two incidences (longitudinal and/or transversal) and it is supposed to be less subjective than the semi-quantitative assessment at least in the counting phase. However, there is a great challenge in selecting the most informative image (WCSI); this is less caused by the transducer’s position (the standard sections are widely accepted and easily acquired by experienced sonographers) but by the continuous variation of the number of PD dots (pixels). The number of colored pixels is related to cardiac activity (that continuously changes the PD loading) and the mechanical proprieties of the local vascular network [11]. The selection of the WCSI frame is usually done manually; the selection process is related to one Region of Interest (ROI). The method is demonstrated to offer a better reliability than the semi-quantitative evaluation [9]. Up to now, in musculoskeletal sonography, cine-loop and manual selection of WCSI method was the preferred technique in Quantitative Doppler assessment. The technique was improved by several US machine producers that implemented on their most advanced machines a simultaneous visualization of Doppler signal and cardiac activity; this offers the sonographer the ability to select the frame corresponding to the highest blood flow [7].

The final aim of both semi quantitative and fully quantitative assessments is to demonstrate a real change of the inflammation level in the selected ROI [2,5]. For the semi quantitative technique, a very small difference in local inflammation is difficult to be demonstrated; for the fully quantitative technique the identifiable difference between two frames are much smaller which create a superiority of this assessment [9]. In order to identify the smallest change of Doppler signal, a very precise selection of the best Doppler frame from a recorded cine-loop is needed; otherwise the selection bias will modify the results. The weak point of the full technique is the subjective selection of the frame with the highest PD signal which is done by a human observer. The question is if the sonographers are able to select the WCSI. Our study explored this hypothesis and examined the ability of both experienced and in-training sonographers to select WCSI from such a cine-loop using a computerized analytical system as the gold standard.

The main objective of our study was to evaluate the ability of experienced and in-training sonographers in identifying the frame with the highest PD level from a package of four different frames extracted from the same cine-loop.

**Materials and methods**

A Continuous Quantitative Color & Power Doppler Assessment technique has been developed (CQCPD – Retina® software) in order to capture the quantitative aspects of vascular images at the same frame rate the machine capture the video clips. The technique was based on the video clips recorded during the US examination and analyzed later with the proprietary software (RETINA® – Bucharest, Romania). The software was made compatible with major formats exported by ultrasound machines.

The technique requires an operator pre-definition of the ROI. The same ROI is used then for every subsequent frame analyzed in each video clip. The ROI should be defined in the first frame both in the freehand (free contour) and in square (square contour) technique (fig 1-3). In the subsequent frames the computer automatically analyses the proprieties of the same frame counting red, blue, and

<table>
<thead>
<tr>
<th>FRAME</th>
<th>Red</th>
<th>Blue</th>
<th>Total</th>
<th>Red + Blue</th>
<th>% color</th>
</tr>
</thead>
<tbody>
<tr>
<td>000.png</td>
<td>180</td>
<td>0</td>
<td>100529</td>
<td>180</td>
<td>0.0017905</td>
</tr>
<tr>
<td>001.png</td>
<td>515</td>
<td>39</td>
<td>100529</td>
<td>554</td>
<td>0.0055108</td>
</tr>
<tr>
<td>002.png</td>
<td>529</td>
<td>47</td>
<td>100529</td>
<td>576</td>
<td>0.0057297</td>
</tr>
<tr>
<td>003.png</td>
<td>732</td>
<td>216</td>
<td>100529</td>
<td>948</td>
<td>0.0094301</td>
</tr>
<tr>
<td>004.png</td>
<td>753</td>
<td>219</td>
<td>100529</td>
<td>972</td>
<td>0.0096689</td>
</tr>
<tr>
<td>005.png</td>
<td>1024</td>
<td>269</td>
<td>100529</td>
<td>1293</td>
<td>0.012862</td>
</tr>
<tr>
<td>006.png</td>
<td>1177</td>
<td>317</td>
<td>100529</td>
<td>1494</td>
<td>0.0148614</td>
</tr>
<tr>
<td>007.png</td>
<td>1183</td>
<td>318</td>
<td>100529</td>
<td>1501</td>
<td>0.014931</td>
</tr>
<tr>
<td>008.png</td>
<td>1263</td>
<td>372</td>
<td>100529</td>
<td>1635</td>
<td>0.016264</td>
</tr>
<tr>
<td>009.png</td>
<td>1382</td>
<td>399</td>
<td>100529</td>
<td>1781</td>
<td>0.0177163</td>
</tr>
<tr>
<td>010.png</td>
<td>1399</td>
<td>413</td>
<td>100529</td>
<td>1812</td>
<td>0.0180246</td>
</tr>
<tr>
<td>011.png</td>
<td>1566</td>
<td>486</td>
<td>100529</td>
<td>2052</td>
<td>0.020412</td>
</tr>
<tr>
<td>012.png</td>
<td>1593</td>
<td>509</td>
<td>100529</td>
<td>2102</td>
<td>0.0209094</td>
</tr>
<tr>
<td>013.png</td>
<td>1775</td>
<td>505</td>
<td>100529</td>
<td>2280</td>
<td>0.02268</td>
</tr>
<tr>
<td>014.png</td>
<td>1854</td>
<td>618</td>
<td>100529</td>
<td>2472</td>
<td>0.0245899</td>
</tr>
<tr>
<td>015.png</td>
<td>1859</td>
<td>624</td>
<td>100529</td>
<td>2483</td>
<td>0.0246993</td>
</tr>
</tbody>
</table>
white pixels. The results are exported in an excel file including the number of total pixels from the analyzed area (Table I). The user has the option to clean data from those pixels that were rather considered to be artifacts (not valuable information) – e.g. isolated unique colored pixels. In addition, when ROI is known (was previously measured by the machine software), RETINA® offers information about the areas covered by various colored pixels (red, blue, non-grey etc.). Excel data might be then analyzed by any statistical software (in this case we used SPSS 21 – IBM – USA).

The development of RETINA® included summarization the “incoming” and “outgoing” signals and separation of the background noise from useful information. The RGB channels have been used to compute three values: 1. Luminance (Y) – representing the total amount of light passing through a particular area, weighted by the effect of each wavelength on a typical human observer; 2. Chroma (C) - representing the “colorfulness” relative to the brightness of a similarly illuminated white; 3. Hue (H) being an attribute of a visual sensation according to which an area appears to be similar to one of the perceived colors: red, yellow, green, and blue, or to a combination of two of them.

For RGB color spaces that use the ITU-R BT.709 primaries (or sRGB, which defines the same primaries), relative luminance was calculated from linear RGB components [1]:

\[ Y = 0.2126 R + 0.7152 G + 0.0722 B \]

The formula reflects the luminous efficiency function which describes the average spectral sensitivity of human visual perception of brightness: green light contributes the most to the intensity perceived by humans, and blue light the least.

For the chroma and hue values, we used the formulas:

\[ H = \text{atan2}(\beta, \alpha) \]
\[ C = \sqrt{\alpha^2 + \beta^2} \]
\[ \alpha = 1/2 \times (2R - G - B) \]
\[ \beta = \sqrt{3/2} \times (G - B) \]

Using these values, we separated the useful information from the background noise by considering “background noise” as all portions of the image with low chroma and low luminance (i.e. small values for Y and C). Afterwards, the remaining parts of the image were classified as either “incoming blood” (relatively high chroma and luminance, yellow-red hue), “outgoing blood” (relatively high chroma and luminance, green-blue hue) or “turbulence” (high luminance, low chroma). This concluded the image segmentation.

In order to allow a facile definition of ROI we decided to offer the operator both the option of a simple, rectangular ROI – or a more complex, customized region (“freeform”). For the customized region, we recorded it as a closed polyline. To determine the interiors of closed polyline, we used an algorithm based on the number of lines crossed by a line that starts from outside the area. Having computed the ROI indicated by the operator, the program performs the segmentation of that region into “background”/“incoming blood”/“outgoing blood”/“turbulence”, displays this information back to the operator, and saves it in memory for later reuse. In this way, the operator can analyze a sequence of images and save all the results in a spreadsheet at the end, for further mathematical analysis. As an additional improvement, the program performs connected-component analy-
and allows the operator to filter-out small components (those that are just 1, 2, or 3 pixels wide), in order to eliminate the artifacts that are introduced by quantization in heavily compressed images (e.g. on some compressed jpeg images). In rectangular ROI mode, the “turbulence” related pixels are computed but not included in the statistics due to a difficult separation from bone contour. In the freeform ROI mode, the bone could be totally excluded and these pixels might be used for statistics. The freeform ROI mode can compute the number of pixels generated by reflection artifacts, too. In this way, the operator is able to clear the data from false Doppler signals.

**Evaluation exercise**

Twenty anonymized cine-loops recorded using standardized scans (dorsal carpal longitudinal according EULAR standard [13]) from the database of a tertiary rheumatology unit were included based on previous patient consent; one region of interest (ROI) was marked on each scan and CTRatio (Power Doppler) was computed for each frame of the loop by using RETINA® software. From each loop 4 different frames were extracted: the one with the highest level of the PD signal (highest CTR-named CTR100 that represents in the same time the WCSI) and 3 frames with 5%, 10%, and respectively 20% lower level of CTR (named CTR95, CTR90 and CTR80). This produced twenty distinct packages each package containing one CTR100 – WCSI. All frames from each package were randomly presented to six experienced and six in training sonographers; the request was to select the frame with the highest CTR from each package. A similar exercise was performed with CTRs decreasing in steps of 2% (CTR100, CTR98, CTR96, CTR94) involving twenty packages, too. The exercises took place during 2015 in a tertiary rheumatological university clinic where all ethical aspects related to research on humans were properly considered; aspects of ethics in industrial research were not considered applicable and not addressed.

**Results**

The first exercise requested less abilities from the sonographers as the difference in Doppler signal was larger (5%) than in the 2nd exercise (2%) and the results confirmed this: in the first exercise the CTR100 (WCSI) was correctly identified in 79.1% cases while in 15.8% of cases CTR95 were wrongly indicated as WCSI and in 5.5% of cases the CTR90 was selected. In the 2nd exercise the subjects indicated as WCSI the following frames: CTR100 (the real WSCI) in 67%, CTR98 in 18.3%, cases, CTR 96 in 11,6% and CTR94 in 3.1% of cases. A two sample t-test confirmed this significant difference: t = 10.05, p<0.001.

The experienced sonographers performed better than in-training examiners in both exercises. The experienced sonographers made a correct selection in 95% to 75% of cases and the in-training sonographers made a correct selection in 90% to 50% of cases in the 1st exercise. In the 2nd exercise both groups performed less correct: the experienced sonographers made the correct selection in 90% to 65% of cases and in-training sonographers in 75% to 45% of cases in the 2nd exercise. The interobserver agreement between experienced and in-trainer evaluators was measured by computing kappa statistics (with 0.41-0.60 moderate agreement, 0.61-0.80 – substantial agreement and 0.81-1 almost perfect agreement [14]). For the 1st exercise the agreement was 0.78 and in the 2nd exercise the agreement was 0.4.

**Discussions**

In addition to US grey scale that offers to the rheumatologist the opportunity to assess the structural changes of the musculoskeletal system, the US Doppler can depict local inflammation. This technology is based on the detection of red cell moving; a higher number of moving cells and faster speed produce a larger number of colorized pixels and different color on the MSUS image. The sonographer has the opportunity to visualize the “inflammation” and this gives him an idea of what is happening in the region of interest. However, the sonographer should be capable not only of affirming the presence or absence of inflammation but also of quantifying the level of inflammation and so, assessing the effectiveness of a therapeutic regime [10]. This is not such an easy process as the number of vessels does not correspond with the number of color pixels visualized by Doppler ultrasound but rather with the area of the already existing vessels [14,15]. A number of studies already have linked the Doppler information with synovial activity in arthritis. Significant improvement in PDUS parameters have been found in parallel with changes in clinical and laboratory indices of disease activity [15-17]. Therapeutic effect seems to be also parallel with Doppler activity: intra-articular administration of corticosteroids has been proven to change the fraction of color pixels and the technique has been proposed to be used for quantitative measurements of the blood flow [16]. Naredo et al [15] demonstrated in quantitative analyses the in-pannus vascularization of small finger joints in patients with rheumatoid arthritis a decrease of the color signal after treatment with soluble tumor necrosis factor α receptor (etanercept). These studies support the PDUS as a valid method for monitoring response to anti-inflammatory and biological therapy in
The issue of selecting the best Doppler frame, the one with the highest number of colored pixels is still unsolved. The selection is more difficult when the focus is represented by one part of the whole image, the so-called Region of interest (ROI): the colored pixels situated outside this ROI have frequently a slightly different pattern and the WCSI for the whole image and the WCSI for a specifically ROI are not the same. In addition, the rheumatologist is particularly interested by the blood flow that is not produced by feeding vessels – with all efforts done for a proper position of the transducer these vessels are not always excluded from the ROI and could lead to a selection bias. There are always differences between computer and human abilities in performing different tasks – human speed, accuracy or analyses are, unfortunately, inferior. The question is related to the amplitude and implications of this difference – is this difference minor and we can fully rely on human performances or is important and a computerized solution should be developed and implemented? In our work, we demonstrated the amplitude of this human vs. computer difference. Our work also demonstrated that experienced sonographers perform better but still not enough to match the software/computer performance. In our opinion we have demonstrated the presence of an end of road in this direction – it seems quite unproductive to try to improve the sonographer’s abilities in the selection of WCSI in Quantitative Doppler analysis as long as the computer performs better. The results of our study promote a better standardization in US Doppler assessment and a more precise description of Doppler changes; these can further promote the utilization of Doppler US in clinical trials (that require a very high level of standardization of the assessment techniques and benefit from instruments with higher discriminative capabilities). In real life, such an assessment may help for a more rapid forecast of the efficacy of a treatment [1] by reducing the time between two evaluations from weeks to days or hours.

There are some limitations of our study: the number of sonographers is not so large – it might be interesting to see what the results of such exercise are when a larger group of sonographers is involved. The cine-loops resulted from the evaluation of the same type of joint and the same incidence – our results should be demonstrated on different scenarios: different joints, tendons etc. All examinations have been performed with the same machine and same settings – different machines (including their different optimal settings) should be also tested. Finally, we did not offer the sonographers simultaneous data regarding cardiac activity – as we emphasized some advanced machines do this and they can help the sonographers in their decision.

**Conclusions**

Using the computerized analyses as the gold standard, we demonstrated the imperfect ability of experienced sonographers to identify the WCSI from a package of images that are clearly different; our work demonstrated the existence of a significant heterogeneity across sonographers regarding their ability to identify WCSI even from a small group of frames. The ability is even more questionable when the differences between grouped pictures are small. When the quantitative Doppler signal is assessed in the same patient at different time points this heterogeneity might induce errors. When computerized analysis is feasible it might offer more precise information than the manual selection. Additional studies are required to compare the semi quantitative method to our continuous quantitative method (CQCPD) in real patient exercises.

**Conflict of interest:** none

**References**


