

Is the Doppler Ultrasound Score More Reliable than the Grey Scale Ultrasound Score in Assessing the Rheumatoid Arthritis Patient's Joints?

Copotoiu Monica¹, Georgescu Lia¹, Poór G², Copotoiu Ruxandra³

¹ Department of Rheumatology, University of Medicine and Pharmacy, Tirgu Mures, Romania

² 1st Department of Rheumatology, National Institute of Rheumatology and Physiotherapy, Budapest, Hungary

³ Department of Anaesthesia and Intensive Care, University of Medicine and Pharmacy, Tirgu Mures, Romania

Background: Rheumatoid arthritis (RA) is a complex disease with multiple faces, one of them being its activity. At the moment, there are multiple-choice tools to assess the activity of the disease, no one gaining the complete status of “gold standard”. In the evaluation of the disease activity biological and non-biological methods are used and interposed. The methods are more or less subjective, thus the validation of measurements in RA is highly relevant.

Objective: The main objective of the study was to evaluate the reliability of the Doppler Ultrasound score versus the Grey scale ultrasound score in the assessment of the disease activity in rheumatoid arthritis patients.

Methods: A number of 66 patients diagnosed with rheumatoid arthritis were assessed with the B-mode and consequently Colour Doppler ultrasound score system. The joints examined were the second and the third metacarpal (MCP) and proximal interphalangeal joints (PIP) on the dorsal and ventral side. Two thousand twelve images were stored and evaluated on both scales. Data were analysed with GraphPad Prism 5.0. In order to assess the validity of the ultrasound methods used in the study as non-biological markers for the activity of rheumatoid arthritis Pearson correlation, Bland-Altman method and linear regression were used.

Results: A positive correlation was found when analyzed the two methods of assessment of the disease activity in rheumatoid arthritis. A single mismatch was found (Pearson correlation and linear regression) at the level of the second metacarpal on the ventral and right side when compared the Doppler ultrasound versus the Grey scale score, but we presumed that this is due to the particularity of the individual outcome. The Bland-Altman method showed that the Grey scale of quantification overrated the scores versus the Color Doppler Ultrasound scale.

Conclusions: The semi-quantitative Color Doppler method which assesses the intra-joint activity in rheumatoid arthritis patients was validated in our research to be further used. Even though none of the statistic methods used for validating the Doppler score in our study showed significant statistic differences between them, it is for the best to apply all of them for validating any intended method to be used in any study.

Keywords: Color Doppler ultrasound, synovial activity, validation

Introduction

Rheumatoid arthritis (RA) is a complex disease with multiple faces, one of them being its activity. At the moment, there are multiple-choice tools to assess the activity of the disease, none of them gaining the complete status of “gold standard”. In the evaluation of the disease activity biological and non-biological methods are used and interposed.

The methods are more or less subjective, thus the validation of measurements in RA is highly relevant. The assessment of the validity is divided into three categories: the content validity, the criterion validity and the construct validity. All categories are mutually related. The definitions of the validity's categories are not totally stringent.

The content validity known as the face validity follows the requirement of the measurement to expose all the important aspects of the disease assessment. In RA, one of the basic and important aspect is the synovial inflammation, thus the blood flow is increased. Corroborating those data with the presence of angiogenesis it can be easy speculated on the usefulness of Color or Power Doppler ultrasound assessment – the ultrasound Doppler reflects the activity of the disease. The Grey-scale (the B mode) ultrasound emphasizes on synovial hypertrophy and erosions, so the content validity is present when combining the two methods.

The criterion validity refers to the correlation of a mea-

surement with an estimated test or a “gold standard” test for the assessment of a certain condition. An important part of the criterion validity is testing the accuracy of the method. The accuracy is defined as the assay precision, on what extend it is affected by random errors or systematic errors. The criterion validity represents the objective, practical term of validity.

In rheumatology, the construct validity is confounded with the criterion validity. It estimates the capacity of the assay to measure what theoretically is supposed to measure.

Methods

The inflammatory and the destructive changes in RA can be visualised by ultrasound. Two types of US modes are used: the grey scale and the Doppler US. The grey scale is capable of depicting the morphological changes. The Doppler US displays blood flow in the tissues. The increased blood flow is a part of the inflammatory process, thus the amount of Doppler activity can be an indirect measure of inflammation [1,2,3].

The grey scale is able to show us a synovial hypertrophy and erosions. There are a number of definitions for US pathology, including the OMERACT one. The scale used for B imaging was a semi-quantitative one. The semi-quantitative scale was grade from 0 to 3 as followed: 0 – normal or no synovial thickening/effusion, 1 – mild thickening/

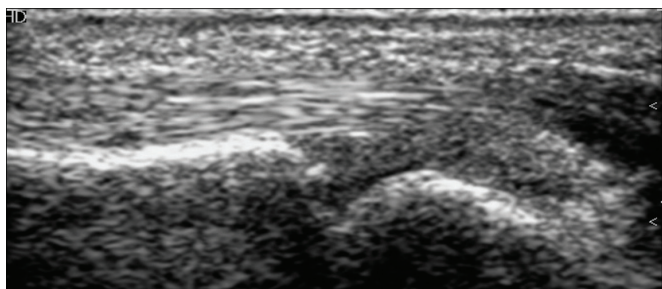


Fig. 1. Grey scale MCP 2 ventral side

effusion, 2 – moderate thickening/effusion, 3 – intense thickening/effusion [4–9].

On the monitor, in the grey scale the tissues are marked in different grey tones, the different nuances of grey representing the reflective ability of the tissue. The bones are good reflectors and their surfaces are shown in white, and the synovial fluid is shown in black, as a lack of reflection on the monitor. An echo is generated when a US beam is crossing two tissues with different acoustic impedance. The acoustic impedance (Z) is the results of the density multiplied by the speed of the sound ($= Z$) and the reflective ability depends on it.

The US Doppler is able to differentiate between a thickened synovium with inflammation and thickened synovium due to previous inflammatory attacks (no Doppler activity) [10,11].

In the semi-quantitative US Doppler score the assessment is done on a subjective basis by the percentage of colour pixels in ROI (region of interest). The 4 grade semi-quantitative scoring system is the following: 0 – no Doppler signals/no blood flow, 1 – single Doppler signals/mild blood flow, 2 – various, confluent Doppler signals/moderate blood flow, 3 – confluent Doppler signals with more than half of the visible synovium showing Doppler signals/intense blood flow.

A Phillips HD7 (High Definition) ultrasound machine was used. The ultrasound exam was performed with a linear array transducer with variable frequency from 3 to 12 MHz (L12-3/38mm, HD7, Phillips, Bothell, WA, USA). The longitudinal scan was used according with the standards imposed in the musculoskeletal ultrasound guidelines for rheumatology.

The patient was positioned in a proper way in order to make the examination comfortable for both the subject and the operator, regardless of the examined anatomical region to ensure as little pressure as possible from the trans-

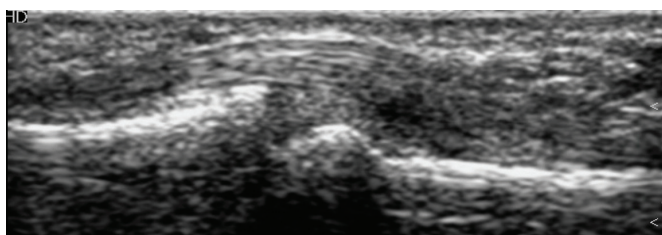


Fig. 3. Grey scale PIP 2 ventral side

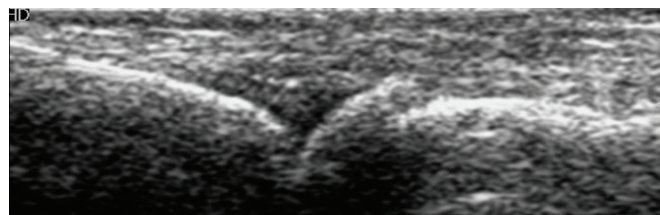


Fig. 2. Grey scale MCP 2 dorsal side

ducer (e.g. the joints were assessed in an extended position with the comfortable resting of the scanned extremity). The angle of the ultrasound beam must be perpendicular to the examined structure; due to the fact the artefacts are easily reproducible in the musculoskeletal ultrasound. The gain was set to a colour per unit (CPU) of 82%, which was selected by manual elevation of the Doppler US gain level until the colour box was almost uniformly filled with the first indication of colour and with the only the minimum of the next highest signal just beginning to appear.

Our prospective study included a number of 66 patients diagnosed with rheumatoid arthritis. The inclusion criteria were: subjects diagnosed with RA according to the new 2010 EULAR/ACR criteria for RA, clinically medium-high activity (DAS28 >3.2) RA patients with non-response or low synovial activity measured by Power Doppler ultrasound (OMERACT criteria), subjects without cardiovascular and metabolic diseases known/recorded, subjects who consent to the study, subjects aged between 18–75 years. The patients were assessed with the B-mode and consequently Color Doppler ultrasound score system. The joints examined were the second and the third metacarpal (MCP) and proximal interphalangeal joints (PIP) on the dorsal and ventral side. Two thousand twelve images were stored and evaluated on both scales.

The statistical analysis of the data was performed with GraphPad Prism 5.0 software (San Diego, California). In order to assess the validity of the ultrasound methods used in the study as non-biological markers for the activity of the rheumatoid arthritis the Pearson correlation, the Bland-Altman method and the linear regression were used. A $p < 0.05$ was considered to be statistically significant.

Results and discussion

Grey scale images for MCP and PIP for ventral and dorsal approach are depicted in figures 1–4. Corresponding Color Doppler images are presented in figures 5, 6.

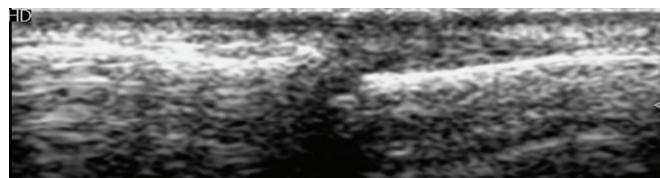


Fig. 4. Grey scale PIP 2 dorsal side

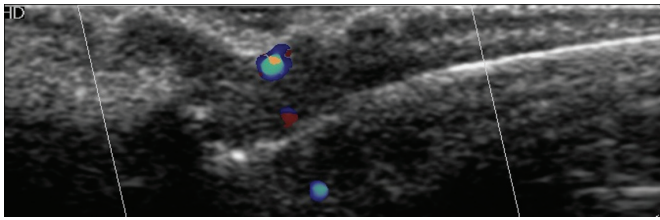


Fig. 5. Color ultrasound MCP 2 ventral side

The Pearson’s correlation

The ultrasound measurements by Grey scale and Color Doppler scale were analyzed applying the Pearson’s correlation, in order to acknowledge the differences (if it exists) between the two non biological methods chosen to measure the activity of the disease (rheumatoid arthritis) and to validate the use of only one method over the other (in our case the Color Doppler method).

A positive correlation was found on the second metacarpal joint on the dorsal side, between the semi-quantitative methods chosen for assessment (the Grey scale vs. the Color Doppler scale) ($p = 0.0252$, 95% CI: 0.03581 to 0.4852) (Figure 7).

This correlation was found in all the joints ($p < 0.05$) on both sides examined with one exception – it didn’t apply to the ventral side ($p: 0.1144$, 95% CI: -0.04484 to 0.4054) on the second metacarpal joint on the right limb (Figure 8).

The only mismatch found was at the ventral level of the second MCP. We could only stipulate that that result is due to the activity of the disease and couldn’t be taken as bias.

A lot of the agreement studies published have shown that using the t test or the Pearson’s correlation is flawed when measuring the agreement or detecting the bias. So the Bland-Altman method proposed graphical techniques in order to analyze method comparison and to validate studies.

The Bland-Altman method

In medical research it is often required or needed to compare two methods of measurement (usually a new one

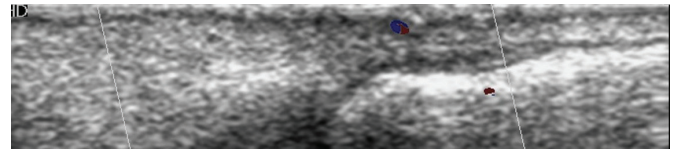


Fig. 6. Color ultrasound PIP 2 dorsal side

versus the established method – the so-called “gold standard”) to determine whether the methods can be used interchangeably or the new method can replace the “gold standard”. The methods compared are known that each provides some errors in their measurements [12–17].

Bland and Altman first described the Bland – Altman method, in the eighties as a method of data plotting in analyzing the agreement between two different assays. The original Bland-Altman method was developed for two sets of measurements done on one occasion (independent data), and so this approach is not suitable for repeated measures data. In the latest years, the medical publication received increased numbers of studies using the Bland-Altman method (the method was cited in more than 11,500 articles till 2007) or similar ones (from 8% in 1986, 14% in 1995 ended at 31–36% in 2007) [12].

It is used a graphical method to plot the difference scores of two measurements against the mean for each subject and if the new method agrees in accepted terms with the established one, the first may be replaced. The Bland-Altman plot (popularised in medical statistics) is widely known as Tukey Mean-Difference Plot. The x-axis is the mean of the two measurements (best guess of the “right” result) and the y-axis is the difference between the two measurement differences. The plot is chosen according to the values of the difference. It may be possible to plot the difference, the ratio or the percent difference. If the difference gets larger as the average gets larger, it can make more sense to plot the ratio or the percent difference. The Bland-Altman plots are generally interpreted informally, without further analyses [10].

The chart highlights the certain types of anomalies in the assays – that one of the methods over-estimates high

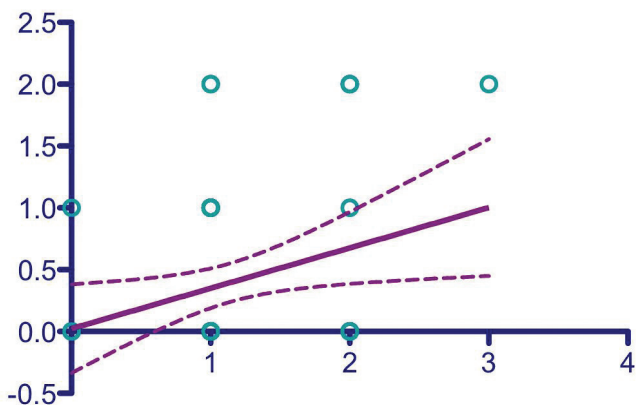


Fig. 7. Correlation for MCP 2 dorsal right side between the grey scale and Doppler

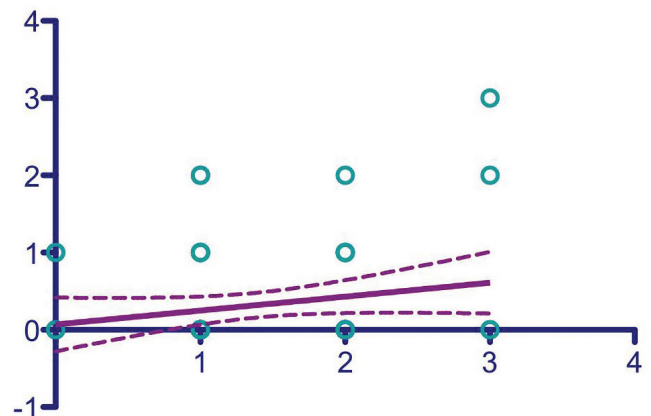


Fig. 8. Correlation for MCP 2 ventral right side between the grey scale and Doppler

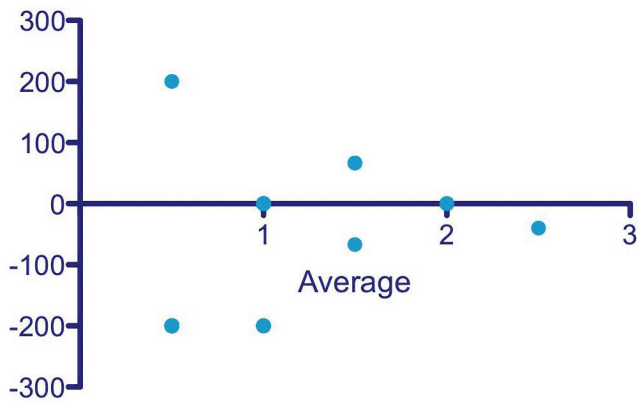


Fig. 9. Bland-Altman analysis for MCP 2, grey scale versus Doppler, dorsal side

values and under-estimates low values. If the plots are scattered all over the place, above and below zero that there is no consistent bias of one approach versus the other – it can't exclude completely the hidden bias. Still, the Bland-Altman method remains a good first step for exploring the data [18,19, 20].

The key word in Bland-Altman method and other methods comparison is agreement. In the Bland-Altman method it is common to compute the limits of agreement specified as bias \pm 1.96 SD (average difference \pm 1.96 standard deviation of the difference) and the estimation of confident intervals for the bias. Those results are often omitted in research papers.

The 95% limits of agreement are to visually judge the data on how the two methods agree. The smaller the range is, the better the agreement is. The size of the average discrepancy between methods (the bias) must be interpreted clinically. Whether the observed discrepancy is large enough or not is a clinical enquire not a statistical one.

The next step is to question the existence of a trend (e.g. Does the differences between the two methods tend to get smaller or larger as the average increases?). The final step is to verify if the variability is consistent across the graph.

A Bland-Altman method was applied at the start of the study in order to decide the best method to assess by ul-

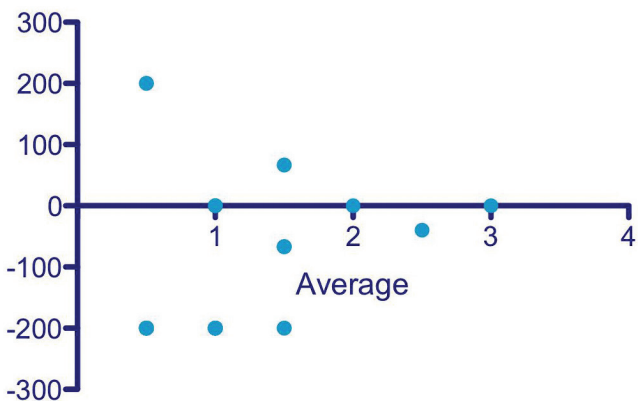


Fig. 11. Bland-Altman analysis for MCP 2, grey scale versus Doppler, ventral side

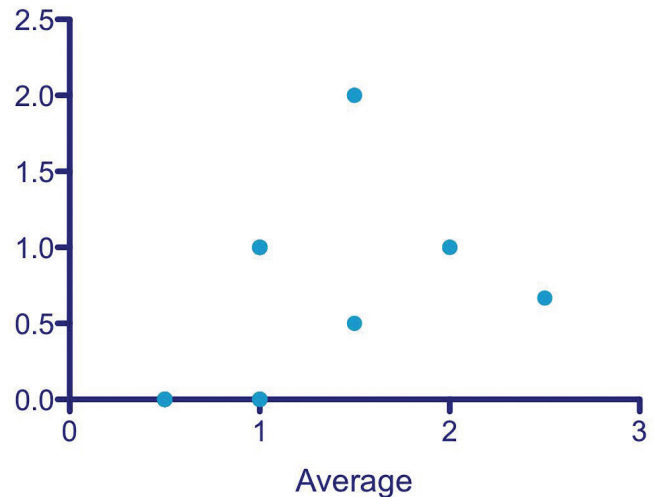


Fig. 10. Bland-Altman analysis for MCP 2, grey scale versus Doppler, dorsal side

trasound the activity of the disease in rheumatoid arthritis patients. The methods compared were the Grey-scale ultrasound versus the Doppler ultrasound in small joints on the dorsal and ventral part.

The measurements at the level of the second metacarpal joint using the Grey scale method versus the Doppler scale method on the dorsal side after being submitted to the Bland – Altman method showed an over-estimated data by using the Grey scale versus Doppler scale, even if the bias are not high statistically (Bias: 0.3306, SD of bias: 0.5327, 95% limits of agreement: -0.7135 – 1.375). The plot used was the percent difference and ratio (Bias: -131.8, SD: 103.8, 95% limits of agreement: -335.3 – 71.65) (Figure 9, 10).

The measurements at the level of the second metacarpal joint using the Grey scale method versus the Doppler scale method on the ventral side after being submitted to the Bland – Altman method showed an over-estimated data by using the Grey scale versus Doppler scale, even if the bias are not high statistically (Bias: 0.2077, SD of bias: 0.4621, 95%

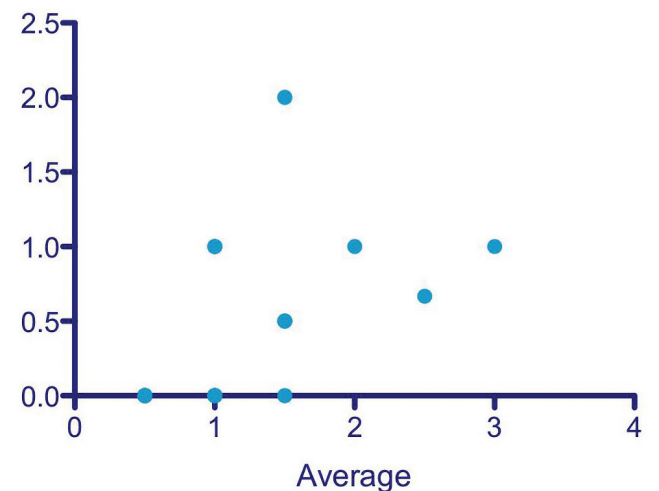


Fig. 12. Bland-Altman analysis for MCP 2, grey scale versus Doppler, ventral side

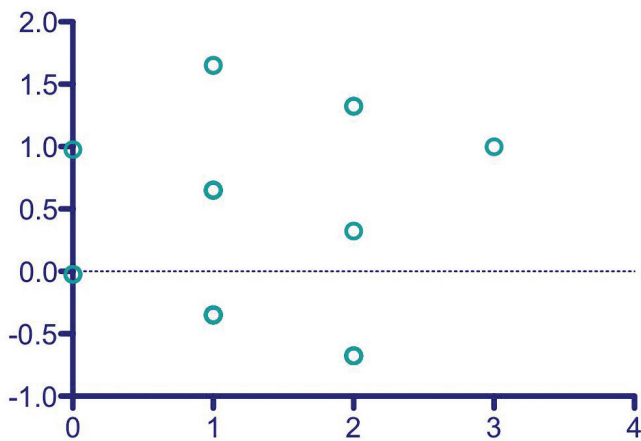


Fig. 13. Linear regression for MCP 2, dorsal right side, grey scale vs. Doppler

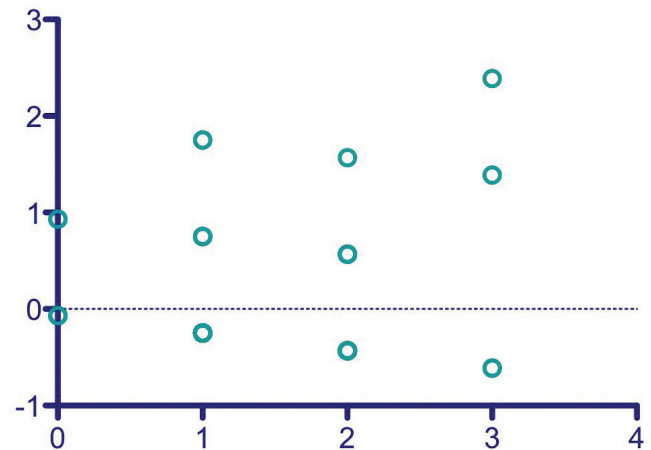


Fig. 14. Linear regression for MCP 2, ventral right side, grey scale vs. Doppler

limits of agreement: $-0.6981 - 1.13$). The plot used was the percent difference and ratio (Bias: -148.8 , SD: 101.1 , 95% limits of agreement: $-346.9 - 49.37$) (Figure 11, 12).

It was observed not only that data were similar to the dorsal measurements at this level, but the fact that the larger the average became the smallest the difference was.

The general trend observed was the smallest the differences were the largest the averages were. The second general observation was the trend of the Grey-scale score to over-estimate values. All the trends could be applied generally. All the biases obtained (nearby zero) supply the idea that there is not a big difference between the two semi-quantitative methods studied and applied.

Linear regression

In 2000, Hopkins debated the flaws (the hidden bias) of the Bland-Altman's graphical method and proposed the use of simple linear regression applied between the two measurements. He emphasized on the nature on the random error in measurements provided by the instruments. The Bland and Altman group recommended the use of a log transformation of the data points [21].

Validating a method is important, so after performing the Bland-Altman method in order to avoid as much as possible the biases, the linear regression was applied to the measurements.

The measurements applied at the second MCP on the dorsal side on the right showed a positive correlation ($p < 0.005$). The results exerted from evaluating the same joint on the ventral side didn't show a positive correlation ($p > 0.005$) (Figure 13, 14). The final results after using the logistic regression were similar with the two ones applied before (the Pearson correlation and the Bland-Altman method).

Conclusions

The semi-quantitative Color Doppler method for assessing the intra-joint activity in rheumatoid arthritis patients was validated in our research and can be further used on.

Even though all the three statistic methods used for the validation of the Doppler score in our study didn't show significant statistic differences between them, it is for the best to apply all of them in order to validate any intended method to be used in any other study.

References

- Albrecht K, Muller-Ladner U, Strunk J – Quantification of the synovial perfusion in rheumatoid arthritis using Doppler ultrasonography. *Clin Exp Rheumatol* 2007, 25: 630–638.
- Schmidt WA – Doppler sonography in rheumatology. *Best Pract Res Clin Rheumatol* 2004, 18: 827–846.
- Ellegaard K – Accuracy of ultrasound Doppler measurements in assessment of disease activity in patients with rheumatoid arthritis, PhD thesis, accessed from www.vbn.aan.dk on August 22, 2011
- Albrecht K, Grob K, Lange U, Muller-Lander U, Strunk J – Reliability of different Doppler ultrasound quantification methods and devices in the assessment of therapeutic response in arthritis, *Rheumatology* 2008, 47: 1521–1526.
- Fiocco U, Ferro F, Vezzu M et al. – Rheumatoid and psoriatic knee synovitis: clinical, grey scale, and power Doppler ultrasound assessment of the response to etanercept. *Ann Rheum Dis* 2005, 64: 899–905.
- Filippucci E, Farina A, Carotti M, Salaffi F, Grassi W – Grey scale and power Doppler sonographic changes induced by intra-articular steroid injection treatment. *Ann Rheum Dis* 2004, 63: 740–743.
- Filippucci E, Iagnocco A, Salaffi F et al. – Power Doppler sonography monitoring of synovial perfusion at the wrist joints in patients with rheumatoid arthritis treated with adalimumab. *Ann Rheum Dis* 2006, 65: 1433–1437.
- Ribbens C, Andre B, Marcelis S et al. – Rheumatoid hand joint synovitis: gray-scale and Power Doppler US quantifications following anti-tumor necrosis factor-alpha treatment: pilot study. *Radiology* 2003, 229: 562–569.
- Salaffi F, Carotti M, Manganelli P et al. – Contrast-enhanced power Doppler sonography of knee synovitis in rheumatoid arthritis: assessment of therapeutic response. *Clin Exp Rheumatol* 2004, 23: 285–290.
- Taylor PC, Steuer A, Gruber J et al. – Comparison of ultrasonographic assessment of synovitis and joint vascularity with radiographic evaluation in a randomized, placebo-controlled study of infliximab therapy in early rheumatoid arthritis. *Arthritis Rheum* 2004, 50: 1107–1116.
- Naredo E, Collado P, Cruz A et al. – Longitudinal power Doppler ultrasonographic assessment of joint inflammatory activity in early rheumatoid arthritis: predictive value in disease activity and radiologic progression. *Arthritis Rheum* 2007, 57: 116–124.
- Myles PS – Using the Bland-Altman method to measure agreement with the repeated measures. *Br J Anaesth* 2007, 99: 309–311.
- Opdam H, Wan L, Bellomo R – A pilot assessment of the FloTrac (TM) cardiac output monitoring system. *Intensive Care Med* 2007, 25: 459–463.
- Niedhart DJ, Kaiser HA, Jacobson E, et al. – Inpatient reproducibility of the BISxp monitor. *Anesthesiology* 2006, 104 (2): 242–248.

15. Anderson RE, Sartipy U, Jakobsson JG – Use of conventional ECG electrodes for depth of anaesthesia monitoring using the cerebral state index: a clinical study in day surgery. *Br J Anaesth* 2007, 98: 645–648.
16. Button D, Weibel L, Reuthebuch O et al. – Clinical evaluation of the Flo/Vigileo system and two established continuous cardiac output monitoring devices in patients undergoing cardiac surgery. *Br J Anaesth* 2007, 99: 329–336.
17. Hanneman SK – Design, analysis and interpretation of method-comparison studies. *AACN Advanced Critical Care* 2008, 19: 223–234.
18. <http://stats.stackexchange.com/questions>, accessed on Sep 24, 2011
19. Bland JM, Altman DG – Applying the right statistics: analyses of measurement studies. *Ultrasound in Obstetrics and Gynaecology* 2003, 22: 85–93.
20. Ludbrook J – Statistical techniques for comparing measurers and methods of measurement: a clinical review. *Clinical and Experimental Pharmacology and Physiology* 2002, 29: 527–536.
21. Hopkins WG – Measures of reliability in sports medicine and science. *Sports Medicine* 2000, 30: 1–15.