STUDIA INFORMATICA 2016
Volume 37 Number 3B (126)

Pawel MIELNIK
Section for Rheumatology; Department for Neurology, Rheumatology and Physical Medicine, Helse Førde, Førde, Norway
Marcin FOJCIK
Faculty of Engineering and Science, Sogn og Fjordane University College, Førde, Norway
Marek KULBACKI, Jakub SEGEN
Polish-Japanese Academy of Information Technology, Warszawa, Poland

CHALLENGES IN INTRODUCTION OF ARTIFICIAL INTELLIGENCE IN MEDICAL PRACTICE – A REVIEW OF CLINICAL TRIALS CONCERNING ADAPTATION OF ARTIFICIAL INTELLIGENCE IN MEDICINE [1]

Summary. An interest in Artificial Intelligence [AI] as science is growing in the last years. It has become gradually more used in the medicine. Methodology of development and testing of AI algorithms is generally well established. Use of AI in medicine requires elaboration of standards of its validation in clinical settings. This paper is a review of literature concerning clinical trials on AI adaptation in medicine

Keywords: Artificial Intelligence, diagnostics, standards, clinical testing, validation

PROBLEMY Z WPROWADZANIEM TECHNOLOGII SZTUCZNEJ INTELIGENCJI DO PRAKTYKI MEDYCZNEJ

Streszczenie. Zainteresowanie technologią sztucznej inteligencji nieustannie wzrasta. Również w medycynie znajduje ta technologia coraz częściej praktyczne zastosowanie. Mimo dynamicznego jej rozwoju brakuje nadal standardów dla klinicznej weryfikacji jej skuteczności w praktyce medycznej. Artykuł jest przeglądem publikacji dotyczących badań klinicznych nad zastosowaniem tej technologii.

Słowa kluczowe: sztuczna inteligencja, diagnostyka

[11]The research project MEDUSA has received funding from the Polish-Norwegian Research Programme operated by the National Center for Research and Development under the Norwegian Financial Mechanism 2009-2014 in the frame of Project Contract No. Pol-Nor/204256/16/2013.

1. Introduction

An interest in Artificial Intelligence [AI] as science is growing in the last years. PubMed search reveals more than 5 thousand publications in the AI area per year in 2013 and 2014 (Fig. 1). For comparison, it was only slightly more than 1 thousand in 2000. AI technology is now widely used in a daily life also in routine medical tasks. We understand the "routine medical tasks" as task performed usually by medical staff as part of routine medical practice. AI can play supporting or subsidiary role in that field.

This paper is an effect of considerations of a research project named MEDUSA (Medical Ultrasound and Power Doppler Examinations using Image Processing and Machine Learning Methods) in the context of prototype verification[1]. The main goal of the MEDUSA project is to develop an automated system able to detect and grade synovitis in joints of the hand. Semiquantitative ultrasound with power Doppler is a widely used method of assessing synovitis. Synovitis is assessed by ultrasound examiner using the scoring system graded from 0 to 3, both for synovitis area and vascularisation, separately for each joint. The method requires trained medical personnel and the results can be affected by a human error. We hope that our system for automation of assessments can reduce human dependent discrepancies in the joint evaluations. Concerns described further in this paper arose during analysis of the methods used to verify the results. We have reviewed available scientific literature to assess methods used by researchers for the purpose of evaluation of AI related prototypes. The paper however does not constitute a systematic review of the literature.

2. Results

We have reviewed PubMed database with keywords 'artificial intelligence', 'automated assessment' and 'computer assisted' for the last 3 years. We included papers that describe practical approach to the introduction of Artificial Intelligence into medical practice. We have excluded publications presenting computer science use in basic medical science, laboratory etc. We have focused our research mostly on systems used in medical imaging, as our project is located in this area. Other clinical diagnostic or therapy projects were included if they encompass studies on real life data. We have analysed methods used in assessments of efficacy.

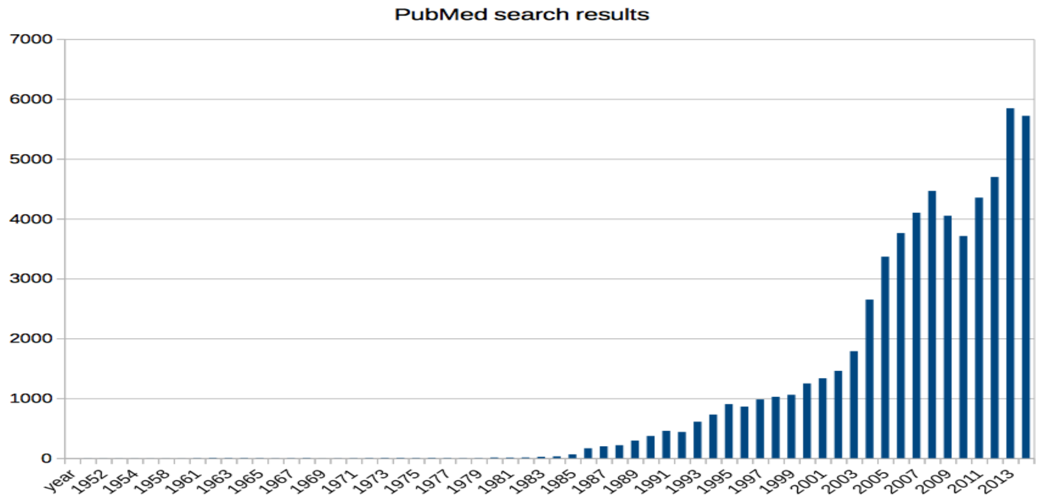

Fig. 1. Results of PubMed search for AI keyword
Rys. 1. Wyniki przeszukiwania bazy PubMed za pomocą słowa kluczowego AI

Many papers have been published about automated image segmentation as a possible supporting technology in radiology. Most of them are related to the central nervous system [CNS] imaging. Both commercial and non-commercial software are available for this type of analysis. An example of the open source software is FreeSurfer suite [2]. FreeSurfer [FS] has a large bibliography of clinical studies which showed that it is a practical, effective and supportive tool in magnetic resonance imaging of the brain [MRI]. FS has become standard for cortical metrics [3]. However, some concerns exist about accuracy and reproducibility of this tool. MacCarthy and colleagues showed that full-automatic measures of cortical thickness differ from manual measurements in some areas. They concluded that further studies are necessary to assess the problem, as some brain areas can be difficult for FS algorithm to segment. Additionally, as it was emphasized in other studies, FS shows larger volume of hippocampus than manual methods [4].

Despite of those discrepancies FS was found useful in clinical trials mainly in chronic, degenerative CNS diseases [5, 6].

Another approach to CNS segmentation was shown by Kim et al. [7]. The authors compared 8 different machine-learning algorithms and found 2 of them the most accurate. They conducted the study on large data set from other clinical trials. The algorithms were compared with manual segmentation. Despite it is not stated in the paper one can assume that manual segmentation was performed in advance.

Comparison of 29 algorithms for computer-assisted dementia diagnosis is presented in work by Bron [8]. The work is the largest and most sophisticated analysis of AI diagnostic framework. 384 subjects from 3 centres were included. The clinical diagnosis was the standard reference in this study. The study population consisted of Alzheimer disease patients, mild cognitive impairment patients and non-dementia subjects. Unfortunately the authors do not provide information about the source of patients for the study (consecutive patients, register source etc.). The highest accuracy (63%) was shown for the algorithm proposed by Sørensen [9].

Segmentation of lymph nodes [LN] in thorax and abdomen in computed tomography [CT] is the subject of the work by Roth and colleagues [10]. The data was manually segmented in advance by a radiologist. Data was randomly divided to the training and testing group. The algorithm used by authors was shown to be more effective than those previously used.

Segmentation algorithms were also used in detecting of anterior irregularity of vertebra [11] and parotid gland demarcation [12].

Automated segmentations was studied in patients with acute cerebral stroke [13]. The researchers segmented manually stroke area in 37 clinical cases. They identified two algorithms as the most accurate in this task (Random Decision Forest and a Convolutional Neural Networks), however none of them could outperform manual segmentation. One of the reasons was relatively large intraobserver variability.

Other group of works on AI in image diagnostic deals with texture analysis. Ardakani and colleagues applied texture analysis to differential diagnostics of thyroid tumours [14]. Authors assessed 26 benign and 44 malignant tumours by ultrasound. They used texture method that showed high specificity and sensitivity in differentiation of lesions. It was compared to FNAB.

Global and local analysis of lung pattern in CT-scans from patients with COPD was used to distinguish subjects susceptible to exacerbation from those non-susceptible [15]. CT scans from 20 patients were analysed. All patients were in the advanced disease stage [GOLD 3]. Half of the patients did not have any exacerbation within one year preceding the assessment. The other had more than 6 exacerbations in the same period of time. The cohort was selected manually to construct the study model with the largest possible difference between groups. The main limitation of the work was, as authors admitted, relatively small size of the analysed population.

Mollina and colleagues used texture analysis to describe prostate carcinoma in 12 patients [16]. They used MR images from patients with known cancer taken before any treatment. MR was randomly blinded. The method used was successful in demarcation of carcinoma regions.

Method that uses visual analysis to predict semantic descriptive term was presented in a paper by Depeursinge et al. [17]. They used a database of 74 liver lesions visualised on CT scans. In the first phase of the study a radiologist described the lesions with standard terms and circumscribed it on the CT scan. The database was used to learn the system to detect the lesion and give ontological results. As the work is in early stage no further results are available.

Comparison of automated texture analysis of prostate MRI and histological prostate carcinoma gradation [Gleason score – GS] was presented in another work [18]. It differs from the other presented studies as it compares two different methods. The aim of the study was to separate the patients assessed as being at GS over 6 from those assessed as GS of 6 and below. Automatic MRI texture assessment achieved accuracy of 92-93% depending on the involved prostate region.

AI methods have also been used in trials in many others medical science areas, not only radiological image analysis.

An automated system for diagnosis of retinopathy of prematurity (ROP) was presented by authors from international consortium [19]. The system task was to classify images in 3 groups: plus, preplus, or normal. “Plus disease” is defined as abnormal dilation and tortuosity of the blood vessels during ROP that may go on to total retinal detachment. The authors presented in details a clinical validation of the system. They used images of 77 prematurely born infants. The performance of the system was evaluated by a k-fold cross-validation procedure. Additionally they compared results of the system assessment with a gold standard. The gold standard was developed as a consensus between 3 experts. The experts assessed images independently. If they assigned an image to different groups the consensus was established in discussion. The reference standard consisted of concordant expert assessment. Kappa statistic was performed to compare results of expert and automatic assessment to the gold standard reference. The system assessment result was not significantly different from expert assessment.

Fergus at al discriminated between seizure and non-seizure EEG records in 24 subjects with known epilepsy [20]. They trained the algorithm with pattern from all 24 patients. The study was focused on detecting seizure records in each patient recording. Results were compared with manual seizure pattern recognition. Another paper on EEC analysis describes sampling methods in recognition of pathological patterns [21]. Author used publicly available EEC database and no prospective evaluation was present. Preliminary results are however promising.

Decision-making support system for nutrition disorder after bariatric surgery was presented by Cruz and colleagues [22]. In the first step they developed a gold standard in collaboration of experts by analysing a nutritional status of 15 patients. After the system training phase the authors compared assessment results with the gold standard and expert opinions. It is not stated in the paper whether the validation was made on the same group of patient data. The high specificity and sensibility was shown.

The support vector machine was used to identify prognostic factors in patients with nasopharyngeal carcinoma [23]. This is not exactly an example of use of AI technology in routine “medical task” but rather an epidemiological study. The authors identified a group with longer median survival time (38 months) and shorter median survival time (13.8 months). All patient were in M1 stage (distant metastases present). Thirty haematological and eleven clinical markers were identified and analysed.

Mehta and colleagues presented in theirs paper a prototype of a wearable monitoring system for diagnostic of voice disorders [24]. Automated system using pattern recognition was able to diagnosed correctly 74% of patients in comparison to clinical diagnosis.

One of the first areas for computer analysis was electrocardiography [ECG]. The first ever publication regarding this subject is from 1964; unfortunately we had no opportunity to find the original paper [25]. The newest publication on this topic came from Taiwan[26]. The researchers developed an automated system for assessment of the ECG data coming from a telemedicine centre. The ECGs from the first year of the study were used to the system training, those from consecutive years for testing. Very high specificity and sensitivity was observed.

Other area where AI technology can be useful is evaluation of the histological specimens [27]. Turkki and colleagues analysed viability of tumour xerograft. Automated assessment was shown to be highly specific in comparison to expert assessment. As xenograft tumour models are used in preclinical trials it does not fully correspond to our paper assumptions. However Turkkis work illustrates well the possibilities of AI in histological analysis.

3. Discussion

Our review is focused on the methodology of clinical validation of AI in medical systems. A lot of papers have been published on that area but only some of them give details of the clinical testing. AI is a relatively young branch of science. It is quite widely used in healthcare, for example in speech recognition for creating clinical notes [28]. Computer analysis is well established in ECG interpretation. As mentioned above, the first report on this topic comes from 1964. In 1973 Methewson reported that works on computer ECG interpretation system were advanced [29]. Subsequent digital revolution led to the development of ECG devices equipped with interpretation software. Since the 90-ties such software are the standard for ECG devices produced by all leading manufacturers. However

PubMed search gives only a little more than 1800 results for “ECG” + ”computer interpretation” keywords. The majority of papers related to clinical studies come from the recent years [26, 30, 31]. It is also difficult to find more detailed, not marketing data from industry source [32, 33]. The FreeSurfer case shows that results obtained from automated systems can differ from reference gold standard [4]. Despite the difference can be systemic and reproducible it does not mean that the results are simply “worse” or “better”. It only means they are different. This leads to the conclusion that there is a need for clinical trials that can validate practical usefulness of an AI system.

In European Union development and use of medical devices and medical software is controlled by number of regulations. Those include among the others: Council Directive 93/42/EEC (concerning medical devices), Directive 93/68/EEC (CE Marking), Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostic medical devices [34]. Harmonized standards apply to standardisation of medical devices. For our purpose the most relevant is the norm EN 62304:2006 concerning medical software. There are also local, national regulations, which can contain more detailed requirements. All those regulations are focused on biological and technical safety. The norm EN 62304:2006 requires clear development process. Validation of the device is necessary but no specific methods of validation are given.

Learning and testing of classifiers is generally standardized [35]. Regardless of the quality and amount of data, learning and testing process is performed in artificial conditions which can be very much different from clinical conditions. Therefor we claim that standard testing of machine learning effect should be followed by clinical validation.

Methods of clinical testing should be adjusted to the medical science area. It is necessary that experts in technical and medical science collaborate at each stage of the development process. It is important to elaborate a reference ‘gold’ standard if this has not yet been well established. The gold standard in medicine is often based on expert assessment/opinion. Such gold standard is biased by definition. Good examples of developing reference standard are presented in works by Ataer-Cansizoglu and Cruz [19, 22]. The study material should be derived prospectively from as natural as most possible conditions. For systems that are widely used more information is publically available. It is applicable for FreeSurfer that has bibliography with hundreds of position. However for many commercially available software only marketing information is available. It seems that that full clinical validation process should be conducted and its results published for such software/tools if used in medicine. At minimum all ethical and legal requirements for clinical studies should be fulfilled.

4. Conclusions

We propose that all computer-assisted systems undergo clinical validation irrespectively of the results from standard classifier tests. Methodology of clinical trials should be used.

BIBLIOGRAPHY

1. Wojciechowski K., Smołka B., Cupek R., Ziębiński A., Nurzyńska K., Kulbacki M., Segen J., Fojcik M., Mielnik P., Hein S.: A Machine-Learning Approach to the Automated Assessment of Joint Synovitis Activity. Computational Collective Intelligence 8th International Conference, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016; Proceedings, Part II, Proceedings, Part II in of the series Lecture Notes in Computer Science Nguyen N.-T., Manolopoulos Y., Iliadis L., Trawiński B. (Eds.).
2. FreeSurfer. [Online]. Available: http://surfer.nmr.mgh.harvard.edu/. [Accessed: 07-Dec-2015].
3. McCarthy C.S., Ramprashad A., Thompson C., et al.: A comparison of FreeSurfer-generated data with and without manual intervention. Front. Neurosci., vol. 9, Oct. 2015.
4. Butts A.: Freesurfer Vs. Manual Tracing: Detecting Future Cognitive Decline In Healthy Older Adults At-Risk For Alzheimer’s Disease. 2013.
5. McCrae C., O’Shea A., Boissoneault J., et al.: Fibromyalgia patients have reduced hippocampal volume compared with healthy controls. J. Pain Res., Jan. 2015, p. 47.
6. Phillips J.L., Batten L.A., Tremblay P., et al.: A Prospective, Longitudinal Study of the Effect of Remission on Cortical Thickness and Hippocampal Volume in Patients with Treatment-Resistant Depression. Int. J. Neuropsychopharmacol., vol. 18, no. 8, Jun. 2015, DOI: 10.1093/ijnp/pyv037.
7. Kim E.Y., Magnotta V.A., Liu D., Johnson H.J.: Stable Atlas-based Mapped Prior (STAMP) machine-learning segmentation for multicenter large-scale MRI data. Magn. Reson. Imaging, vol. 32, no. 7, Sep. 2014, p. 832÷844.
8. Bron E.E., Smits M., van der Flier W.M., et al.: Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. NeuroImage, vol. 111, May 2015, p. 562÷579.
9. Sorensen L., Pai A., Anker C., et al.: Dementia diagnosis using MRI cortical thickness, shape, texture, and volumetry. Proc MICCAI Workshop Chall. Comput.-Aided Diagn. Dement. Based Struct. MRI Data, 2014, p. 111÷118.

10. Roth H.R., Lu L., Seff A., et al.: A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations, [in:] Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer, 2014, p. 520÷527.
11. Mustapha A., Hussain A., Samad S.A., et al.: Design and development of a content-based medical image retrieval system for spine vertebrae irregularity. Biomed. Eng. Online, vol. 14, no. 1, 2015, p. 6.
12. Yang X., Wu N., Cheng G., et al.: Automated Segmentation of the Parotid Gland Based on Atlas Registration and Machine Learning: A Longitudinal MRI Study in Head-and-Neck Radiation Therapy. Int. J. Radiat. Oncol., vol. 90, no. 5, Dec. 2014, p. 1225÷1233.
13. Maier O., Schröder C., Forkert N.D., et al.: Classifiers for Ischemic Stroke Lesion Segmentation: A Comparison Study. PloS One, vol. 10, no. 12, 2015, DOI: 10.1371/journal.pone.0145118.
14. Ardakani A.A., Gharbali A., Mohammadi A.: Application of Texture Analysis Method for Classification of Benign and Malignant Thyroid Nodules in Ultrasound Images. Iran. J. Cancer Prev., vol. 8, no. 2, 2015, p. 116.
15. Bragman F.J., McClelland J.R., Modat M., et al.: Multi-scale Analysis of Imaging Features and Its Use in the Study of COPD Exacerbation Susceptible Phenotypes, [in:] Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer, 2014, p. 417÷424.
16. García Molina J.F., Zheng L., Sertdemir M., et al.: Incremental Learning with SVM for Multimodal Classification of Prostatic Adenocarcinoma. PLoS ONE, vol. 9, no. 4, Apr. 2014, DOI: 10.1371/journal.pone.0093600.
17. Depeursinge A., Kurtz C., Beaulieu C., et al.: Predicting Visual Semantic Descriptive Terms From Radiological Image Data: Preliminary Results With Liver Lesions in CT. IEEE Trans. Med. Imaging, vol. 33, no. 8, Aug. 2014, p. 1669÷1676.
18. Fehr D., Veeraraghavan H., Wibmer A., et al.: Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proc. Natl. Acad. Sci., vol. 112, no. 46, Nov. 2015, p. E6265÷E6273.
19. Ataer-Cansizoglu E., Bolon-Canedo V., Campbell J.P., et al.: Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity: Performance of the "i-ROP" System and Image Features Associated With Expert Diagnosis. Transl. Vis. Sci. Technol., vol. 4, no. 6, 2015, p. 5.
20. Fergus P., Hignett D., Hussain A., et al.: Automatic Epileptic Seizure Detection Using Scalp EEG and Advanced Artificial Intelligence Techniques. BioMed Res. Int., vol. 2015, 2015, p. 1÷17.

21. Siuly S., Kabir E., Wang H., Zhang Y.: Exploring Sampling in the Detection of Multicategory EEG Signals. Comput. Math. Methods Med., vol. 2015, 2015.
22. Cruz M.R., Martins C., Dias J., Pinto J.S.: A Validation of an Intelligent Decision-Making Support System for the Nutrition Diagnosis of Bariatric Surgery Patients. JMIR Med. Inform., vol. 2, no. 2, Jul. 2014, p. e8.
23. Jiang R., You R., Pei X.-Q., et al.: Development of a ten-signature classifier using a support vector machine integrated approach to subdivide the M1 stage into M1a and M1b stages of nasopharyngeal carcinoma with synchronous metastases to better predict patients' survival. Oncotarget, Nov. 2015.
24. Mehta D.D., Van Stan J.H., Zañartu M., et al.: Using Ambulatory Voice Monitoring to Investigate Common Voice Disorders: Research Update. Front. Bioeng. Biotechnol., vol. 3, Oct. 2015.
25. Pipberger H.V., Stallmann F.W.: Use of computers in ECG interpretation. Am. Heart J., vol. 64, Aug. 1962, p. 285÷286.
26. Ho T.-W., Huang C.-W., Lin C.-M., et al.: A telesurveillance system with automatic electrocardiogram interpretation based on support vector machine and rule-based processing. JMIR Med. Inform., vol. 3, no. 2, 2015, p. e21.
27. Turkki R., Linder N., Holopainen T., et al.: Assessment of tumour viability in human lung cancer xenografts with texture-based image analysis. J. Clin. Pathol., May 2015, p. jclinpath-2015-202888.
28. Forsiden - Max Manus AS. [Online]. Available: http://www.maxmanus.no/. [Accessed: 11-Jan-2016].
29. Mathewson F.A.: Electrocardiogram interpretation by computer. Can. Med. Assoc. J., vol. 108, no. 10, May 1973, p. 1207÷1208.
30. Park J., Kang K.: Intelligent Classification of Heartbeats for Automated Real-Time ECG Monitoring. Telemed. E-Health, vol. 20, no. 12, Dec. 2014, p. 1069÷1077.
31. Rautaharju P.M.: Eyewitness to history: Landmarks in the development of computerized electrocardiography. J. Electrocardiol., Nov. 2015.
32. ECG - Philips. [Online]. Available: http://www.healthcare.philips.com/main/products/patient_monitoring/products/ecg/. [Accessed: 28-Dec-2015].
33. Diagnostic ECG - Products. [Online]. Available: http://www3.gehealthcare.com/en/products/categories/diagnostic_ecg. [Accessed: 28-Dec-2015].

34. Medical devices - European Commission. [Online]. Available: http://ec.europa.eu/growth/single-market/european-standards/harmonised-standards/medical-devices/index_en.htm. [Accessed: 29-Dec-2015].
35. Dietterich T.G.: Machine-learning research. AI Mag., vol. 18, no. 4, 1997, p. 97.

Omówienie

Niniejsza praca powstała jako efekt rozważań nad metodyką walidacji klinicznej prototypu oprogramowania do klasyfikacji zapalenia stawów, jakiego powstanie jest końcowym celem projektu MEDUSA. Przeszukiwanie dostępnej literatury medycznej w bazie danych PubMed pod kątem słów kluczowych „sztuczna ineligencja" daje powyżej 5000 rezultatów rocznie (rys 1). Pomimo dostępności licznych norm aparatury oraz oprogramowania medycznego zasady kontroli ich klinicznej przydatności nie są wystandaryzowane. W naszej pracy przeanalizowaliśmy literaturę z ostatnich 3 lat dotyczącą tego problemu. Wybór literatury został dokonany przez przeszukanie bazy danych PubMed za pomocą angielskich słów kluczowych „artificial intelligence", „automated assessment" i „computer assisted". Po przeanalizowaniu dostępnych publikacji doszliśmy do wniosku, że w walidacji technik sztucznej inteligencji w medycynie należy zastosować metodykę jak w badaniach klinicznych nad nowymi metodami leczniczymi. Kluczowa jest współpraca pomiędzy ekspertami z aktualnej dziedziny medycznej oraz inżynierami.

Acknowledgment

The research leading to these results has received funding from the Norwegian Financial Mechanism 2009-2014 under Project Contract No. Pol-Nor/204256/16/2013

Addresses

Pawel MIELNIK: Section for Rheumatology, Department for Neurology, Rheumatology and Physical Medicine, Helse Førde, Førde 6800, Norway, pawel.franciszek.mielnik@helse-forde.no

Marcin FOJCIK: Faculty of Engineering and Science, Sogn og Fjordane University College, Svanehaugvegen 2, 6812 Førde, Norway, marcin.fojcik@hisf.no

Marek KULBACKI: Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland, kulbacki@pjwstk.edu.pl

Jakub SEGEN: Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland, segen@pjwstk.edu.pl