Increasing Trust Through the Design of Algorithm-Based Lesion Segmentation Support Systems

Emilia Gryska[1], Katerina Cerna[2], and Rolf A. Heckemann[1]

[1] Department of Radiation Physics, Gothenburg University Institute of Clinical Sciences, Gothenburg, Sweden
[2] Department of Information Systems and New Media, University of Siegen, Siegen, Germany

Abstract. The adoption rate of algorithm-based lesion segmentation support systems in clinical practice is very low. This is partly due to low trust levels radiologists have in such systems. To increase the trust, the design and validation of the support tools must comply with the needs and expectations of radiologists. We interviewed four clinicians who work with brain images on a daily basis to understand the needs, current methods and practices of image interpretation, and their opinion of automatic brain lesion segmentation tools. In the interviews, we identified the necessity to state the error of the automated decision support tool and its clinical relevance in a given context.

Keywords: algorithm-based support systems · brain lesion segmentation · trust · design for clinical practice.

1 Introduction

Technological advances in medical imaging and increased accessibility of imaging equipment lead to a significant increase of both the number and information content of images that need to be analysed. The time available to radiologists for reading each image is thus reduced, which can have an impact on the quality of image interpretation [8, 9]. A number of efforts are currently directed at alleviating this problem through the use of automatic image analysis [7]. Beyond the purely algorithmic challenges, a number of complex issues can obstruct the implementation path for systems designed to support medical image interpretation in a clinical setting.

In an ongoing scoping review, we are studying automatic brain lesion detection methods [3]. Such methods implement algorithms that delineate brain lesions on MR images without user interaction. A literature search of three databases (PubMed, IEEE Xplore, and Scopus) returned around 1500 articles that propose or evaluate such methods. In articles reporting pertinent data, we see increasing performance results over recent years, with accuracy levels matching or exceeding the variability measured between ratings by different experts (Figure 1).

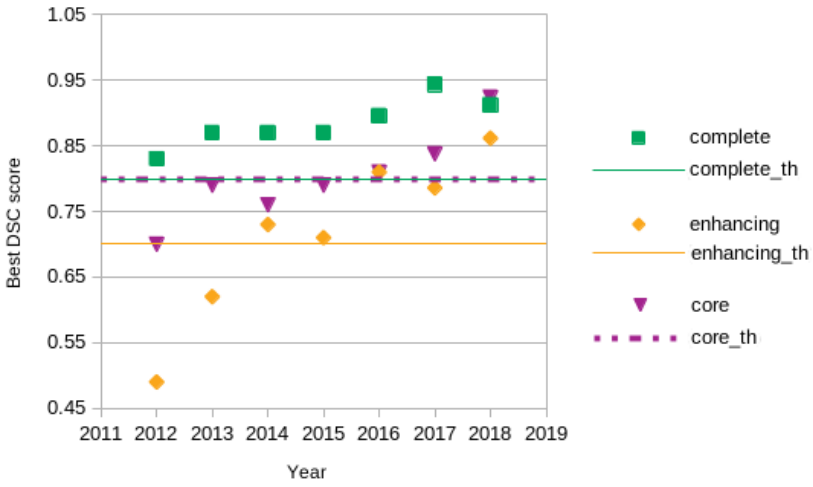

Fig. 1. Accuracy scores (Dice index) achieved for each tumor component in the yearly BRATS challenge [7]. "Complete", "enhancing", and "core" indicate the best score obtained for segmenting respective component of a tumor by an automated segmentation method. The "_th" suffix indicates variability measured between ratings by different experts for the respective tumor components.

Despite the growing evidence base and the strong performance of automatic segmentation methods, adoption into systems used in clinics is still rare. One factor that explains the situation is that despite the existing evidence, radiologists generally have low levels of trust in automated decision support systems [5]. Assuming that the low trust level is a significant factor [1], [2] obstructing adoption in clinical practice, a closer assessment of the design and validation framework of such systems is warranted.

1.1 Research question

In the present study, we seek to explore how we can draw on radiologists competence to improve the design and validation of algorithm-based lesion segmentation (ABS) support systems to increase the level of trust in such systems.

2 Methods

In the course of the scoping review referred to above, we conducted semi-structured interviews with four specialists (a neurosurgeon, an oncologist, a radiation oncologist, and a neuroradiologist). We chose these specialists because they process brain images in their daily clinical practice and are potential target users for the automatic brain lesion segmentation methods. The first author, who has a

background in biomedical engineering and medical image processing, conducted the interviews. The questions were tailored for each specialist to prompt them to describe their workflow. We aimed to understand some of the clinical needs and challenges pertaining to medical image interpretation. We posed questions about needs, current methods and practices of image interpretation, as well as their opinion on automatic brain image segmentation methods. The first and second author analysed the interview transcripts to identify challenges and suggestions for improving the design of image-analysis based support systems for clinical use. The second author has background in educational sciences.

3 Findings

Through the interviews, we identified three themes that impact trust and relate to the design of automated brain lesion segmentation tools. The three features of an ABS system are:

- accuracy level of the segmentation is given
- the context and procedure for calculating accuracy is provided
- clinical relevance of the accuracy level is estimated

The following excerpts come from an interview with the neuroradiologist and each is followed by our interpretation through which we identified the features stated above. Although we are using excerpts from only one interview, our interpretation resonates well with what we learned through the remaining interviews.

During the interview, the neuroradiologist told us about the criteria that would allow her to evaluate ABS:

Excerpt 1: "What do you think is better, your validated marginal, or my guesstimate (paraphrased a gesture we interpret means guessing)? It depends on what the margin is and that is in the validation step of the method. If you tell me in this disease and in this pipeline we have a segmentation error of 20% (thats a lot). Is 20% of clinical relevance? If yes, the tool is not good enough. Does your program allow for a second check? (...) Can you show me the segmentation, is it reasonable? Then you can accept larger margins."

Our interpretation of this excerpt is that neither a radiologist nor an automated algorithm-based segmentation tool is 100% accurate. For the tool to be trusted, however, the level of accuracy must be stated. The context of accuracy calculation (segmentation pipeline and type of lesion) and clinical relevance of the accuracy level should also be considered. The neuroradiologist also expresses a need for being able to verify the results visually as opposed to providing only quantitative results.

In the following excerpt, the neuroradiologist provides us with insights about the relevance of the accuracy level in the clinical context:

Excerpt 2: "In some questions any change is of clinical significance. For example a 2mm change can be a huge difference in one clinical case but not relevant at all in another."

In this excerpt, we see that measurements are relative with respect to particular cases. We can infer that if the accuracy level of the tool is not high enough to capture a significant level of change to be detected then the tool is not good enough.

The particular context is not important only for the measurements but also for the acceptance of the error:

Excerpt 3: "The acceptance of the error will depend on the question posed (). In some cases any change is of clinical significance."

The third excerpt points to the importance of considering clinical relevance of the error margin. We can also see that the neuroradiologist has a need for tools that are proficient at detecting change rather than accurately quantifying it.

Finally, the neuroradiologist's needs are summarized in the following excerpt:

Excerpt 4: "Sometimes engineers have an impression for the need for a very low error margin. But they are wrong because what we are doing subjectively I am 50% times wrong. What we need is a software that helps us accepting the margin as long as we have the idea on the margin. Am I perfect? No, should I not be allowed to interpret images? Well you can tell me but nobody else will interpret them perfectly anyways."

From this excerpt we infer that radiologists need to know in a given case how a support system makes a decision and with what error to trust the system enough to use it.

4 Discussion

In the interview we identified issues that support what Jorritsma et al. suggested in his paper [5]. The conclusions of Jorritsma's study identified four improvements of ABS system (referred to as CAD in the quotation) to increase the trust level radiologists have in the system: *"(1) presenting a confidence rating for the decisions made by the CAD system, (2) providing a global rationale for the decision-making process used by the CAD system, (3) providing a local rationale for each specific CAD decision, and (4) informing radiologists of the performance levels of the CAD system in different contexts."* Global rationale refers to a general principle by which a system makes a decision and local rationale to reasons for a particular decision in individual cases.

Considering the analysis of our interview as well as the review paper by Jorritsma et al. [5], we propose two general themes that need to be considered

in a system design to maximize trust: confidence rating, or level of accuracy, and clinical relevance of that level in a given context. Confidence rating refers to the reported performance of a segmentation method used for a given system. The level of accuracy will be influenced by a segmentation validation framework, characteristics of study population and images used for the validation, experts segmentation used to compare the methods output and metrics used for the comparison. Clinical relevance refers to a particular question that we want to answer with the help of an ABS system and what accuracy level is clinically acceptable in that question.

5 Future work

This study will serve as a pilot for a larger study to identify the information that is necessary to create and validate a trusted system. In the future we plan to interview a larger sample of radiologists in structured interviews to collect additional input regarding a trustworthy design of automatic decision support systems, and to propose a validation framework that takes into account radiologists needs and builds on their experience.

Another aspect of a future study that we want to consider is whether the conclusions of Jorritsma's review need to be revisited in view of the rapid growth of the field. Most of the papers reviewed in [5] were published before 2010. Since then we observed rapid developments in the algorithms and their performance, as well as an increase in the number of annotated images for testing the algorithms, as this increase has implications for the error margin estimations. In particular, we aim to learn whether the claims need to be updated with respect to deep learning algorithms. On the one hand, the complexity of these algorithms poses a problem for rationalizing the decisions made by such an algorithm. On the other hand, we acknowledge the substantial future potential of deep learning algorithms and their reported strong performance [6].

6 Conclusions

There is a gap between available methods for automatic brain lesion segmentation and their adoption in clinical practice. This is partially a result of low levels of trust in such methods on the part of radiologists. We must involve radiologists and use their knowledge to establish guidelines for design and validation of systems so they are trustworthy and end up being used in clinics.

References

1. Kim Drnec, Amar R. Marathe, Jamie R. Lukos, and Jason S. Metcalfe. From Trust in Automation to Decision Neuroscience: Applying Cognitive Neuroscience Methods to Understand and Improve Interaction Decisions Involved in Human Automation Interaction. *Front. Hum. Neurosci.*, 10, 2016.

2. Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718, June 2003.
3. Emilia Agnieszka Gryska, Justin Schneiderman, and Rolf A. Heckemann. Automatic brain lesion segmentation on standard MRIs of the human head: a scoping review protocol. *BMJ Open*, 9(2):e024824, February 2019.
4. Saurabh Jha and Eric J. Topol. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA*, 316(22):2353–2354, December 2016.
5. W. Jorritsma, F. Cnossen, and P. M. A. van Ooijen. Improving the radiologist-CAD interaction: designing for appropriate trust. *Clinical Radiology*, 70(2):115–122, February 2015.
6. Konstantinos Kamnitsas, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, February 2017.
7. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, October 2015.
8. Andrew B. Rosenkrantz, Tarek N. Hanna, James S. Babb, and Richard Duszak. Changes in Emergency Department Imaging: Perspectives From National Patient Surveys Over Two Decades. *Journal of the American College of Radiology*, 14(10):1282–1290, October 2017.
9. Rebecca Smith-Bindman, Diana L. Miglioretti, and Eric B. Larson. Rising Use Of Diagnostic Medical Imaging In A Large Integrated Health System. *Health Aff (Millwood)*, 27(6):1491–1502, 2008.