The Journal of Arthroplasty 36 (2021) e28–e29

Contents lists available at ScienceDirect

The Journal of Arthroplasty

journal homepage: www.arthroplastyjournal.org

Letter to the Editor

Response to Letter to the Editor on "Artificial Intelligence to Identify Arthroplasty Implants From Radiographs of the Hip"

In Reply

We thank the authors of the letter to the editor for their questions and comments and appreciate the opportunity to further discuss our work. We will respond to each paragraph individually.

In reply to the authors' discussion regarding the sensitivity and specificity of our results: As mentioned in our article, approximately 10% (206) of implants were used during the testing phase (approximately 80% used for training and 10% used for validation). These values are approximations as not all implants were perfectly divisible by 10. Using the example provided by the authors, our testing sensitivity for DePuy Trilock was 66.7%, with two true positives and one false negative. The authors next cite our reported value for specificity (99.8%), which is different than sensitivity and cannot be used to calculate false negatives. Using the reported *macro*-averaged sensitivity value of 94.3% gives approximately 12 expected false negatives. The *micro* average of sensitivity was 96.6%, which gives approximately 7 expected false negatives. The actual number of false negatives during testing was 7. We reported the *macro* average of our metrics to err on the side of caution and to avoid overstating our outcomes [1]. The largest difference between the macro and micro average of all of our outcome metrics was for sensitivity with a difference of 0.023.

Our implementation, using TensorFlow, outputs probabilities as correctly stated by the authors of the letter. It is correct to interpret these softmax probabilities as how confident the model is in its answer. There is a difference between confidence and credibility (which implies the use of contextual information based on prior distributions), which is why we only used the former in our publication.

We thank the authors of the letter for suggesting vector convoluted neural network (CNN) outputs, which would allow us to cluster output vectors into different higher-dimensionality clusters, each of which should belong to an individual implant. This is an interesting area of future work, although it has not been included in our original study.

We feel that it is important to recognize that CNNs are no longer completely opaque black boxes. While it is difficult to analyze information flow through a CNN, it is possible to, on a large scale, determine which pixels in a given image are most important. It also is important to note that our implementation is a supervised machine learning problem, which allows us to use class activation maps as described in our methods [2]. We did not intentionally "focus" the CNN on the femoral component. The model autonomously made this decision as all of the important pixels for identifying implants should reside in the area of the implant. While it is true that other areas of the image are used by the model, using a large sample size and robust image processing techniques (e.g., augmentation) should make the model use features relevant to each unique implant design. As stated in our article, 1.4% of images in our testing set had gradients that were not focused on pixels containing the implant and instead were spread throughout the image. Most likely, these implants did not have satisfactory class activation maps due to low number of implants in those groups. Based on analysis of our activation maps, we do not believe that the model was influenced by acetabular cup inclusion in this study. Future work will focus on acetabular cups.

We thank the authors of this letter to the editor for their interest in our work and hope that this discussion reinforces the findings outlined in our article and spurs interest for further study in this area.

Jaret M. Karnuta, MS
Orthopaedic Machine Learning Laboratory
Cleveland Clinic
Cleveland, OH

Heather S. Haeberle, MD
Orthopaedic Machine Learning Laboratory
Cleveland Clinic
Cleveland, OH

Department of Orthopaedic Surgery
Hospital for Special Surgery
New York, NY

Bryan C. Luu, BS
Orthopaedic Machine Learning Laboratory
Cleveland Clinic
Cleveland, OH

Department of Orthopaedic Surgery
Baylor College of Medicine
Houston, TX

Prem N. Ramkumar, MD, MBA*
Orthopaedic Machine Learning Laboratory
Cleveland Clinic
Cleveland, OH

DOI of original article: https://doi.org/10.1016/j.arth.2020.12.043.
Investigation performed at the Cleveland Clinic, Cleveland, Ohio.

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.arth.2020.12.046.

Department of Orthopaedic Surgery
Brigham & Women's Hospital, Boston, MA

*Reprint requests: Prem N. Ramkumar, MD, MBA, 9500 Euclid Avenue, Cleveland, OH.

https://doi.org/10.1016/j.arth.2020.12.046

References

[1] Sci-Kit Learn Developers. Metrics and scoring: quantifying the quality of predictions. Sci-Kit Learn API Doc. https://scikit-learn.org/stable/modules/model_evaluation.html; 2020. [accessed 24.12.20].
[2] Zhou B, Khosla A, A L, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. CVPR. https://arxiv.org/pdf/1512.04150.pdf; 2016. [accessed 24.12.20].