European Journal of Cancer 130 (2020) 262–264

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Letter to the Editor

Reply to the letter to the editor: 'Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images'

Achim Hekler [a], Jochen S. Utikal [b,c], Wiebke Solass [d], Max Schmitt [a], Joachim Klode [e], Dirk Schadendorf [e], Wiebke Sondermann [e], Cindy Franklin [f], Felix Bestvater [g], Dieter Krahl [h], Christof von Kalle [a], Stefan Fröhling [a], Titus J. Brinker [a,*]

[a] *National Center for Tumor Diseases, German Cancer Research Center, Heidelberg, Germany*
[b] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[c] *Skin Cancer Unit, German Cancer Research Center, Heidelberg, Germany*
[d] *Institute of Pathology and Neuropathology, Eberhard-Karls-University Tuebingen and National Center for Pleura and Peritoneum, University of Tuebingen, Germany*
[e] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[f] *Department of Dermatology, University Hospital Cologne, Cologne, Germany*
[g] *Core Facility Unit Light Microscopy, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany*
[h] *Private Laboratory of Dermatohistopathology, Mönchhofstraße 52, 69120, Heidelberg, Germany*

Received 21 December 2019; accepted 21 December 2019
Available online 30 January 2020

Dear Editor,

At the end of the recruitment period, 14 addresses had acknowledged the invitation by e-mail, and these physicians were counted as invited to the survey. As described in the manuscript, participants with and without board certification were eligible because pathologists with and without board certification assess histologic slides in clinical routine.

As described in the Methods section of the manuscript, the primary end-point of the study was the superiority of the average result of the 11 training and test runs achieved by the algorithm compared with the average result achieved by 11 pathologists (with and without board certification) assessing histologic slides in clinical routine. Thus, the title correctly describes the study's predefined statistics.

While it is formally correct that we used a Google form, we do not suspect fraudulent participation in the survey because only colleagues personally known to the recruiters were contacted and received the link.

The estimated 20 min necessary for participation in the invitation letter were not a limit that we would have set, and the participants had all the time they needed. The 20-min time frame was merely an estimate of how long one would need to participate. It was meant to help

DOIs of original article: https://doi.org/10.1016/j.ejca.2019.09.018, https://doi.org/10.1016/j.ejca.2019.06.012.

* *Corresponding author*: National Center for Tumor Diseases, German Cancer Research Center, Im Neuenheimer Feld 460, Heidelberg, 69120, Germany.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

https://doi.org/10.1016/j.ejca.2019.12.024

the participants to assess if and when they would have time to classify cropped melanoma images, which are very small (0.06% of the whole slide on average, as described in the manuscript). Based on the feedback we received, it appears that less than 20 min is realistic because the images were very small, the decision was binary and no clinical information was included.

In the e-mails we sent, we let the participants know what the experiment was about, so that they would feel respected and involved in the study. We do not think that this has impacted the results or their diligence in evaluating the images.

The quotes by Geraud and Griewank on 'statements', we would have made with regard to the outperformance are incomplete or taken out of context. We feel that we have clearly acknowledged the limitations of our study throughout the manuscript. The entire title reads 'Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images', correctly referring to the classification of images instead of whole slides. The abstract further states that cropped images were assessed. In the Methods and Limitations section, we have extensively discussed that the experiment does not reflect clinical practice but represents a first direct comparison between deep learning and human pathologists with regard to a defined classification task on histologic images of melanomas and nevi where both classifiers have identical information. All results are described as 'preliminary' throughout the manuscript.

We did not have the primary aim to show superiority. It was our hypothesis when we planned the study that deep learning could outperform pathologists on such a task.

We fully agree with the statement 'A major issue of the Hekler *et al.* study is applying cropped images showing only a small random portion of the actual lesion. No dermatopathologist would make a real routine patient diagnosis being able to inspect only a small piece of one section of the entire lesion.', made by Geraud and Griewank and correctly discuss this caveat in the Limitations section of the manuscript: 'Even though H&E slides are also evaluated in daily routine, in a normal setting, a pathologist is able to look at the whole slide instead of just part of a section and order additional immunostaining.'

We also discuss potential explanations for the shown effects in the Discussion section of the manuscript: 'The clear outperformance may be explained by the ability of artificial intelligence to mine 'sub-visual' image features that may not be visually discernible by a pathologist.' This might also be true for near visual threshold lateral extensions of intraepidermal melanocytic proliferations (Madabushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal. 2016; 33:170–175).

We have provided the test set and both the majority decision of the deep learning algorithm and the majority decision of the pathologists as Appendix 1 to this letter. Partial representation of melanocytic nevi in our scans because of the randomly selected image crops of the whole slides could, in some cases, evoke the impression of papillomatous nevi. However, those nevi showed dysplastic features when viewed as a whole or in another cutting direction and were, therefore, included in the specimen. Per definition, 'Clark's' nevi can consist of a central papillomatous portion with lateral 'shoulders' showing the criteria used for diagnosis of dysplastic nevi. Thus, this type of composite nevus architecture can be missed when only showing the central papillomatous, 'Unna-like' part in a random image crop.

We have correctly discussed the limitation of high discordance rates of single pathologists in the Limitations section of the manuscript: "Finally, it should be noted that the defined ground truth has to be interpreted with caution: While the procedure that led to the definition of the ground truth is the standard of care in histopathological melanoma diagnosis, around 25–26% of discordance is found between two pathologists who assess the same slides for melanoma vs. nevus. Possible alternatives for future research to improve the ground truth include consent decisions of groups of histopathologists, genomic analyses and the integration of cancer registry data.' The lesions chosen for these two studies were from clinical routine and included mostly thin melanomas, which show relatively high discordance rates.

In addition, we do not understand how the conclusion that the estimated concordance rates are higher can be drawn from the studies by Elmore et al [3] and Elder et al [4] cited by Geraud and Griewank. As mentioned correctly, five classes were considered [4]; however, the results presented do not allow statistical analyses related to the binary problem in our opinion. The situation is slightly different in Elmore et al [3] where detailed results are presented (Table 4) that allow approximate calculations, which yield interobserver concordance values in a similar order of magnitude as described in our manuscript.

When preparing the article, we especially looked for studies that considered a binary classification problem. Accordingly, the two studies cited in our introduction, which report discordance rates of 25–26%, are: (a) Corona *et al.* [5]: this study investigated the concordance among four pathologists on a set of 140 slides containing cutaneous melanoma and benign pigmented skin lesions. The participating pathologists agreed on 103 of 140 cases (74%). (b) Lodha *et al.* [6]: the authors examined the concordance of pathologists evaluating difficult cases of nevus vs. melanoma. A total of 143 cases were evaluated by two pathologists. In a total of 36 cases (25%), the highest level of disagreement (definite melanoma vs. definite nevus) occurred.

We fully agree with the statement that a review panel would further improve the ground truth and the quality

of the study. However, at the time, we did not have the resources available to conduct such a study. For a higher level claim, we are currently conducting a follow-up investigation with a test set consisting of whole slides, which were verified by a review panel of renowned board-certified histopathologists. At this point of time, we were very thankful to be supported by Dr. Heinz Kutzner, MD (Friedrichshafen, Germany) and national/international colleagues of Dr. Kutzner, which made this possible.

We agree with the statement that a binary decision does not reflect clinical practice but think that a binary decision reflects the highest level of relevance to the patient at hand. A more fine-grained approach would be possible at later points in time for a definitive study, once the binary problem has been sufficiently investigated. However, at the time, we did not have the resources available to conduct such a study, and it was not the aim of our study to reflect World Health Organisation schemes.

We do not anticipate that there are too few authors on the publication, nor that they are inexperienced in designing studies. We discussed the limitations of our study in great detail and did not draw other conclusions than that outperformance is possible on the histopathological images (and not on whole slides). We call our results 'preliminary' multiple times in the manuscript.

The authors view their work in this new field as what it is, the first two pilot studies implementing deep learning into the histopathologic analysis of melanoma. They are no definitive studies, and the limitations have been extensively discussed in both manuscripts [1,2]. While we agree that the data are preliminary data and acknowledge this in both manuscripts, we argue that our studies are important as they reveal the potential of deep learning in this field and introduce a new application. At the same time, it is clear that pilot studies cannot provide definitive findings. We report a direct comparison between specialists and artificial intelligence with obvious limitations as to the representation of clinical practice but clearly demonstrate the potential of artificial intelligence with regard to a specific task with identical conditions for human and computer assessment, that is, the same amount of information for both classifiers.

Conflict of interest statement

The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2019.12.024.

References

[1] Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019;118:91–6.
[2] Hekler A, Utikal JS, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur J Cancer 2019;115:79–83.
[3] Elmore JG, Barnhill RL, Elder DE, Longton GM, Pepe MS, Reisch LM, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ 2017;357:j2813.
[4] Elder DE, Piepkorn MW, Barnhill RL, Longton GM, Nelson HD, Knezevich SR, et al. Pathologist characteristics associated with accuracy and reproducibility of melanocytic skin lesion interpretation. J Am Acad Dermatol 2018;79:52–9. e5.
[5] Corona R, Mele A, Amini M, De Rosa G, Coppola G, Piccardi P, Faraggiana T. Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions. J Clin Oncol 1996;14(4):1218–23.
[6] Lodha S, Saggar S, Celebi JT, Silvers DN. Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting. J Cutan Pathol 2008;35(4):349–52.