European Journal of Cancer 130 (2020) 259–261

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Letter to the Editor

Re: Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images

Cyrill Géraud [a,b,c,*], Klaus G. Griewank [d,e,**]

[a] *Department of Dermatology, Venereology and Allergology, University Medical Center and Medical Faculty Mannheim, Heidelberg, Mannheim, Germany*
[b] *Section of Clinical and Molecular Dermatology, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[c] *European Center for Angioscience, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[d] *Department of Dermatology, University Hospital Essen, West German Cancer Center, University Duisburg-Essen and the German Cancer Consortium (DKTK), Essen, Germany*
[e] *Dermatopathologie bei Mainz, Nieder-Olm, Germany*

Received 28 August 2019; accepted 10 September 2019
Available online 23 October 2019

To the editor,

We read with great interest the recent publication by Hekler *et al.* [1]. The authors compared analysis of histology images from melanomas and nevi by deep learning with assessment by 11 humans. The results are interesting; however, we have considerable concerns regarding the title, study design and the conclusions made.

Reviewing the invitations we received for participation (provided to the editors), this was sent to a considerable number of physicians including resident physicians without board certification in dermatology, dermatopathology or pathology by email. The list of people invited and the listed study participants we know of already exceeds the 14 reported in the manuscript. As only eight of the 11 study participants were board-certified, and the manuscript reports two participants performed on par with deep learning, the study title 'outperformed 11 pathologists' is misleading. Participant information was based on a freely accessible google form not restricted to invited participants making verification impossible. The 20 min allotted in the invitation are insufficient for a pathologist to assess 100 images of melanocytic lesions with the required prudence (=12 seconds per image). The senior author's invitation letter explicitly stated that the study design, applying cropped images, provided image material inappropriate for a human pathologist, the hypothesis being deep learning could better distinguish nevus and melanoma under these circumstances. We find it problematic that the manuscript makes statements such as 'Deep learning outperformed 11 pathologists' and 'The aim of this study is to perform such a first direct comparison' without explicitly stating the caveat mentioned in the invitation letter that the study design intentionally impeded diagnosis by a human pathologist. It is also scientifically questionable to conceive studies with the primary aim of showing superiority of a new technique.

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

DOI of original article: https://doi.org/10.1016/j.ejca.2019.06.012.
* *Corresponding author*: University Medical Center Mannheim, Theodor-Kutzer-Ufer 1-3, Mannheim 68163 , Germany.
** *Corresponding author*: University Hospital Essen, Hufelandstr. 55, 45147 Essen, Germany.
E-mail addresses: cyrill.geraud@umm.de (C. Géraud), klaus.griewank@uk-essen.de (K.G. Griewank).

https://doi.org/10.1016/j.ejca.2019.09.018

real routine patient diagnosis being able to inspect only a small piece of one section of the entire lesion. Melanomas can be highly heterogeneous or nevus-associated which is why their diagnosis (at least for humans), relies on a number of criteria that can only be assessed when the complete lesion is visible. A key choice a pathologist has when facing incomplete melanocytic lesions is deciding whether the visible features are insufficient for diagnosis, and further work-up or re-excision of the complete lesion is required. Unfortunately, the study design did not provide this option. We personally opted against participation in the study feeling that for most images included the binary approach forcing participants to choose a diagnosis of nevus or melanoma was problematic. One can argue deep learning algorithms may recognise specific signs not readily recognised by a human pathologist. However, reviewing the 100 histology pictures included in the trial, we identified more than 15 (>15%) having no recognisable melanocytic lesion whatsoever presumably representing perilesional normal skin tissue. Certainly, a human pathologist cannot make a distinction of nevus or melanoma if a melanocytic lesion is not present in the image demonstrated. If Hekler *et al.* believe deep learning can distinguish nevus from melanoma based solely on perilesional tissue, they need to show specific data to support this theory. The authors should provide access to all cropped images with the corresponding diagnoses and distribution of choices made by deep learning and participants to allow the readership an unbiased assessment of the data.

Interestingly, despite citing discordance rates among expert histopathologist of 25–26% [1,2], Hekler *et al.* performed studies where melanoma or nevus were defined by a single histopathologist. The discordance rates cited are problematic as they predominantly refer to difficult to classify melanocytic lesions, not included in the Hekler *et al.* studies [1,2]. Concordance rates estimated at a population level are higher [3,4]. Comparisons are problematic however, as existing studies apply multiple different diagnostic classes (3–5 or more), whereas despite claiming 'pathologist-level classification' [2], the Hekler *et al.* studies are binary. It is nonetheless true that melanoma is currently a pathologist-defined entity, and lesions can be designated differently by different histopathologists. In the Hekler *et al.* studies, deep learning and the other histopathologists were in fact compared with a single pathologist's diagnostic view. A well-controlled study would require a large cohort of histopathology images containing melanomas and different nevi designated such by a number of experienced histopathologists (in our opinion at least 3, preferably more). Additionally, pathology and biology are not binary. Borderline lesions exist which cannot be unequivocally designated as melanoma or nevus [5]. Omitting such cases from the analysis and using binary choices instead of more sophisticated classifications such as the MPATH-DX (Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis) [6] or World Health Organisation [7] scheme may simplify study design and facilitate training of artificial intelligence. However, it does not reflect the current state of classification and makes application to unselected melanocytic lesions encountered in routine practice impossible.

Addressed are the concerns we found most problematic, many of which were communicated to the Brinker group prior to publication (by us and others [personal communication]). We believe several criteria proposed to identify distortion or misinterpretation ('spin') in biomedical research studies [8,9] are present in the Hekler *et al.* studies. One wonders if most of the participating histopathologists chose not to be listed as an author or participant because of similar concerns regarding study design and/or the conclusions drawn.

We do anticipate that computer-driven approaches including deep learning and artificial intelligence have the potential to revolutionise histopathology and provide an enormous diagnostic aid. However, the considerable hype regarding this topic and its potential should not alleviate the requirement to perform well-designed studies with critical scrutiny if the conclusions drawn are justified. Studies need to be performed on test sets with well-characterised tumour cohorts, image material suitable for histopathologic assessment and more adequate options for classification. A very fundamental flaw of the Hekler *et al.* study design is it impairs the diagnostic capabilities of the human pathologist, making a direct comparison with artificial intelligence problematic. Before entering patient care, computer-driven approaches will need to prove their value when compared with or supplementing human pathologists in an optimal diagnostic setting.

Conflicts of interest statement

The authors have no conflicts of interest to declare.

Acknowledgements

No third-party funding was applied to the current manuscript. Funding agencies did not influence the content of the manuscript.

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] Shain AH, Yeh I, Kovalyshyn I, Sriharan A, Talevich E, Gagnon A, et al. The genetic evolution of melanoma from precursor lesions. N Engl J Med 2015;373:1926–36.
[6] Lott JP, Elmore JG, Zhao GA, Knezevich SR, Frederick PD, Reisch LM, et al. Evaluation of the melanocytic pathology assessment tool and hierarchy for diagnosis (MPATH-Dx) classification scheme for diagnosis of cutaneous melanocytic neoplasms: results from the international melanoma pathology study group. J Am Acad Dermatol 2016;75:356–63.
[7] Elder DE, Massi D, Scolyer RA, Willemze R. WHO classification of skin tumours. 4th ed. Lyon: IARC; 2018.
[8] Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A 2018;115:2613–9.
[9] Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of “spin”. Radiology 2013;267:581–8.