Letter to the Editor

Reply to "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists" by H. A. Haenssle et al

In a recently published paper in the Annals of Oncology[1], Haenssle et al compare the performance of a deep learning model to that of 58 dermatologists. The article was of high general quality, yet their methodology requires clarification.

First, they underestimate human performance by using a metric that they call the ROC area. This is not the same metric as the ROC-AUC, which they compare it to. The ROC-AUC is the calculated area under the ROC curve, whereas the ROC area is the average of sensitivity and specificity at a given operating point. Comparing two different metrics as if they are the same is inappropriate.

In this paper, we as readers cannot calculate the ROC-AUC for the dermatologist group with the data provided, but we can calculate the ROC-area for the model at the specified operating points. These are presented in Table 1, and shows no difference between the model and dermatologists in these experiments.

The authors also present sensitivity and specificity results at the level of human sensitivity. Second is that the mechanism for selecting this operating point is not stated, but it is likely this occurred post-experiment. We see evidence for this in the section “Diagnostic accuracy of CNN versus dermatologists”, where several operating points are chosen for the AI system which appear to *exactly* match the level of human sensitivity. If this decision was made using the training data, the sensitivity on the test data would almost certainly be slightly different than the human level.

I note that in figure 2a from Haenssle et al. the ROC curve is very steep in both directions in the region of interest, and a very small change in operating point could lead to a very large reduction in either specificity or sensitivity (into the 70s for both metrics). This suggests that the model performance may be significantly overestimated.

I expect the model of Haenssle et al. performs very well, but the methods applied overestimate the performance of the model and underestimate the performance of the

human experts. The methodologies used require clarification and may raise questions about the validity of the results and the conclusions of the paper.

The authors have declared no conflicts of interest.

L Oakden-Rayner*

The School of Public Health, University of Adelaide, South Australia, Australia

(*E-mail: lukeoakdenrayner@gmail.com)

Funding: None declared.

Conflicts: The authors have declared no conflicts of interest.

References:

1. Haenssle H, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018; 29: doi.org/10.1093/annonc/mdy166.

	Sensitivity	Specificity	AUC	ROC area
CNN (0.5 threshold)	95	63.8	86	79*
Derm L1	86.6	71.3	-	79

Table 1: The performance of the CNN and dermatologists on the task. *ROC area for the model (not presented in the paper).