TO THE EDITOR

Although deep learning–based algorithms demonstrate excellent results (Brinker et al., 2019; Esteva et al., 2017; Haenssle et al., 2018; Liu et al., 2020; Tschandl et al., 2019) for a trained task, uncertainty concerns exist even for an easy task. The accuracy demonstrated in previous studies may not be reproduced when the test environment changes (Cook, 2019; Han et al., 2019; Liu et al., 2019; Zech et al., 2018). Therefore, external tests must be performed, and the model must be improved by reflecting the feedback. The top-1 accuracy of our algorithm for the International Skin Imaging Collaboration dataset is low (39%), as reported by Navarrete–Dechent et al. (2020). However, in this letter, we explain why we should not conclude that the algorithm’s performance is poor on the basis of only this result.

First, for an accurate assessment of the diagnostic performance of the algorithm, a cross-sectional or comparative study with respect to dermatologists is a better design than a small case study because in a small case study, the selected population cannot represent the population that the test will be applied to in the clinical setting (Chassé and Fergusson, 2019). The performance of an algorithm can be overestimated or underestimated according to given test sets; therefore, it should be evaluated in comparison with the performance of dermatologists. A cohort study, where all cases in a certain clinical setting are tested, better reflects the real world (Park and Han, 2018).

At first glance, the top-1 accuracy for the International Skin Imaging Collaboration dataset (39%) appears to be unacceptably low. If this is the case, what about the top accuracies of dermatologists? The diagnostic accuracy of dermatologists for biopsied cases is lower than that we generally expect. In a recent study (Han et al., 2019) using the cases of Severance Hospital (Seoul, Korea), we tested almost all single lesion–biopsied cases (10,426 cases) from January 2008 to March 2019. We reviewed the clinical diagnoses of the attending physicians, which were performed using the information obtained by reviewing the history and examining patients and even making reference to a referral note. The top-1 and top-3 accuracies in actual practice were 65.4% and 74.7%, respectively, and the sensitivities for malignancy derived from the top-1 and top-3 diagnoses were 70.2% and 88.1%, respectively. The accuracy further decreased when dermatologists diagnosed using only images. The top-1 and top-3 accuracies of the reader test were 37.7% and 53.4%, respectively, and the sensitivities for malignancy were 65.8% and 84.9%, respectively.

It is inherently difficult to find an exact diagnosis match among various diseases. In an Australian study, general physicians’ diagnostic sensitivities for malignant melanoma was 33.8% (Heal et al., 2008). The difficulty further increases when only a single image is available for diagnosis. In general, the top-1 and top-3 (39% and 63%, respectively) accuracies and sensitivity (77%) of our algorithm for the International Skin Imaging Collaboration dataset are within a range similar to those of readers in the Severance study.

Second, the algorithm should be adjusted according to the target population, and thresholds should be adjusted with those results. Between the melanomas of Asian individuals and those of White individuals, the degree of bizarreness in the morphology differs significantly. In Figures 1 and 2, three benign melanoma cases of Asian individuals and one melanoma case from the International Skin Imaging Collaboration dataset were presented, and all cases were pathologically confirmed. If Figure 1d represents a case of an Asian individual, the initial impression may be intradermal nevus rather than melanoma. Conversely, if Figure 1a–c represent cases of White individuals, then melanoma should be considered.

Because a tradeoff between sensitivity and specificity exists, we must investigate a proper threshold depending on whether a given task is screening or confirmation. The threshold in the White individuals may be lower than that in Asian individuals because of the high prevalence of melanoma. If we use the same threshold for Asian individuals, a significant number of patients will be recommended to undergo an unnecessary biopsy. Both sensitivity and specificity are important for the detection algorithm. It is particularly true considering that algorithms can be used much more frequently than dermatologic clinic visits. To calibrate the threshold for the target population, a cohort study such as the Severance study is necessary.

Third, the performance of the algorithm is affected by the composition of...
Because the algorithm has been trained with the assumption that the user selects the region of interest that centers the lesion and is 80% covered by the lesion, the optimal result can be obtained when the algorithm is used as instructed. When using the detection algorithm (http://rcnn.modelderm.com), the sensitivity can be varied according to the number of trials. We set the threshold to the setting where the user captured multiple photographs. In our previous studies, an average of 3.9 (Han et al., 2019) and/or 4.2 (Han et al., 2020a) photographs were captured to document one lesion. Because the detection algorithm selects the highest value as the final output, if only one photograph were used for the analysis, the sensitivity could have been lower, although the specificity could have been higher. The user should have been instructed to capture two close ups and two macro photographs with different angles in the web-DEMO; however, no instructions were provided in the early DEMO.

A recent meta-analysis of deep learning algorithms showed that the performances of those algorithms were comparable with those of specialists in the same setting (Liu et al., 2019). However, if an algorithm was not trained with all clinical information, the performance was inferior to that of dermatologists in actual practice (Han et al., 2019). Therefore, it is more appropriate to combine conventional dermatologic care with algorithms rather than to rely entirely on algorithms. However, many patients cannot visit hospitals owing to the lack of dermatologists, high medical costs, and negligence of suspected lesions. In our previous studies (Han et al., 2020a, 2020b), laypersons missed
approximately 50% of cases of malignancy.

In conclusion, to more completely understand the sensitivity and specificity of an algorithm, knowledge of the comparative performance of dermatologists would be helpful. Instruction for the usage of algorithms is important, and a misusage will deteriorate their performances. To clarify the performances of algorithms, prospective randomized controlled studies need to be conducted in the future with various ethnicities in various environments.

Data availability statement

Datasets related to this article can be found at https://doi.org/10.6084/m9.figshare.12478238

ORCIDs

Seong Seog Han: http://orcid.org/0000-0002-0500-3628
Seong Hwan Kim: http://orcid.org/0000-0001-6831-5621
Jung-Im Na: http://orcid.org/0000-0002-5717-2490

CONFLICT OF INTEREST

The authors state no conflicts of interest.

AUTHOR CONTRIBUTIONS

Conceptualization: SSH, SHK, JIN; Resources: SHK, JIN; Writing – Original Draft Preparation: SSH, JIN; Writing - Review and Editing: SSH, JIN

Seong Seog Han1,4, Seong Hwan Kim2,4 and Jung-Im Na3,*

1Department of Dermatology, 1 Dermatology Clinic, Seoul, Korea; 2Department of Plastic and Reconstructive Surgery, Kangnam Sacred Hospital, Hallym University College of Medicine, Seoul, Korea; and 3Department of Dermatology, Seoul National University Bundang Hospital, Seongnam, Korea
4These authors contributed equally to this work.
5Corresponding author e-mail: jina1@snu.ac.kr

REFERENCES


NCSTN Deficiency and Depigmentation: All About Tyrosinase?


TO THE EDITOR

In their recent publication in the Journal of Investigative Dermatology, Hsu et al. (2020) employ studies on a zebrafish (ZF) mutant to support their hypothesis that NCSTN deficiency induces tyrosinase-dependent depigmentation. Although we acknowledge the large data set, we are not entirely convinced by their presentation and/or interpretation of the data and certain aspects of their experimental approach.

NCSTN is a type 1 integral membrane protein and a member of the γ-secretase endoprotease complex. Among others, this complex is responsible for the intracellular cleavage of Notch receptors, a process required for the activation of the Notch signaling pathway. Besides NCSTN, γ-secretase comprises three further transmembrane protein subunits: PSEN1/PSEN2, PEN-2, and APH-1α/APH-1b (Wolfe, 2020). NCSTN acts as a substrate receptor and is encoded by the NCSTN gene, which is located on chromosome 1q23.2. Two previous reports have described an association between NCSTN deficiency secondary to NCSTN mutations and pyoderma gangrenosum, acne, and suppurative hidradenitis syndrome. However, this association remains uncertain owing to the low number of cases reported and an overlap of the clinical phenotype with that of other nonsyndromic and syndromic forms of