Dermatology Practical & Conceptual

Application of an Interactive Diagnosis Ranking Algorithm in a Simulated Vignette-based Environment for General Dermatology

Antonia Wesinger[1], Elisabeth Riedl[1], Harald Kittler[1], Philipp Tschandl[1]

1 Department of Dermatology, Medical University of Vienna, Vienna, Austria

Key words: logic, diagnosis, algorithm, ranking, human-computer interaction

Citation: Wesinger A, Riedl E, Kittler H, Tschandl P. Application of an interactive diagnosis ranking algorithm in a simulated vignette-based environment for general dermatology. *Dermatol Pract Concept*. 2022;12(3):e2022117. DOI: https://doi.org/10.5826/dpc.1203a117

Accepted: November 24, 2021; **Published:** July 2022

Funding: This research was supported by a grant of the Austrian Research Promotion Agency (FFG = Forschungsförderungsgesellschaft). The investigation is part of the FFG-project: "A novel decision support tool for training and diagnosis in dermatology" (project number: 840281). The development of the CDRA was realized by the Medical University of Vienna, and its project partners, the Vienna University of Technology and emergentec biodevelopment GmbH.

Competing interests: PT reports fees from Silverchair, grants from MetaOptima Technology Inc. and Lilly, and speaker honoraria from Lilly, FotoFinder and Novartis, all outside the submitted work. ER is currently an employee and minor stockholder of Eli Lilly and Company. HK reports royalties or licenses from Casio, Barco and MetaOptima, speaker honoraria from FotoFinder, and receipt of equipment for testing from FotoFinder, 3Gen, DermaMedicalSystems, Heine and Casio; all outside the submitted work.

Authorship: All authors have contributed significantly to this publication.

Corresponding author: Philipp Tschandl, MD, PhD, Department of Dermatology Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria E-mail: philipp.tschandl@meduniwien.ac.at

ABSTRACT

Introduction: Diagnostic algorithms may reduce noise and bias and improve interrater agreement of clinical decisions. In a practical sense, algorithms may serve as alternatives to specialist consultations or decision support in store-and-forward tele-dermatology. It is, however, unknown how dermatologists interact with algorithms based on questionnaires.

Objectives: To evaluate the performance of a questionnaire-based diagnostic algorithm when applied by users with different expertise.

Methods: We created 58 virtual test cases covering common dermatologic diseases and asked five raters with different expertise to complete a predefined clinical questionnaire, which served as input for a disease ranking algorithm. We compared the ranks of the correct diagnosis between users, analyzed the similarity between inputs of different users, and explored the impact of different parts of the questionnaire on the final ranking.

Results: When applied by a board-certified dermatologist, the algorithm top-ranked the correct diagnosis in the majority of cases (median rank 1; interquartile range: 1.0; mean reciprocal rank 0.757). The median rank of the correct diagnosis was significantly lower when the algorithm was applied by four dermatology residents (median rank 2-5, $P < 0.01$). The lowest similarity between inputs of the residents and the board-certified dermatologist was found for questions regarding morphology. Sensitivity analysis showed the highest deterioration in performance after omission of information on morphology and anatomic site.

Conclusions: A simple questionnaire-based disease ranking algorithm provides accurate ranking for a wide variety of dermatologic conditions. When applied in clinical practice, additional measures may be needed to ensure robustness of data entry for inexperienced users.

Introduction

Skin diseases have a profound impact on public health as they are estimated to account for a large fraction of all primary care visits [1,2]. Skin diseases are the fourth most common form of illness and affect almost one-third of the world population at any time [3,4]. Furthermore, because of the rising incidence of skin cancer in most countries, accurate diagnosis and treatment of cutaneous neoplasms are required to maintain a high standard of care in the future. Recent developments in the field of artificial intelligence (AI) propelled machine learning algorithms in the center of image-based diagnostic dermatology [5,6], but this development was also the target of substantial critique [7-9]. The main points of critique include lack of robustness and interpretability of current machine learning algorithms as well as failure to include relevant diagnostic information beyond what is captured in images. A more complete view of the patient including contextual information may lead to better and more robust diagnoses for neoplastic and inflammatory diseases [10,11]. Attempts to incorporate multimodal information in machine learning models for automated diagnosis are emerging slowly [12-17]. Only a few digital tools employ a bottom-up approach starting with the description of the appearance and distribution of primary lesions and additional symptoms [18,19].

Objectives

We recently described an interactive diagnosis ranking algorithm based on high-level, symbolic representations of structured descriptions of dermatologic conditions by human readers [20]. Herein, we want to assess this algorithm in a vignette-based study simulating a potential application in tele-dermatology decision support. The major goals of this pilot study were to assess the baseline performance of such an algorithm and to explore typical problems of human-computer interaction.

Methods

A reasoning based clinical diagnosis-ranking algorithm (CDRA) was used as an example for an interactive diagnostic system based on high-level, human readable, symbolic logic [20]. Five physicians with varying experience in clinical dermatology independently rated 58 consecutive patient vignettes (virtual test cases). The raters input consisted of structured descriptions of the dermatologic conditions presented in the vignettes. The descriptions were entered into the software via a simple multiple-choice questionnaire, resulting in ranked lists of differential diagnoses.

Clinical diagnostic ranking algorithm

The CDRA uses a custom dermatological knowledge database, containing 620 different dermatologic diagnoses at the time of conducting the study, as described recently [20]. Briefly, it provides probability-ranked differential diagnoses through a reasoning component, based on computational logic. The user interface in this study was a simple questionnaire that allowed users to enter the following information: 1) basic epidemiologic information (patient sex, age, skin type, number of lesions); 2) arrangement of lesion(s) (information regarding multiplicity, distribution and arrangement of the lesions); 3) localization of lesion/s in anatomic areas (including special sites such as sun-exposed areas); 4) morphology of lesion(s); 5) color of lesions; 6) timing and onset of the disease; 7) additional non-cutaneous signs and symptoms. The participants did not receive any additional information or exemplar cases of primary lesions. After completing the input, the algorithm creates a ranked list of all 620 diagnoses in the background. The software generates up to 8 “top-ranked” diagnoses and an arbitrary number of “excluded diagnosis”. No correction of data entry after the first submission was permitted or possible, and users did not see ranked lists at any point.

Rater characteristics and training

Four dermatologists-in-training and 1 board-certified dermatologist from a single center served as independent raters (Supplementary Table 1). Dermatologists in-training were ranked by post-graduate years (PGY-1 to PGY-4). Before entering any study-specific information, all raters were trained on the technical data entry process of the software. Raters received individual user access for the software and a pdf-file containing all virtual patients in random order. Every rater had a separate computer workstation and no time constraints for entering the information into the CDRA.

Table 1. Performance of the CDRA with different users. MRR: Mean Reciprocal Rank. P-Value denotes paired Wilcoxon Signed-Rank test, comparing the diagnosis ranks of a dermatology resident to those of a board-certified dermatologist (Reference).

Rater	Median Rank Position of the Correct Diagnosis	MRR	P
PGY-1	2.00 (IQR: 21.50)	0.514	< 0.001
PGY-2	5.00 (IQR: 239.00)	0.355	< 0.001
PGY-3	2.00 (IQR: 2.75)	0.597	0.003
PGY-4	2.00 (IQR: 4.00)	0.557	0.003
Board-certified	1.00 (IQR: 1.00)	0.757	Reference

IQR = interquartile range; PGY = post-graduate year of dermatology residency.

Vignettes

The convenience sample was collected from educational material of the Medical University of Vienna, and contained 58 virtual patient cases including common dermatologic diseases but also more rare conditions, if they seemed relevant for a primary care setting. Fitzpatrick skin types, as assessed by a single author based on digital images, were 93.1% I-II (N = 54), 5.2% III-IV (N = 3), and 1.7% V-VI (N = 1). A complete list of diagnoses alongside basic patient information is shown in Supplementary Table 2. Vignettes included a brief medical history covering only the main points, and between one and four representative clinical images involving overviews of different body parts and, if necessary, close-up images of individual skin lesions. The views were selected to allow evaluation of morphologic features of the primary lesions as well as their distribution, arrangement and color. The 58 vignettes covered a range of different disease categories including allergic, autoimmune, benign neoplastic, exogenous, hereditary, infections, inflammatory, malignant neoplastic, and other diseases like melasma or amyloidosis). Of the 58 vignettes, 31 contained information about non-cutaneous signs and symptoms.

Statistical analysis

A single correct diagnosis served as the ground truth for each vignette. We used the median correct ranking position and the Mean Reciprocal Rank (MRR) to estimate the ranking ability of the algorithm. The Reciprocal Rank is defined as 1/k, where k is the rank position of the correct diagnosis as predicted by the CDRA. The MRR is the mean across all cases. We calculated the Sørensen-Dice-coefficient (Dice) to measure the similarity between descriptions. Paired comparisons of rank positions were performed with the Wilcoxon Signed-Rank test. Confidence intervals (CI) and interquartile range (IQR) are reported where applicable. We used R Statistics (version 4.1.0) for all statistical analyses and applied a Bonferroni-Holm correction to all p-values [21,22]. A two-sided P value < 0.05 indicates statistical significance. Plots were created using ggplot2 [23].

Table 2. Similarity of descriptions of residents compared to the corresponding descriptions of a board-certified dermatologist according to subsections. Results are pooled over all users and cases, lowest values are highlighted in bold.

Questionnaire Section	Dice (mean)
Arrangement	0.75 (95% CI: 0.72-0.79)
Color	0.75 (95% CI: 0.71-0.79)
Epidemiology	0.96 (95% CI: 0.94-0.97)
Localization	0.72 (95% CI: 0.69-0.75)
Morphology	**0.57 (95% CI: 0.54-0.60)**
Signs and Symptoms	0.85 (95% CI: 0.81-0.89)
Time	**0.62 (95% CI: 0.58-0.65)**

CI = confidence interval

Results

Fifty-eight vignettes described by five raters, with one entry of rater PGY-2 missing through a technical error, resulted in 289 probability-ranked diagnosis lists. For all raters, the correct diagnosis was top-1 ranked in most vignettes (Figure 1). While most rankings following inputs of more experienced users fell into the top-8 ranks, inputs by younger participants (PGY-1 & PGY-2) frequently resulted in a very low ranking of the correct diagnosis (> 128; Figure 1). The mean reciprocal rank of the algorithm was 0.757 when applied by the board-certified dermatologist, and significantly lower when applied by residents (Table 1). The highest MRR was measured for benign (0.68; 95% CI: 0.35-1.01) and inflammatory (0.68, 95% CI: 0.46-0.90), the lowest for autoimmune (0.39; 95% CI: 0.02-0.76) and exogenous (0.43; 95% CI: 0.30-0.56) diseases.

Similarity of data entry

The vignettes' descriptions of the four residents were compared with those of the board-certified dermatologist for similarity. The median Dice-score ranged from 0.64 (95% CI: 0.60-0.67; PGY-2) to 0.74 (95% CI: 0.71-0.77; PGY-3; Suppl. Figure 1). Furthermore, we analyzed the similarities

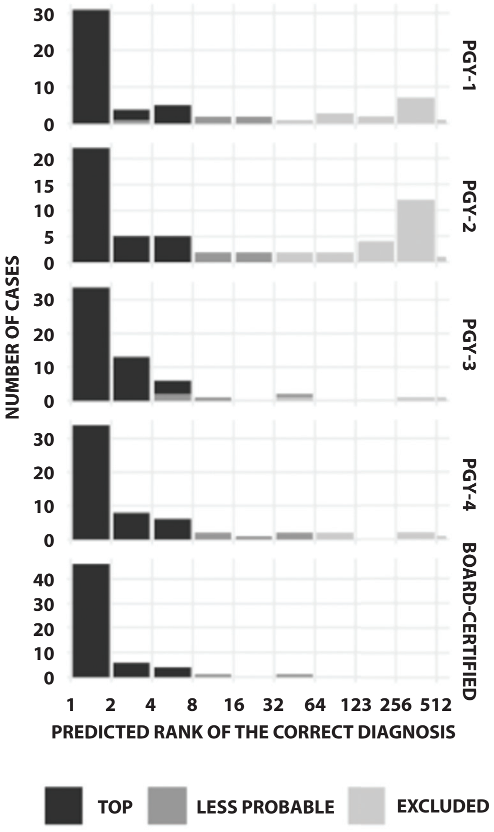

Figure 1. Histogram of ranking positions of the correct diagnosis. Black color denotes the categorization as a highly relevant differential diagnosis, the brightest gray as "excluded".

between the pooled ratings of the residents and the board-certified dermatologist for different subsections of the questionnaire. Inputs for the sections "epidemiology" and "signs & symptoms" obtained the highest average similarity between residents and the board-certified dermatologist (Dice 0.96, [95% CI: 0.94-0.97] and 0.85 [95% CI: 0.81-0.89], respectively). We observed the lowest Dice scores for descriptions of morphology, arrangement and time (Table 2).

Similarity of inputs regarding morphology and time

Descriptions for primary lesions ("elevation", "plane", "even") and surface changes ("crust", "erosion") were used consistently, whereas descriptions of consistency ("firm", "soft", "indurated") were more ambiguous (Supplementary Figure 2A). Regarding the section of time and disease course (Suppl. Figure 2 B), the terms "recurrent" and "progressive" were used consistently, while the similarity of inputs for the terms "limited", "self-limited", and "transient" was rather low.

Influence of users input on performance of the algorithm

Complete omission of subsections of the questionnaire deteriorated ranking results. The decrease in performance was most pronounced for the subsections on anatomic site and morphology (Figure 3; Supplementary Table 3). In a small

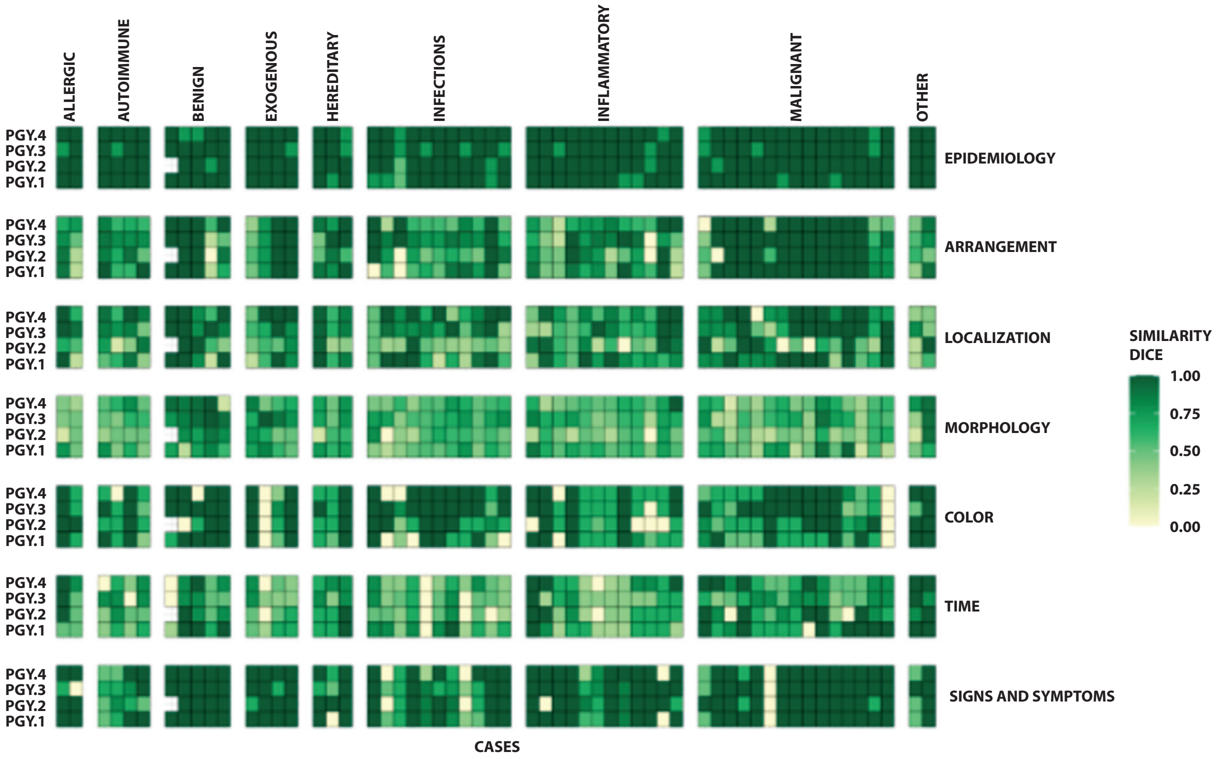

Figure 2. Similarity of descriptions of residents with the corresponding descriptions of a board-certified dermatologist according to subsections. Columns denote a single case, column groups denote grouping of cases to a diagnostic category. Row groups denote description groups within the data entry form.

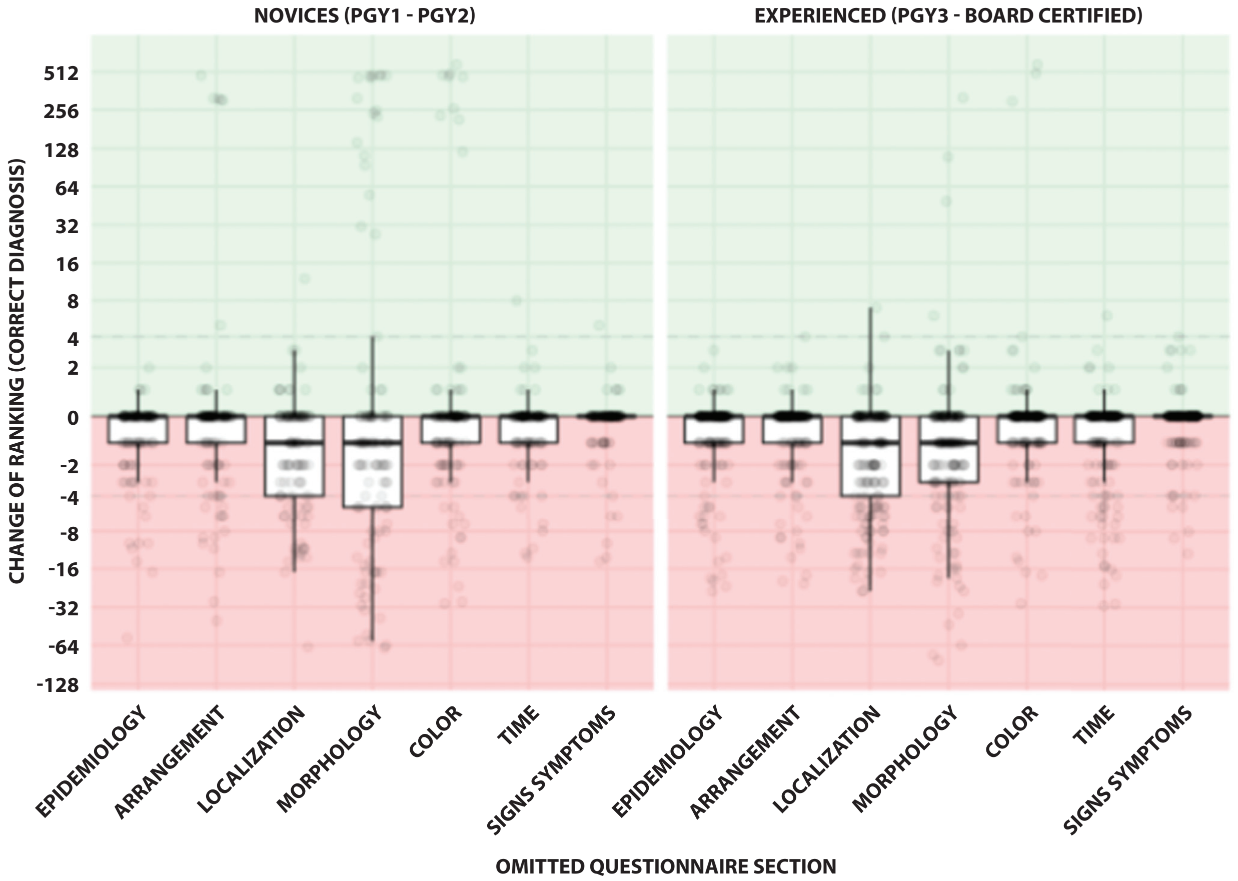

Figure 3. Rank changes after omission of specific subsections of the questionnaire. Participants were grouped according to experience into novices (left panel; PGY-1 and PGY-2) and experienced users (right panel; PGY-3, PGY-4 and board-certified dermatologist). Dots denote change of rank of the correct diagnosis for one query of one user, boxplots denote median and IQR. The green area highlights changes to a better position, the red area to a worse position.
IQR = interquartile range; PGY = post-graduate year of dermatology residency.

subgroup of vignettes, omission of inputs on morphology and color from novices improved the rankings.

Conclusions

In this pilot study we conducted an experimental validation of a simple diagnosis ranking algorithm based on comprehensive and structured clinical descriptions provided by physicians. If applied by an experienced user, the algorithm top-ranked the correct diagnosis in the majority of cases. The median rank of the correct diagnosis was not below the fifth position even for the least experienced participant (Figure 1). This means that in a typical use case the correct diagnosis will be included in the first eight ranked diagnoses. The measured accuracy of our approach outperformed similar algorithms in general medicine, in which the top-5 results included the correct diagnosis in about 50% of cases [24]. Our results are in line with other promising reports of clinical decision support systems for dermatology and provide further evidence that a logic-based, interactive diagnosis ranking algorithm may be a useful tool in clinical practice, especially for primary care and in the setting of store-and-forward tele-dermatology [18,25].

We further demonstrate that human understandable symbolic AI fed by human inputs could be a worthwhile alternative to deep learning algorithms for image based diagnostic dermatology. In contrast to deep learning, the rules of symbolic AI are derived from expert knowledge, which facilitates explainability. Aside from that, the rules can be easily adjusted, if errors occur. The disadvantage of this approach, however, is the reproducibility of human inputs.

To better pin down potential sources of noise and bias introduced by user inputs, we studied the impact of user expertise on ranking and the similarity of user inputs for corresponding cases. Finally, we also performed a sensitivity analysis to test the robustness of rankings if parts of the clinical description are either missing or misleading. In this respect, we found that the algorithm is most vulnerable to omissions of sections regarding morphology and localization.

We further found that the rank of the correct diagnosis significantly decreased if less experienced users were responsible for the input. In some cases, the correct diagnosis was even excluded (Figure 1). As the CDRA is structured in sections simulating a bottom-up dermatologic work-up starting with descriptions of primary lesions, we were able to decipher the reasons for these errors in most cases. Considering the inputs of the board-certified dermatologist as the reference standard, the residents' descriptions were most similar to the reference standard for questions regarding epidemiology, age group, skin type, and additional symptoms (Figure 2). Not unexpectedly, the most ambiguous parts of the questionnaire were the subsections covering morphology and timing.

Haptic elements such as induration were used inconsistently, which can be easily explained by the virtual setting which makes palpation impossible (Supplementary Figure 2A). Follow-up studies with live patients will be necessary to determine whether such elements should be entirely removed from the algorithm or omitted only in image-based case presentations. Analysis of user inputs referring to timing demonstrated that terms describing the course of the disease ("recurrent", "progressive", "chronic"; Supplementary Figure 2B) were used rather consistently, but not terms related to resolution ("transient duration", "self-limited", "limited"). The explanation may be that experienced users will already know the correct diagnosis and may fabricate a description that is in line with the correct diagnosis, even if the information given in the vignette or by the patient is ambiguous. This points to a limitation of our study since we did not compare diagnostic rankings of users with and without support by the algorithm. The aims of this pilot study, however, were to investigate whether the algorithm is principally feasible for clinical use and to improve the logic of the algorithm and the composition of the questionnaire upon the results of this small-scale experiment, if necessary. Our results show that the performance of the algorithm will depend on the quality of user inputs. To improve the evolution of this and similar algorithms, developers need to focus not only on machine learning issues but also on the user interface and how to minimize noise and bias. The results of our study indicate that it is crucial to select variables that are equally robust and relevant. The number of variables and the time spent for data input will significantly impact the user friendliness of the interface. Poor user friendliness and time efficacy constitute important barriers for deploying such systems in primary care [26]. Furthermore, we learnt from this pilot study that the number of variables and the granularity of descriptions were probably too high, which had the adverse effect of increasing noise while decreasing accuracy.

This was a small-scale pilot study using a convenience sample and vignettes instead of live consecutive patients. The study included only dermatologists, either board-certified dermatologists or dermatology residents, and did not include main target users such as primary care physicians or nurses. Because of the small number of raters included, especially quantitative findings should be verified in larger follow-up studies. As baseline accuracy of raters was not measured, applicability and added value in a clinical setting could not be estimated. The range of skin types of patients in the vignettes was biased towards lighter skin and skin of color was underrepresented, which may limit the generalizability of our findings [27].

In conclusion, we demonstrated that our previously described clinical diagnosis ranking algorithm performed well across a wide range of dermatologic. In our small rater group, we found inconsistent input from inexperienced users, who are an important target population of this algorithm, introduced noise and bias and decreased its performance.

Acknowledgements

We want to thank the residents that participated in this study, Arno Lukas from emergentec biodevelopment GmbH for the help in creating the initial graphical user interface of the algorithm, and Gernot Salzer as well as Rodriguez Dominguez Rosa María for the algorithm testing tool.

Ethics approval

The study was reviewed and approved by the Institutional Review Board of the Medical University of Vienna, Austria (protocol-no: 1758/2013), and conducted in accordance with the Helsinki Declaration of 1975, as revised in 1983.

Patient consent

Not required.

References

1. Lowell BA, Froelich CW, Federman DG, Kirsner RS. Dermatology in primary care: Prevalence and patient disposition. *J Am Acad Dermatol*. 2001;45(2):250-255. DOI: 10.1067/mjd.2001.114598. PMID: 11464187.
2. Verhoeven EWM, Kraaimaat FW, van de Kerkhof PCM, et al. Prevalence of physical symptoms of itch, pain and fatigue in patients with skin diseases in general practice. *British Journal of Dermatology*. 2007;156(6):1346-1349. DOI:10.1111/j.1365-2133.2007.07916.x. PMID: 17535233.
3. Hay RJ, Johns NE, Williams HC, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *J Invest Dermatol*. 2014;134(6):1527-1534. DOI: 10.1038/jid.2013.446. PMID: 24166134.
4. Hay RJ, Augustin M, Griffiths CEM, Sterry W, Board of the International League of Dermatological Societies and the Grand Challenges Consultation groups. The global challenge for skin health. *Br J Dermatol*. 2015;172(6):1469-1472. DOI: 10.1111/bjd.13854. PMID: 26036149.
5. Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med*.

2020;26(8):1229-1234. DOI: 10.1038/s41591-020-0942-0. PMID: 32572267.

6. Polesie S, Gillstedt M, Kittler H, et al. Attitudes towards artificial intelligence within dermatology: an international online survey. *Br J Dermatol*. 2020;183(1):159-161. DOI: 10.1111/bjd.18875. PMID: 31953854.
7. Lallas A, Argenziano G. Artificial intelligence and melanoma diagnosis: ignoring human nature may lead to false predictions. *Dermatol Pract Concept*. 2018;8(4):249-251. DOI: 10.5826/dpc.0804a01. PMID: 30479851. PMCID: PMC6246056.
8. di Ruffano LF, Takwoingi Y, Dinnes J, et al. Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults. *Cochrane Database Syst Rev*. 2018; 12(12):CD013186.. DOI: 10.1002/14651858.CD013186. PMID: 30521691. PMCID: PMC6517147.
9. Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of Computer-Aided Diagnosis of Melanoma: A Meta-analysis. *JAMA Dermatol*. 2019;155(11):1291-1299. DOI: 10.1001/jamadermatol.2019.1375. PMID: 31215969. PMCID: PMC6584889.10.
10. Ferrara G, Argenyi Z, Argenziano G, et al. The influence of clinical information in the histopathologic diagnosis of melanocytic skin neoplasms. *PLoS One*. 2009;4(4):e5375. DOI: 10.1371/journal.pone.0005375. PMID: 19404399. PMCID: PMC2671836.
11. Cerroni L, Argenyi Z, Cerio R, et al. Influence of evaluation of clinical pictures on the histopathologic diagnosis of inflammatory skin disorders. *Journal of the American Academy of Dermatology*. 2010;63(4):647-652. DOIi:10.1016/j.jaad.2009.09.009. PMID: 20846566.
12. Binder M, Kittler H, Dreiseitl S, Ganster H, Wolff K, Pehamberger H. Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process. *Melanoma Res*. 2000;10(6):556-561. DOI: 10.1097/00008390-200012000-00007. PMID: 11198477.
13. Yap J, Yolland W, Tschandl P. Multimodal Skin Lesion Classification using Deep Learning. *Exp Dermatol*. 2018;27(11):1261-1267. DOI:10.1111/exd.13777. PMID: 30187575.
14. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26(6):900-908. DOI: 10.1038/s41591-020-0842-3. PMID: 32424212.
15. Höhn J, Hekler A, Krieghoff-Henning E, et al. Integrating Patient Data Into Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. *J Med Internet Res*. 2021;23(7):e20708. DOI: 10.2196/20708. PMID: 34255646. PMCID: PMC8285747.16.
16. Pacheco AGC, Krohling R. An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE J Biomed Health Inform*. 2021;25(9):3554-3563. DOI: 10.1109/JBHI.2021.3062002. PMID: 33635800.17.
17. Tognetti L, Bonechi S, Andreini P, et al. A new deep learning approach integrated with clinical data for the dermoscopic differentiation of early melanomas from atypical nevi. *J Dermatol Sci*. 2021;101(2):115-122. DOI: 10.1016/j.jdermsci.2020.11.009. PMID: 33358096.
18. Chou W-Y, Tien P-T, Lin F-Y, Chiu P-C. Application of visually based, computerised diagnostic decision support system in dermatological medical education: a pilot study. *Postgrad Med J*. 2017;93(1099):256-259. DOI: 10.1136/postgradmedj-2016-134328. PMID: 27591194.
19. Holubar K. Ferdinand von Hebra 1816--1880: on the occasion of the centenary of his death. *Int J Dermatol*. 1981;20(4):291-295. DOI: 10.1111/j.1365-4362.1981.tb04341.x. PMID: 7016771.
20. Salzer G, Ciabattoni A, Fermüller C, et al. Dermtrainer: A Decision Support System for Dermatological Diseases. *arXiv [csIR]*. Published online July 1, 2019. DOI: 10.48550/arXiv.1907.00635
21. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2019. Available from: https://www.R-project.org/
22. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand Stat Theory Appl*. 1979;6(2):65-70.
23. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2016.
24. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. *PLoS One*. 2021;16(7):e0254088. DOI: 10.1371/journal.pone.0254088. PMID: 34265845. PMCID: PMC8282353.
25. Gilbert S, Mehl A, Baluch A, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. *BMJ Open*. 2020;10(12):e040269. DOI: 10.1136/bmjopen-2020-040269. PMID: 33328258. PMCID: PMC7745523.
26. Burke MD, Savard LB, Rubin AS, Littenberg B. Barriers and facilitators to use of a clinical evidence technology in the management of skin problems in primary care: insights from mixed methods. *J Med Libr Assoc*. 2020;108(3):428-439. DOI: 10.5195/jmla.2020.787. PMID: 32843874. PMCID: PMC7441913.
27. Groh M, Harris C, Soenksen L, et al. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021;1820-1828. DOI: 10.48550/arXiv.2104.09957.