DOI: 10.1111/jdv.16002 JEADV

ORIGINAL ARTICLE

Observer-independent assessment of psoriasis-affected area using machine learning

N. Meienberger,[1] F. Anzengruber,[1] L. Amruthalingam,[2] R. Christen,[3] T. Koller,[3] J.T. Maul,[1] M Pouly,[3] V. Djamei,[1] A.A. Navarini[1,2,*]

[1]Department of Dermatology, University Hospital Zurich, Zurich, Switzerland
[2]Department of Dermatology, University Hospital of Basel, Basel, Switzerland
[3]Lucerne University for Applied Sciences and Arts, Lucerne, Switzerland
*Correspondence: A.A. Navarini. E-mail: alexander.navarini@usb.ch

Abstract

Background Assessment of psoriasis severity is strongly observer-dependent, and objective assessment tools are largely missing. The increasing number of patients receiving highly expensive therapies that are reimbursed only for moderate-to-severe psoriasis motivates the development of higher quality assessment tools.

Objective To establish an accurate and objective psoriasis assessment method based on segmenting images by machine learning technology.

Methods In this retrospective, non-interventional, single-centred, interdisciplinary study of diagnostic accuracy, 259 standardized photographs of Caucasian patients were assessed and typical psoriatic lesions were labelled. Two hundred and three of those were used to train and validate an assessment algorithm which was then tested on the remaining 56 photographs. The results of the algorithm assessment were compared with manually marked area, as well as with the affected area determined by trained dermatologists.

Results Algorithm assessment achieved accuracy of more than 90% in 77% of the images and differed on average 5.9% from manually marked areas. The difference between algorithm-predicted and photograph-based estimated areas by physicians was 8.1% on average.

Conclusion The study shows the potential of the evaluated technology. In contrast to the Psoriasis Area and Severity Index (PASI), it allows for objective evaluation and should therefore be developed further as an alternative method to human assessment.

Received: 16 June 2019; Accepted: 20 September 2019

Conflicts of interest

The authors declare no conflict of interest.

Funding sources

Regierungsrat Kanton Zürich, Bruno Bloch Foundation, Promedica Stiftung Chur, Forschungskredit Universität Zürich, Novartis.

Introduction

Multiple new biologicals have revolutionized the treatment of psoriasis patients. But even though the patents for some of the drugs have expired and biosimilars are in development, the costs are expected to remain very high.[1] So, it will stay economically impossible for all the patients to receive such treatments. Current guidelines recommend to base treatment decisions on the body surface area (BSA) and Psoriasis Area and Severity Index (PASI), and in most countries, only patients with moderate-to-severe psoriasis will receive reimbursement of the more expensive drugs.[2] The threshold for moderate-to-severe psoriasis is >10% BSA affected by psoriasis. It has however already been shown by multiple researchers that these scores have significant weaknesses, the most severe of all not being objective.[3–6]

An objective, computer-based and automatic scoring method would be fairer, furthermore timesaving, and could even be more exact than human evaluation in the long run. Esteva *et al.*,[7] who have already achieved dermatologist-level results by using a machine learning algorithm for detection of skin cancer have shown that neural networks are the future of skin-pattern analysis.[7] There is however still a lack of research on neural network-based machine learning algorithms to assess psoriatic skin.

In this study, we propose and evaluate a neural network especially trained to detect psoriatic lesions on photographs and compare the results to the affected areas estimated by physicians, which is the basis of the PASI.

Material and methods

Study design

This retrospective, non-interventional, single-centred, interdisciplinary study was performed as a cooperation between the Department of Dermatology at the University Hospital of Zurich and the School of Information Technology at the Lucerne University of Applied Sciences and Arts (HSLU).

The ethics application (BASEC-Number 2017-01388) was approved by the cantonal ethics committee of Zurich on 10 January 2018.

Study objectives

The primary objective of the study was to compare psoriasis lesion detection done by neural networks, one trained with an unweighted objective function and one trained with a penalty factor on false-predictions of diseased regions, to the manually marked psoriasis lesions on the same images using accuracy, F1-score and difference in area.

Secondary outcomes were as follows: (i) the comparison of psoriasis lesion detection to manually marked area by the different weight algorithm on images with 50% of the original quality using the scores above, (ii) the comparison of psoriasis lesion detection by the different weight algorithm on images with 25% of the original quality to manually marked area using the scores above, (iii) the comparison of live estimated affected area, photograph-based estimation of affected area, manually marked affected area and algorithm-predicted affected area using intraclass correlation (ICC) and mean absolute difference in area.

Data set (inclusion/exclusion criteria)

A total of 203 photographs of Caucasian patients, aged between 18 and 80 years old and suffering from plaque-type psoriasis, were selected. The photographs included were taken with a Nikon D700 camera by the in-house photographer and had a resolution of 8–14 megapixel. To be included, the photographs had to be standardized frontal or dorsal shots of either the lower body or the upper body without head, in a neutral position. Twenty-eight patients that had a frontal and a dorsal shot of either the lower or the upper body, fulfilling the criterias above, and a precise PASI taken the same day, were selected. This resulted in a test set of 56 photographs. All physicians performing PASI assessment had more than 3 years of experience and were supervised by a senior dermatologist. All the patients chosen for the test set were not yet featured in one of the 203 photographs from the training set.

Data preparation

The psoriatic areas on all of the photographs chosen were marked using SkinWebApp, developed and made available by HSLU.

Convolutional networks, like ours, benefit from parallel processing of many pixels on the Graphical Processing Units (GPU). However, it is undesirable to process the pixels from only one image in one training step, as it would only optimize for this image. In order to mix pixels from different images, but still benefit from the parallel processing, we divided the images into smaller image patches of size 64×64 pixels and processed a batch of those image patches in each step.

Patches with background coverage of more than 95% were discarded before being processed by the algorithm, so only skin surface would be assessed.

Neural network architecture and hyperparameters

We used a supervised deep-learning approach, designed by the School of Information Technology at HSLU, for this study. This fully convolutional neural network called Net16 uses a residual connection architecture as introduced by He *et al.*[8] It consists of a 3×3 convolutional layer with depth 16, followed by five residual blocks with the same depth 16 and a final convolutional layer with depth 32 before the 1×1 logits and the softmax layer as before. The number of patches in one batch of data chosen was 512, and the neural network was trained for 1200 epochs.[9] In machine learning, learning tends to benefit when under-represented features are given more weight. We therefore trained the algorithm with threshold 0.5, once using different weights (background = 1.0, healthy = 1.0, psoriasis = 2.5) and once using same weights (background = 1.0, healthy = 1.0, psoriasis = 1.0) in the objective function to see if this influences our results.

Evaluation

A fivefold cross-validation was done, where 80% of the 203 marked photographs were used for training and the remaining 20% for validation of the trained model. The 56 marked photographs of the 28 patients set aside were only as a final test data set and nor for training nor validation. To test algorithm performance on lower quality images, the test set photographs were additionally scaled down to 25% and 50% of their original resolution by decreasing the pixels and used again as a second and third test set for the algorithm. The algorithm detected individual pixels, so each pixel was to be a true positive (TP) if was correctly assigned to be non-healthy, true negative (TN) if correctly assigned healthy and false positive (FP) and false negative (FN) if mistakenly assigned non-healthy and healthy, respectively. To test for accuracy and the F1-score, manually marked areas were regarded as gold standard when compared to algorithm-predicted areas. We assume manually marked area to be the most accurate photograph-based assessment method, as other researchers improved the power of clinical trials through computer-aided skin lesion assessment using manual selection.[10] To put results into context, live estimated affected areas displayed in the images were retrieved from the precise PASI scores as mentioned before and compared with manually marked areas, algorithm-predicted areas and areas estimated based on photographs, respectively. For this, mean affected areas of the corresponding frontal and dorsal shots were calculated for manually marked and predicted areas. In leg shots, results could be directly compared with physicians' estimates. For upper body images, physicians' estimations of arms and torso were combined using the ratios (2*(0.2*area arms)(0.3*area torso)), as established in the precise PASI.

Figure 1 Example of algorithm predictions compared with manual marked area. ■ green = false negative, not algorithm predicted but manually marked. ■ pink = true positive, algorithm predicted and manually marked. ■ purple = false positive, algorithm predicted but not manually marked.

Data analysis

The accuracy of predictions can be calculated on a single pixel level as TP + TN/(TP + TN + FP + FN). Because our data set has much more healthy, than non-healthy pixels, this is not a sufficient measure.

We therefore collected data for sensitivity (TP/(TP + FN)) and precision (TP/(TP + FP)) and calculated the F1-score, being the harmonic mean of the two measures.

To assess the agreement between all of our four assessment methods, a Bland–Altman plot was used, which shows the differences between measurement methods for each measure against the mean of the measurement methods.

Results

Full resolution, weights: 1,1,2.5

The overall F1-score for the full resolution images was 0.71 with the algorithm trained with different weights (background = 1.0, healthy = 1.0, psoriasis = 2.5). This was the best F1-score achieved in all our tests. The overall accuracy achieved on single pixel level was 0.91. When accuracy was calculated for each image individually, the mean accuracy was 0.92 (95% CI 0.89–0.94). When the same was done for the F1-score, a mean of 0.53 (95% CI 0.47–0.61) was reached. The mean difference in area was 5.9 percentage points (95% CI 3.8–8.1) on a single image basis. On the test set overall, a mean area of 13.2% was manually marked, whilst the algorithm tended to overestimate, predicting a mean area of 16.9% over the whole test set.

Figure 1 shows a good assessment result using this algorithm, with an accuracy of 0.88 and an F1 being 0.89. The manually marked area in this figure was 49.3%, whilst the predicted area was 60.1%, resulting in a difference of 10.8 percentage points.

Full quality, weights: 1,1,1

The overall accuracy on a single pixel level was with 0.93 slightly better in the same weight test. However, the overall F1-score of 0.69 achieved by the algorithm trained with same weights was slightly lower than that for the algorithm using different weights. On average, the area manually marked was 13.1% with this algorithm, whilst the average area predicted was 10.5% across the whole test set. However, the mean difference in area was 5.2 percentage points (95% CI 3.3–7.2) on a single image basis. When accuracy was calculated for each image individually, the mean accuracy was 0.93 (95% CI 0.90–0.93). When the same was done for the F1-score, a mean of 0.51 (95% CI 0.43–0.57) was reached.

50% resolution images results

In this third test, the test set images were scaled down to 50% of their original resolution and then evaluated by the algorithm trained with different weights. Overall accuracy of psoriasis lesion detection in this test was 0.92, thus higher than for the full quality images. The overall F1-score, however, was 0.69 and thus slightly lower than in the full quality test set. Overall, the manually marked area was on average 13.2% in this test, whilst the average area predicted was 12.9% across the whole test set. However, the mean difference in area was 5.1 percentage points (95% CI 3.3–6.8) on a single image basis. When accuracy was calculated for each image individually, the mean accuracy was 0.92 (95% CI 0.90–0.94). When the same was done for the F1-score, a mean of 0.48 (95% CI 0.41–0.56) was reached.

25% resolution images results

When the images were scaled down to 25% of their original resolution and evaluated by the algorithm trained with different weights, the overall F1-score decreased significantly to 0.47. The overall accuracy was still high, being 0.90, as the average manually marked area was 13.5% in this test, whilst the average predicted area was 5.2%. When accuracy was calculated for each image individually, the mean accuracy was still 0.90 (95% CI 0.86–0.93). However, when the same was done for the F1-score, a mean of only 0.26 (95% CI 0.20–0.33) was reached. The mean difference in area on a single image basis was with 8.9 percentage points (95% CI 5.7–12.0) the highest of all our test set-ups.

Marked area vs. predicted area vs. live estimated area vs. photograph-based estimation of area

The areas compared in this section are retrieved from a dorsal and frontal shot for each patient as explained in the methods section and not on single image basis anymore. Thus, manually marked areas (=BSA marked) of the patients can be compared with the areas predicted by the algorithm trained with different weights on the full quality images (=BSA predicted) and to the areas estimated live during the treatment session (=BSA live), as well as to the areas estimated based on the evaluated photographs. As can be seen in Table 1, all the comparisons of assessment methods resulted in a ICC of 0.78 or more. The primary objective of this study, the comparison of algorithm predicted to area marked, showed an ICC of 0.88 (95% CI 0.76–0.94). Only the comparison of photograph-based estimation of area made by a psoriasis expert compared with manually marked area showed a slightly higher ICC, being 0.91 (95% CI 0.82–0.96). When mean differences in areas were compared on a single patient level, the comparison of algorithm predicted to manually marked area showed a mean absolute difference of 5.6 percentage points with a standard deviation (SD) of 6.9. Meanwhile, the comparison of photograph-based area estimation by an expert to manually marked area showed a mean absolute difference in area of 4.8 percentage points (SD 5.7). The findings of these comparisons are further visualized in Figs 2 and 3.

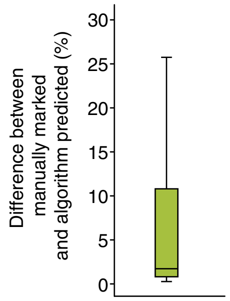

Figure 2 Discrepancy of algorithm-predicted and manually marked area.

Discussion

Main findings

Our algorithm, trained with different weights to detect psoriasis lesions, resulted in a good overall F1-score of 0.71 and an excellent accuracy of 0.91. The overall F1-score from the 50% resolution test set was 0.69 and thus comparable to the results of the full quality images. Only when the images were scaled down to 25% of their original resolution, the quality of the psoriasis lesion detection significantly dropped to a low F1-score of 0.47, demonstrating that a certain resolution is necessary for good results. The algorithm using the same weights achieved nearly as good results as the one using different weights in our tests. This shows that the setting of weights did not influence our outcome parameters much, even as the data were unbalanced. Our comparison of algorithm-predicted area to manually marked area resulted in an ICC of 0.87 and a mean absolute difference of 5.6 percentage points, whilst photograph-based assessment by an expert compared with manually marked area resulted in a ICC of 0.91 and a mean absolute difference of 4.8 percentage points. This is a very good result, as an ICC ranging from 0.75 to 1.00 can according to literature be interpreted as an excellent inter-rater agreement.[11]

Table 1 Table showing intraclass correlations and mean absolute differences between the different assessment methods

	ICC (95% CI)	MAD (95% CI)
Area predicted vs. area marked	0.88 (0.76–0.94)	5.6 (3.0–8.2)
Area predicted vs. live estimated area	0.78 (0.58–0.99)	8.8 (5.8–11.8)
Area predicted vs. photo based estimation	0.82 (0.64–0.91)	8.1 (5.2–11.0)
Area marked vs. live estimated area	0.87 (0.74–0.94)	6.1 (3.7–8.5)
Area marked vs. photo based estimation	0.91 (0.82–0.96)	4.8 (2.7–7.0)
Live estimated vs. photo based estimation	0.85 (0.70–0.93)	6.4 (3.6–9.2)

CI, Confidence Interval; ICC, intraclass correlation; MAD, mean absolute difference (in percentage points); SD, standard deviation.

Data analysis

In our images, the psoriatic area covered only 13% of the skin surface on average. As a consequence, if the algorithm would have graded the complete surface in all the test set images as healthy, a good accuracy of 87% would have already been achieved, even though the algorithm would be useless. This can be seen in the test set with 25% of the original resolution, where a high accuracy of 0.90 was reached by just marking nearly everything as healthy. So even though accuracy is a great assessment parameter for many statistical analyses, a second parameter, like the F1-score, also displaying precision and sensitivity, is needed in machine learning.

Results in context

To put our results into clinical context, marked affected area was compared with algorithm-predicted affected area, photograph-based estimation of affected area and live estimated affected area. As some areas can be lost trough the photographing process, live assessment cannot be directly compared with the photograph assessment methods. However, photograph-based area assessment by an expert, manual selection and predicted area have the same basis for their analysis and are therefore directly comparable. It can further be assumed that out of the photograph

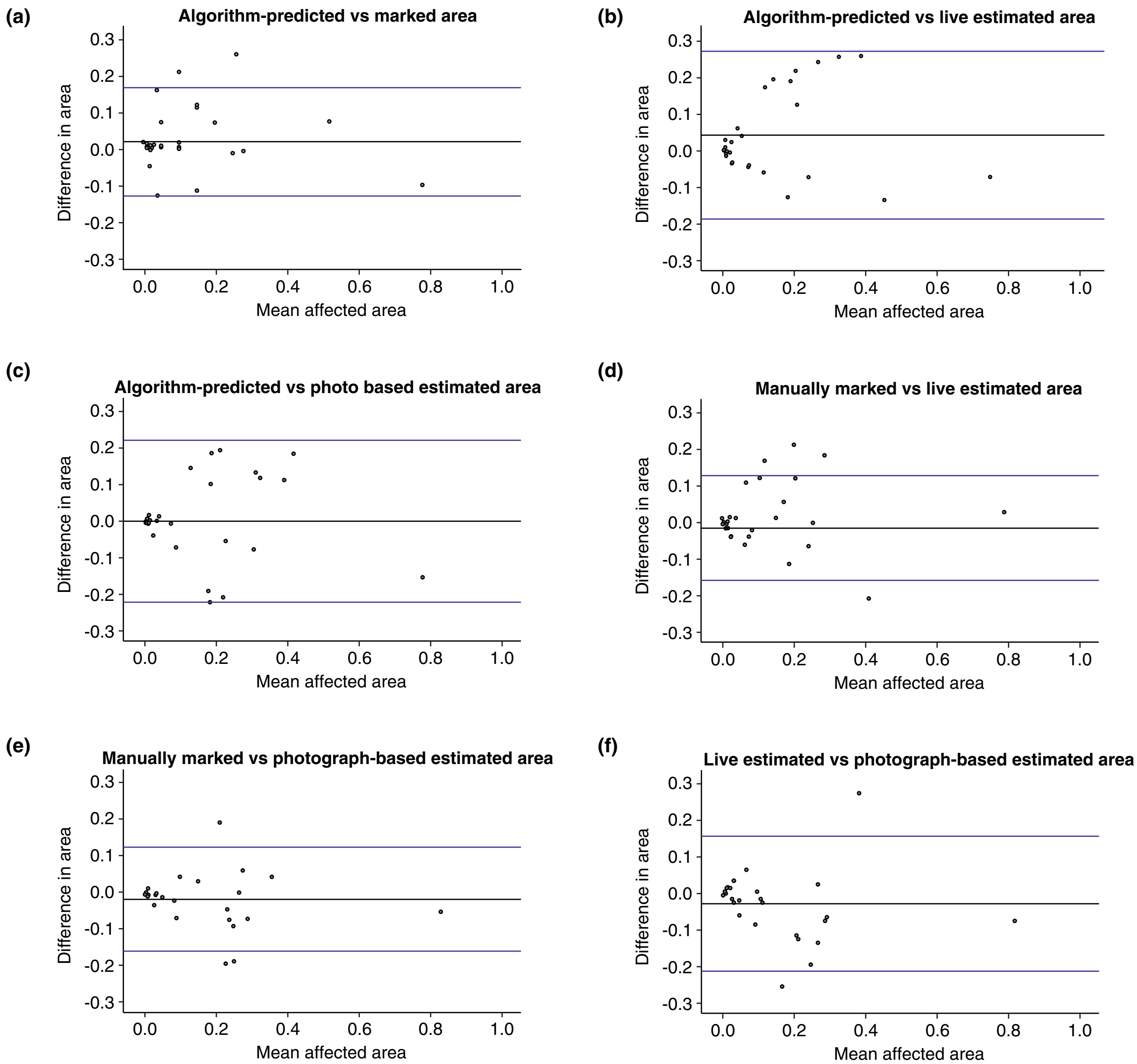

Figure 3 Bland–Altman plots showing the comparison of the different assessment methods. (a) Bland–Altman plot comparing: algorithm-predicted and manually marked area. (b) Bland–Altman plot comparing: algorithm-predicted and live estimated area. (c) Bland–Altman plot comparing: algorithm-predicted and photograph-based estimated area. (d) Bland–Altman plot comparing: manually marked and live estimated area. (e) Bland–Altman plot comparing: manually marked and photograph-based estimated area. (f) Bland–Altman plot comparing: live estimated and photograph-based estimated area.

evaluation methods the manual selection is more accurate than estimation, even if done by an expert. It is thus the gold standard in comparison with the photograph evaluation methods, but not for the comparison to live assessment. We found, that on a single patient basis, the difference of algorithm-predicted area to marked area was 5.6 percentage points on average, whereas the difference of photograph-based estimation to marked area was 4.8 percentage points on average. As the differences in areas, the ICC, of the two methods were also in the same range, with the psoriasis expert being only slightly superior to the algorithm, we believe that the machine learning approach is a legitimate alternative for psoriasis area assessment.

Strengths and limitations

Our results show that our algorithm produces adequate results, comparable to human assessment. The strongest advantage of the evaluation using artificial intelligence is its objectivity. Also, an algorithmic evaluation is always reproducible, with no inter-rater or intra-rater variability as in human assessment.

It must be taken into account that area only is measured as outcome, whilst other factors like induration, scaling and redness are neglected. However, this limitation holds true for the BSA itself, which as well does not consider the severity of the lesions and is thus doubted to be alone sufficient for psoriasis assessment.[12] Another restriction of our data is the inclusion of mostly Caucasian patients. Because the manifestation of psoriasis differs depending on the skin type, including only a few images of other skin types would have led to an highly imbalanced data set.[13] But even though solutions have been proposed to learn from imbalanced data sets, it still remains an issue in machine learning.[14] We therefore focused our study on patients of Caucasian skin tone only, and our algorithm is thus not trained for other skin types.

A further limitation that needs to be discussed is that only two images were used to assess psoriasis-affected area of either upper or lower body. This is of course not enough to depict the full surface of the human body, which is a complex structure of convex and concave areas. Since both comparison of manually marked area to live estimated area and comparison of live estimated area to photograph-based estimation resulted in an excellent ICC of 0.87 (95% CI 0.74–0.94) and 0.88 (95% CI 0.70–0.93), respectively, we speculate, however, that the areas mostly neglected from frontal and dorsal perspectives are concave areas like the axilla that are not predilection sites of plaque-type psoriasis. The mean difference between manually marked and live estimated area, as well as between live estimated and photograph-based estimated area, was also low, with 6.1 percentage points (95% CI 3.7–8.5) and 6.4 (95% CI 3.6–9.2), respectively. Further, Kreft *et al.* showed that computer-aided area assessment based on four images, dorsal and frontal shots of both upper and lower body, already improved the clinical relevance of a psoriasis study, compared to visual grading through a physician. However, a more precise solution was introduced by Fink *et al.*,[15] where 16 overlapping images were taken and overlapping areas recognized and discarded automatically, so the complete body surface would be displayed. This technology could also lead to a much more precise psoriasis area assessment in a machine learning approach. As a more simple and time efficient alternative, we propose, however, to retrieve a full BSA out of only four images. This could be done by using the ratio already established and widely accepted for the calculation of the precise PASI, thus multiplying the mean affected area of the lower body shots, displaying the legs, by 0.4 and adding the mean affected area of the upper body shots multiplied by 0.6.

Implications for research

When compared to studies creating classifiers rather than segmentation approaches, such as Esteva *et al.*,[7] our data set is small. Since it has been shown that machine learning results correlate with the size of data, a larger data set would be needed to achieve optimal results.[16,17] The collection of sufficient amounts of data is however difficult, and the labelling of data, as required in supervised machine learning, is time consuming and expensive. We suggest however that photographs of patients with a higher BSA score combined with photographs of completely healthy patients would provide more learning data, and thus cross-correlate with a better training result (F1-score), without taking much more time for labelling.

Further, we found in our qualitative analysis of the images that the algorithm had problems recognizing scaling as a psoriatic area. We suspect that this is due to the similar colour features of Caucasian skin and scaling. Lu *et al.*[18] already recognized this problem in 2012 and proposed an innovative algorithm focusing on skin texture rather than colour.[18] Also recognizing palms as healthy was rather difficult for the algorithm as can be seen in Fig. 1. We assume however that with a data set of sufficient size, the algorithm would be able to learn this without adding a further algorithm for the assessment.

Implications for practice

As taking four standardized photographs is a swift task, often included in the clinical routine and does not necessary have to be done by the physician himself, we believe that machine learning has the potential to reduce costs in dermatology through timesavings, whilst improving documentation of course of disease. This could also become interesting for the application in pharmaceutical studies. Therefore more attention and resources should be given to the collection of good standardized images, as it is a crucial investment for any future research using artificial intelligence. Singh *et al.*[19] already showed that bad photograph quality impacted a physician's image-based psoriasis assessment. Our results on the images with only 25% of the original image resolution show that image quality influences results in machine learning as well. Good quality, full body photography that avoids both, neglect of lesions and double-checked lesions, is thus needed to enable research and development. An aspect regarding photograph quality that still needs to be investigated is how much variations in photographing perspective influence the outcome of the area assessment.

Conclusion

A machine learning algorithm could simplify the time-consuming psoriasis assessment, and since psoriasis is a very common skin disease, with a prevalence of about 2% in Europe and North America, this could also lead to relevant reductions in health expenditure.[20] Assessment tools like the PASI and especially BSA have high overall inter-observer variation and are difficult to be reproduced correctly by others.[6,21] An artificial intelligence approach like ours would potentially annul such bias and therefore be a more adequate criterion for treatment decisions and evaluation in pharmaceutical studies. It has been shown that machine learning has the potential to even surpass human assessment, when trained with an adequate amount of data.[7] Correspondingly, machine learning has already been applied in several fields of medicine.[22,23] Our results show that even though further training and research are still needed for optimal results, machine learning should be noticed as a legitimate and objective alternative method for the assessment of psoriasis-affected area with immense potential, already achieving results comparable to human expert assessment, whilst missing inter-rater variability and being more time efficient.

References

1 Radtke MA, Augustin M. Biosimilars in psoriasis: what can we expect? *J Dtsch Dermatol Ges* 2014; **12**: 306–312.
2 Alexander N, Lasse A, Matthias A *et al.* S3 – Leitlinie zur Therapie der Psoriasis vulgaris Update 2017. 2017.
3 Jacobson CC, Kimball AB. Rethinking the Psoriasis Area and Severity Index: the impact of area should be increased. *Br J Dermatol* 2004; **151**: 381–387.
4 van de Kerkhof PC. On the limitations of the psoriasis area and severity index (PASI). *Br J Dermatol* 1992; **126**: 205.
5 Otero ME, van Geel MJ, Hendriks JC, van de Kerkhof PC, Seyger MM, de Jong EM. A pilot study on the Psoriasis Area and Severity Index (PASI) for small areas: presentation and implications of the Low PASI score. *J Dermatolog Treat* 2015; **26**: 314–317.
6 Puzenat E, Bronsard V, Prey S *et al.* What are the best outcome measures for assessing plaque psoriasis severity? A systematic review of the literature. *J Eur Acad Dermatol Venereol* 2010; **24**(Suppl 2): 10–16.
7 Esteva A, Kuprel B, Novoa RA *et al.* Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **546**: 686.
8 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Computer Vision Foundation, Las Vegas Valley, NV, 2016: 770–778.
9 Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 2010: 807–814.
10 Kreft S, Kreft M, Resman A, Marko P, Kreft KZ. Computer-aided measurement of psoriatic lesion area in a multicenter clinical trial–comparison to physician's estimations. *J Dermatol Sci* 2006; **44**: 21–27.
11 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994; **6**: 284.
12 Schmitt J, Wozel G. The psoriasis area and severity index is the adequate criterion to define severity in chronic plaque-type psoriasis. *Dermatology* 2005; **210**: 194–199.
13 Kaufman BP, Alexis AF. Psoriasis in skin of color: insights into the epidemiology, clinical presentation, genetics, quality-of-life impact, and treatment of psoriasis in Non-White Racial/Ethnic Groups. *Am J Clin Dermatol* 2018; **19**: 405–423.
14 Chawla NV, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor Newsl* 2004; **6**: 1–6.
15 Fink C, Alt C, Uhlmann L, Klose C, Enk A, Haenssle HA. Precision and reproducibility of automated computer-guided Psoriasis Area and Severity Index measurements in comparison with trained physicians. *Br J Dermatol* 2019; **180**: 390–396.
16 Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? *Clin Cancer Res* 2008; **14**: 108–114.
17 Kalayeh HM, Landgrebe DA. Predicting the required number of training samples. *IEEE Trans Pattern Anal Mach Intell* 1983; **5**: 664–667.
18 Lu J, Kazmierczak E, Manton JH, Sinclair R. Automatic segmentation of scaling in 2-D psoriasis skin images. *IEEE Trans Med Imaging* 2013; **32**: 719–730.
19 Singh P, Soyer HP, Wu J, Salmhofer W, Gilmore S. Tele-assessment of Psoriasis Area and Severity Index: a study of the accuracy of digital image capture. *Australas J Dermatol* 2011; **52**: 259–263.
20 Christophers E. Psoriasis–epidemiology and clinical spectrum. *Clin Exp Dermatol* 2001; **26**: 314–320.
21 Tiling-Grosse S, Rees J. Assessment of area of involvement in skin disease: a study using schematic figure outlines. *Br J Dermatol* 1993; **128**: 69–74.
22 Choy G, Khalilzadeh O, Michalski M *et al.* Current applications and future impact of machine learning in radiology. *Radiology* 2018; **288**: 318–328.
23 Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. *Prog Retin Eye Res* 2018; **67**: 1–29.