ORIGINAL ARTICLE

Image-based automated Psoriasis Area Severity Index scoring by Convolutional Neural Networks

M.J. Schaap,[1,*] N.J. Cardozo,[1,2] A. Patel,[2] E.M.G.J. de Jong,[1] B. van Ginneken,[2] M.M.B. Seyger[1]

[1]Department of Dermatology, Radboud University Medical Center, Nijmegen, The Netherlands
[2]Department of Medical Imaging, Radboud University Medical Center, Nijmegen, The Netherlands
*Correspondence: M.J. Schaap. E-mail: mirjam.schaap@radboudumc.nl

Abstract

Background The Psoriasis Area and Severity Index (PASI) score is commonly used in clinical practice and research to monitor disease severity and determine treatment efficacy. Automating the PASI score with deep learning algorithms, like Convolutional Neural Networks (CNNs), could enable objective and efficient PASI scoring.

Objectives To assess the performance of image-based automated PASI scoring in anatomical regions by CNNs and compare the performance of CNNs to image-based scoring by physicians.

Methods Imaging series were matched to PASI subscores determined in real life by the treating physician. CNNs were trained using standardized imaging series of 576 trunk, 614 arm and 541 leg regions. CNNs were separately trained for each PASI subscore (erythema, desquamation, induration and area) in each anatomical region (trunk, arms and legs). The head region was excluded for anonymity. Additionally, PASI-trained physicians retrospectively determined image-based subscores on the test set images of the trunk. Agreement with the real-life scores was determined with the intraclass correlation coefficient (ICC) and compared between the CNNs and physicians.

Results Intraclass correlation coefficients between the CNN and real-life scores of the trunk region were 0.616, 0.580, 0.580 and 0.793 for erythema, desquamation, induration and area, respectively, with similar results for the arms and legs region. PASI-trained physicians ($N = 5$) were in moderate–good agreement (ICCs 0.706–0.793) with each other for image-based PASI scoring of the trunk region. ICCs between the CNN and real-life scores were slightly higher for erythema (0.616 vs. 0.558), induration (0.580 vs. 0.573) and area scoring (0.793 vs. 0.694) than image-based scoring by physicians. Physicians slightly outperformed the CNN on desquamation scoring (0.580 vs. 0.589).

Conclusions Convolutional Neural Networks have the potential to automatically and objectively perform image-based PASI scoring at an anatomical region level. For erythema, desquamation and induration scoring, CNNs performed similar to physicians, while for area scoring CNNs outperformed physicians on image-based PASI scoring.

Received: 2 June 2021; Accepted: 14 September 2021

Conflict of interest

M.J. Schaap has carried out clinical trials for Amgen, Celgene, Janssen and Lilly and has acted as a paid speaker for AbbVie; fees were paid directly to the institution. N.J. Cardozo has nothing to disclose. A. Patel has nothing to disclose. E.M.G.J. de Jong has received research grants for the independent research fund of the department of dermatology of the Radboud university medical center, Nijmegen, the Netherlands, from AbbVie, Novartis, Janssen Pharmaceutica, Leo Pharma and UCB and has acted as consultant and/or paid speaker for and/or participated in research sponsored by companies that manufacture drugs used for the treatment of psoriasis including AbbVie, Janssen Pharmaceutica, Novartis, Lily, Celgene, Leo Pharma, UCB and Almirall. All funding is not personal but goes to the independent research fund of the department of dermatology of Radboud university medical center, Nijmegen, the Netherlands. B. van Ginneken is co-founder and shareholder of Thirona. M.M.B. Seyger received grants from/was involved in clinical trials from Abbvie, Amgen, Celgene, Eli Lilly, Janssen, Leo Pharma and Pfizer. She served as a consultant for Abbvie, Eli Lilly, Janssen, Leo Pharma, Novartis, UCB and Pfizer; fees were paid directly to the institution.

Funding sources

This study was funded by the Artificial Intelligence Innovation Voucher awarded by the Radboud University, Nijmegen, the Netherlands.

Introduction

Psoriasis is a common chronic inflammatory skin disease characterized by raised, erythematous and scaling lesions and is prevalent in 2–3% of the Western population.[1] Psoriasis severity is often measured with the Psoriasis Area Severity Index (PASI) score.[2] The PASI score is determined through visual inspection of the skin in four anatomical regions: the head, trunk, arms and legs. For each anatomical region, subscores are determined for erythema (0–4), desquamation (0–4), induration (0–4) and affected body surface area (0–6) resulting in an overall PASI score (0–72). The PASI score is used to determine treatment efficacy in clinical practice and clinical trials and is supportive for treatment decisions. However, the PASI score is time consuming and is known to suffer from inter- and intra-observer variability.[3,4]

Automating the PASI score could enable efficient and objective psoriasis severity assessment. Few studies report computer-aided diagnosis systems to grade the severity of psoriasis, primarily focused on image processing using different colour spaces.[5–7] Moreover, Fink *et al.* recently proposed a filter-based image processing pipeline to determine the PASI score from full body images.[8,9] In addition, for automatic segmentation and severity scoring of a single psoriasis lesion, several attempts have been made to leverage traditional machine learning algorithms, both supervised and unsupervised.[10–16] More recently, deep learning-based Convolutional Neural Networks (CNNs) have obtained state-of-the-art results in computer vision problems including image classification and image segmentation.[17] Given the superior performance of these CNNs, some attempts for deep learning-based automated severity scoring in psoriasis have been reported,[18,19] but these studies were based on the analysis of a single psoriasis plaque and only focused on one or a subset of severity scores (i.e. erythema, induration, desquamation or area).[11–13,15,16,18,19] Since the PASI score is decomposed into four subscores at an anatomical region level, lesion level analysis is inadequate for PASI scoring.

In this study, we aimed to (i) assess the performance of automated PASI scoring by CNNs using images at an anatomical region level, (ii) determine the performance and inter-rater agreement of image-based PASI scoring by PASI-trained physicians and (iii) compare the performance of image-based PASI scoring between physicians and CNNs.

Materials and methods

Data set

In this study, 5844 anonymized images were retrieved from psoriasis patients included in the Child-CAPTURE registry (Continuous Assessment of Psoriasis Treatment Use Registry), a prospective, observational, daily clinical practice cohort that follows paediatric and young adult psoriasis patients. All patients were diagnosed by a dermatologists. Patients visited the outpatient clinic of the Department of Dermatology at the Radboud university medical center, Nijmegen, the Netherlands, between October 2011 and September 2020. Standardized images were taken in an imaging studio by a medical photographer as part of clinical practice, consisting of 10 images including the arms, trunk and legs region (Fig. S1, Supporting Information). The head region was excluded and jewellery, tattoos and large naevi were cropped out to maintain anonymity. Images of each anatomical region were concatenated to represent the region as a whole (Fig. 1).

The PASI scores that were matched to the imaging series were determined in real life by the treating physician prior to capturing the images. These real-life PASI scores were considered the golden standard, also referred to as the 'ground truth'. Since the PASI scores were collected as part of daily practice, subscore distributions were imbalanced with low scores being more common than high scores (Table S1, Supporting Information). The subscore '4' for erythema, desquamation and induration, and '6' for area were excluded since the numbers were too low to accurately train CNNs. This resulted in a final data set of concatenated images of 576 trunk, 614 arm and 541 leg regions. The data sets were divided into training and test sets using an 80%:20% stratified split to ensure an identical severity distribution in each data set. Information on image preprocessing can be found in the Appendix S1 and Figure S4 (Supporting Information). All data were handled in accordance with the General Data Protection Regulation Act.

Network architectures

The performance of six deep learning models using CNN was explored and described in the Appendix S1 (Supporting Information). The Consistent Rank Logits ($\mathrm{CNN}_{\mathrm{CORAL}}$)[20] outperformed the other explored CNN architectures. The $\mathrm{CNN}_{\mathrm{CORAL}}$ is a CNN structure that takes ordinal scales into account, which is applicable to the PASI subscores. More specifically, the $\mathrm{CNN}_{\mathrm{CORAL}}$ converts a K class classification problem into K-1 binary classification sub-problems while ensuring consistent predictions between the binary sub-problems.

Training the CNNs

For training the CNNs, a single task learning approach showed superior performance compared to a multi-task learning approach. Therefore, a separate network was trained for each subscore (erythema, induration, desquamation and area) in each anatomical region (trunk, arms and legs), resulting in 12 trained CNNs. The pretrained ResNet-18[21] network initialized with the ImageNet[22] weights was used and the CNNs were fine-tuned on the current task. The Adam optimization algorithm[23] was used with a learning rate of 0.0001, weight decay of 0.01 and batch size of 8. Early stopping was used if the validation loss did not improve for more than five epochs to prevent overfitting on the training set. Horizontal flips and small random rotations (±5°) were used to make the model more robust and to avoid

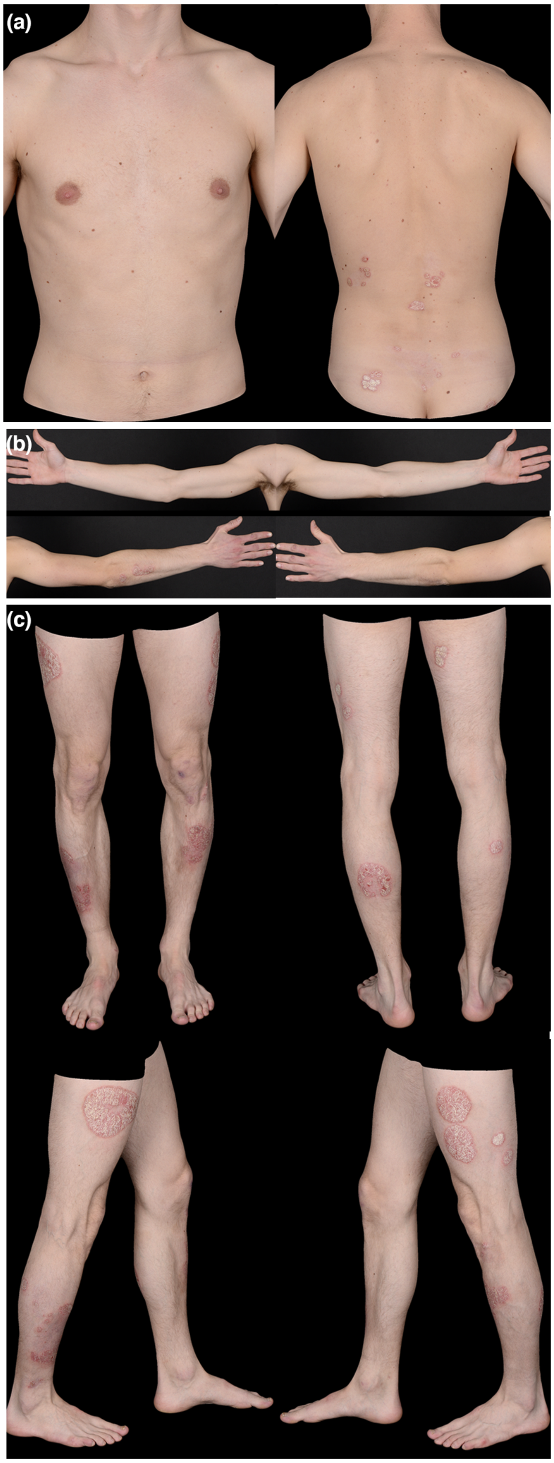


Figure 1 Concatenated images used as input for the Convolutional Neural Networks. The dorsal and ventral views of the trunk were concatenated side by side to represent the trunk region (a). The dorsal and ventral views of the right and left arm were concatenated in a 2 x 2 grid for the arms region (b). The dorsal, ventral and side views of the legs were concatenated in a 2 x 2 grid to represent the legs region (c). The images shown in this figure were pre-processed by extracting the skin and cropping the images of the arms. The patient gave written informed consent for publication of the images.

overfitting. The networks were implemented using the PyTorch[24] (version 1.8.1, CUDA version 10.2) deep learning library and were trained on a Nvidia GTX 1080Ti Graphical Processing Unit machine with 10GB of VRAM.

Image-based PASI scoring by trained physicians

A reader study was performed using the Grand Challenge online platform[25] to assess the image-based PASI scoring performance of PASI-trained physicians and to compare this to the performance of the CNNs. The trunk region was chosen for this reader study, since the CNNs achieved the highest overall performance in this anatomical region. All test set images of the trunk region, consisting of 109 images of each PASI subscore (erythema, induration, desquamation and area), were assessed by five PASI-trained physicians between 25 February and 16 March 2021. The images were shown in a random order and assessors were instructed to perform the PASI subscores at the same pace as they would do in daily practice. Physicians were blinded to the PASI subscores that were given by the treating physician based on real-life assessment.

Statistical analyses

The CNN performance was evaluated using the accuracy, mean absolute error (MAE) and intraclass correlation coefficient (ICC) with 95% confidence intervals (CI). The accuracy (0–1) measures the absolute agreement between the PASI subscores assigned by the CNN and the real-life scores (ground truth). However, the accuracy ignores the ordinal relationship between classes and is less suitable if class imbalance is present. The MAE, more widely used in ordinal regression problems, is measured as the average of the absolute difference between each of the severity subscores assigned by the CNN and the ground truth and therefore takes the ordinal relationship between classes into account.[19,20,26] The ICC reflects the agreement with the real-life scores, defined as: < 0.50, poor agreement; 0.51–0.75, moderate; 0.75–0.90, good; 0.91–1, excellent.[27]

The agreement between image-based PASI scoring by physicians and the real-life scores was calculated for each physician (N = 5) with the ICC (95% CI). For overall agreement between image-based scoring by physicians and the real-life scores, the mean ICC and mean accuracy were used. The inter-rater agreement for image-based PASI scoring between physicians was calculated with the ICC (95% CI) as well. ICCs of the subscores of the trunk region were compared between the CNN and image-based scoring by physicians (mean ICC). ICCs were calculated with two-way random effects models. All analyses were performed using Python programming language (version 3.7.1) or statistical programming language R (version 3.6.2, R Foundation for Statistical Computing, Vienna, Austria) with the package irr (version 0.84.1; https://cran.r-project.org/web/packages/irr/irr.pdf).

Results

Patient characteristics

Data of 326 patients were included in this study, resulting in 655 included imaging series as patients could have multiple imaging series over the years. These imaging series yielded images of 576 trunk, 614 arm and 541 leg regions, since some patients denied part of the anatomical region imaging. At the time of imaging, the mean age was 13.0 (±4.5) years, the median total PASI score was 6.7 (range 0.0–42.4) and there was an equal gender distribution [50.8% female (Table 1)]. Most imaging series were of patients with Fitzpatrick skin type II–III (74.5%).

CNN performance

The $\text{CNN}_{\text{CORAL}}$ reached accuracies in the trunk region of 0.660, 0.663, 0.743 and 0.734 for erythema, desquamation, induration and area scoring respectively. For the arms region, these accuracies were 0.603, 0.612, 0.681 and 0.707. For the legs region, accuracies of 0.667, 0.618, 0.676 and 0.794 were reached (Table 2). The MAE was below 0.42 for all subscores, indicating that predictions made by the $\text{CNN}_{\text{CORAL}}$ were robust (Table 2). Additionally, the performance in terms of the ICC ranged from 0.541 to 0.804 (Table 2). The ICC for erythema, desquamation and the induration subscores was in moderate agreement with the real-life scores, while the agreement between the area subscore and the real-life scores was good. In terms of accuracy, the area and the induration subscores reached the best performance, followed by the erythema and desquamation subscores (Table 2). The performance of other explored CNN architectures is shown in Table S2 (Supporting Information).

Table 1 Characteristics of included imaging series (*N* = 655)

	Imaging series (*N* = 655)
Age, years, mean (±SD)	13.0 (±4.5)
Gender, *n* (%) female	333 (50.8%)
Psoriasis severity, median (range)	
PASI†	6.7 (0.0–42.4)
BSA‡	8.0 (0.0–76.0)
PGA§	3.0 (0.0–5.0)
Fitzpatrick skin type, *n* (%)	
I	65 (9.9%)
II	330 (50.4%)
III	158 (24.1%)
IV	86 (13.1%)
V	5 (0.8%)
VI	11 (1.7%)

The 655 imaging series were yielded in 326 individual patients, who could have multiple imaging series over the years. The characteristics at the time of each imaging series are presented.
SD, Standard deviation.
†Psoriasis Area and Severity Index (PASI; range 0–72).
‡Affected Body Surface Area (BSA; range 0–100%).
§Physician Global Assessment (PGA; range 0–5).

Confusion matrices computed for the $\text{CNN}_{\text{CORAL}}$ subscore predictions of the trunk region showed that a large proportion of the test set was correctly classified (Fig. 2). Images that were misclassified were mostly classified as a neighbouring class. No evident pattern was observed regarding under- or overestimation of the severity by the $\text{CNN}_{\text{CORAL}}$ compared to the real-life scores (Table S3, Supporting Information). Additionally, the underrepresented classes (highest or lowest subscores) were more prone to misclassification. Confusion matrices of subscore predictions in the arm and leg regions yielded similar results (Figs S2 and S3, Supporting Information).

Image-based PASI scoring by trained physicians

Five PASI-trained physicians performed image-based severity scoring for 436 PASI subscores (109 examples of each subscore) of the trunk region. The ICC revealed that the PASI-trained physicians were in moderate to good agreement with each other on erythema [ICC 0.793; CI (0.739–0.842)], desquamation [ICC 0.753; CI (0.679–0.815)], induration [ICC 0.769; CI (0.708–0.824)] and area scoring [ICC 0.706; CI (0.601–0.788) (Table 2)]. In contrast, the inter-rater agreement between image-based scoring by physicians and the real-life scores was found to be lower with mean ICCs of 0.558, 0.589, 0.573 and 0.694 for erythema, desquamation, induration and area subscores respectively (Table 2). Comparing the image-based mean prediction of physicians to the real-life scoring revealed that physicians were more prone to underestimating the severity based on images, except for the score '0' (Table S3, Supporting Information).

Comparison of image-based scoring by the CNN and physicians

To be able to see the $\text{CNN}_{\text{CORAL}}$ performance in the perspective of human performance, ICCs reached by the $\text{CNN}_{\text{CORAL}}$ and physicians (image-based scoring) were compared for each subscore of the trunk region. With respect to the real-life scores, the $\text{CNN}_{\text{CORAL}}$ reached slightly higher ICCs for erythema (0.616 vs. 0.558), induration (0.580 vs. 0.573) and an evidently higher ICC for area scoring (0.793 vs. 0.694) compared to overall performance of PASI-trained physicians (Table 2). Additionally, the $\text{CNN}_{\text{CORAL}}$ outperformed all five physicians on the erythema subscore and four physicians on the induration and area subscores. The $\text{CNN}_{\text{CORAL}}$ achieved a slightly lower ICC (0.580 vs. 0.589) on the desquamation subscore than the physicians as whole, but the $\text{CNN}_{\text{CORAL}}$ still performed better than one out of five physicians (Fig. 3). Additionally, the (mean) accuracies reached by the PASI-trained physicians were considerably lower compared to the $\text{CNN}_{\text{CORAL}}$ (Table 2).

Table 2 Performance of imaged-based PASI scoring by the CNN and physicians

	CNN			**Physician image-based scoring (*N* = 5)**		
	Agreement with real-life scores			**Agreement with real-life scores**		**Inter-rater agreement**
	Accuracy†	**MAE‡**	**ICC§ (95% CI)**	**Mean Accuracy†**	**Mean ICC§**	**ICC (95% CI)**
Trunk region						
Erythema	0.660	0.367	0.616 (0.485–0.721)	0.436	0.558	0.793 (0.739–0.842)
Desquamation	0.633	0.376	0.580 (0.441–0.692)	0.533	0.589	0.753 (0.679–0.815)
Induration	0.743	0.266	0.580 (0.442–0.692)	0.612	0.573	0.769 (0.708–0.824)
Area	0.734	0.275	0.793 (0.712–0.854)	0.634	0.694	0.706 (0.601–0.788)
Arms region						
Erythema	0.603	0.408	0.614 (0.486–0.717)	–	–	–
Desquamation	0.612	0.393	0.568 (0.431–0.680)	–	–	–
Induration	0.681	0.318	0.655 (0.537–0.747)	–	–	–
Area	0.707	0.293	0.799 (0.722–0.856)	–	–	–
Legs region						
Erythema	0.667	0.352	0.711 (0.599–0.795)	–	–	–
Desquamation	0.618	0.401	0.590 (0.447–0.703)	–	–	–
Induration	0.676	0.323	0.618 (0.482–0.725)	–	–	–
Area	0.794	0.205	0.832 (0.761–0.883)	–	–	–

Performance of the Coral Convolutional Neural Network ($\text{CNN}_{\text{CORAL}}$ architecture) and PASI-trained physicians (*N* = 5) on image-based PASI scoring. The performance of the $\text{CNN}_{\text{CORAL}}$ is expressed as the accuracy, the mean absolute error (MAE) and the intraclass correlation coefficient (ICC) with respect to real-life scores. The $\text{CNN}_{\text{CORAL}}$ was separately trained for each task (erythema, desquamation, induration and area scoring) in each anatomical region. Therefore, the performance is shown on the test set images of each subscore in each anatomical region. The test sets consisted of concatenated imaging series of 109 trunk, 116 arm and 102 leg regions. Performance of image-based scoring by the PASI trained physicians was only assessed for the trunk region test set images. Agreement between image-based scoring by physicians and real-life scoring by the treating physician was calculated using the mean accuracy and mean ICC. Inter-rater agreement of the five physicians on image-based scoring is expressed by the intraclass correlation coefficient (ICC) and corresponding 95% confidence intervals (CI).

†The accuracy (0–1) measures the absolute agreement between the severity subscores assigned by the CNN and the ground truth (real-life PASI scoring). A score of 1 means that all scores were correctly classified.

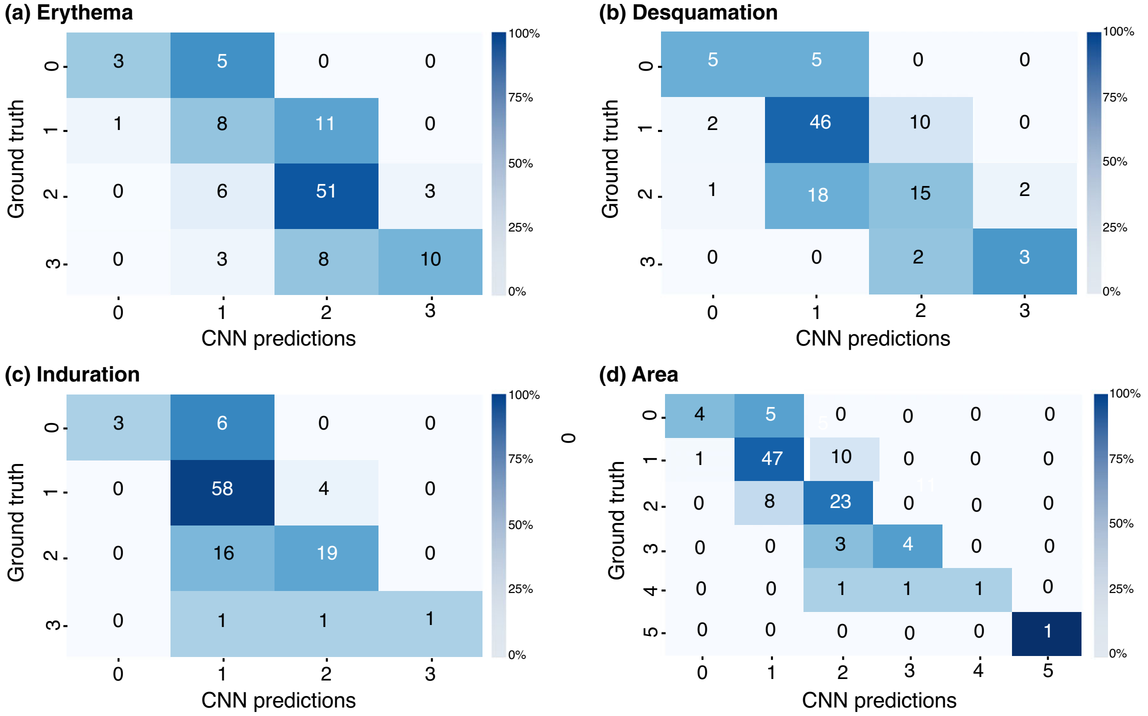
‡The mean absolute error reflects the absolute difference between the severity subscores assigned by the CNN and the ground truth averaged over the entire set. A score of 0 indicates perfect agreement.

§The ICC measures the inter-rater agreement between the severity subscores assigned by the CNN or PASI-trained physicians and the real-life scores. A score of 0 indicates no agreement and a score of 1 indicates full agreement.

Discussion

This study showed the potential of automated PASI scoring by CNNs based on images at an anatomical region level. For the $\text{CNN}_{\text{CORAL}}$, the best performing CNN architecture, the ICCs for erythema (range of ICCs: 0.614–0.711), desquamation (range of ICCs: 0.568–0.590) and induration (range of ICCs: 0.580–0.655) were in moderate agreement with real-life scoring, while the agreement between the area subscore and real-life scoring was good for all anatomical regions (range of ICCs: 0.793–0.832). The $\text{CNN}_{\text{CORAL}}$ reached higher accuracies than PASI-trained physicians on image-based severity scoring, achieving highest performance for area and induration scoring, followed by erythema and desquamation scoring. Differences in terms of ICC were less distinct, but showed the $\text{CNN}_{\text{CORAL}}$ outperformed (for area scoring) or performed as well as physicians (for erythema, induration and desquamation scoring). PASI-trained physicians were in good agreement with each other on image-based PASI scoring, but were in moderate agreement with respect to real-life PASI scores. To the best of our knowledge, this is the first attempt to perform PASI scoring by CNNs in anatomical regions, which is an important step towards objective and image-based automated PASI scoring by deep learning models.

Only two previous studies reported deep learning-based automated scoring of erythema, desquamation and induration on lesion level with accuracies ranging from 0.611 to 0.635.[18,19] Despite our challenging approach to determine psoriasis severity at an anatomical region level, accuracies reached by the $\text{CNN}_{\text{CORAL}}$ were predominantly higher on all subscores. We are the first to report area scoring in an anatomical region by means of CNN, reaching accuracies up to 79.4%. Our results emphasize the future potential of image-based automated and objective PASI scoring by CNNs. Currently, one system is able to perform automated PASI scoring, albeit without using a deep learning approach.[8] However, a direct comparison between the performance of that system and our results is not feasible, since the performance of that system was not reported on the level of single subscores, but rather on the level of combined subscores (which increases the chance of reaching higher ICCs).[9] We did not report performance on combined subscores or total PASI scores, since the single-task learning approach resulted in


(a) Erythema

(b) Desquamation

(c) Induration

(d) Area

Figure 2 Confusion matrices of $\text{CNN}_{\text{CORAL}}$ for (a) erythema, (b) desquamation, (c) induration and (d) area scoring of the trunk region. The confusion matrices for each subscore show the frequencies and percentage of correct classifications and misclassifications for each class in the test data set of the trunk region. The presented numbers are low, since the test data set is a 20% subset of the total database and included 109 concatenated imaging series of the trunk region. Most misclassifications are shown to be classified in the adjacent classes of the ground truth (real-life scores).

different images in each stratified random test set, making it impossible to combine subscores of the same anatomical region. Moreover, reporting performance on the level of single subscores gives the opportunity to evaluate the (CNN) performance in detail. Still, the ICCs reached on single subscore level by the $\text{CNN}_{\text{CORAL}}$ were comparable to superior to the ICCs of the combined subscores reported by Fink *et al.* Moreover, our deep learning method has the potential to increase performance as the data set increases. In contrast to the system reported by Fink *et al.*, image-based automated scoring by CNNs does not require a specialized capturing device, increasing the accessibility for clinical practice.

Physicians were in good agreement with each other on image-based PASI scoring, which is also reported by Fink *et al.*,[4] but the agreement with real-life scoring by the treating physician was averagely less than the CNN, indicating that it is challenging for physicians to perform image-based scoring. Previously reported ICCs between real-life and image-based scoring by physicians are higher than ours, reaching good agreement.[9] However, these ICCs were reported for combined subscores, so a direct comparison to our results is not possible. One could hypothesize that the expertise on real-life scoring of physicians complicates image-based scoring, since the psoriasis severity may look different on images. In contrast, the CNN was specifically trained to perform image-based PASI scoring in agreement with real-life PASI scores, which most likely explains why the $\text{CNN}_{\text{CORAL}}$ was in better agreement with the real-life scores than physicians.

We note several limitations, including potential errors in the real-life scores since they were collected as part of clinical practice by (PASI-trained) treating physicians. The CNN performance could be increased by improving the quality of the real-life scores by consensus on real-life assessment by several physicians. Given the retrospective nature of this study, we were unable to use this approach. To preserve anonymity, the head region was not included in our study. Furthermore, the sub score '4' was excluded from training for erythema, induration and desquamation, and the subscore '6' was excluded for the area score due to an insufficient number of images. Exclusion of the head region and most severe scores degrades the performance of the currently trained CNN in clinical practice. Even though our trained $\text{CNN}_{\text{CORAL}}$ outperforms previous deep learning attempts and used an anatomical region approach, performance is not yet adequate for implementation in clinical

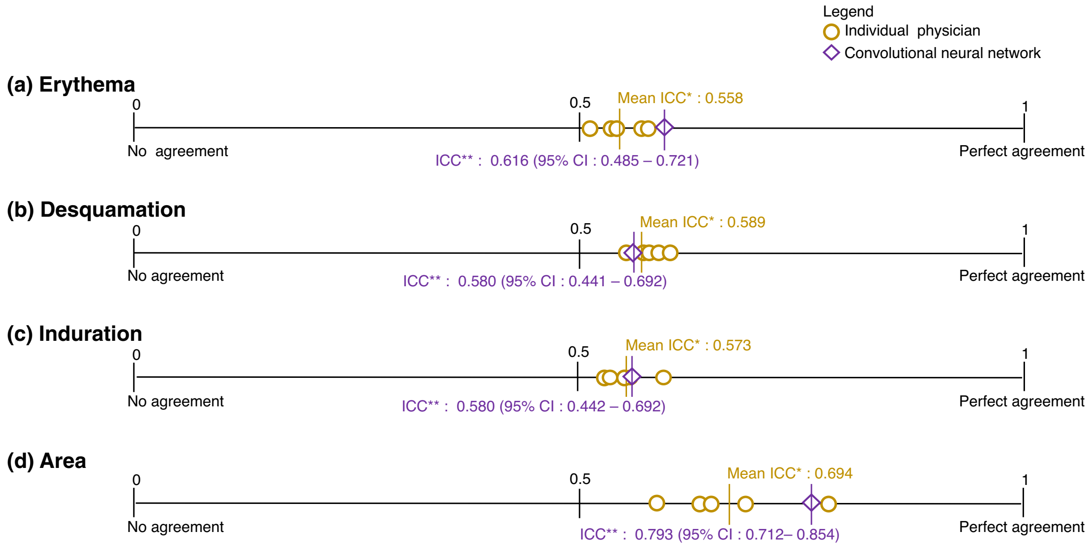


Figure 3 Comparison of agreement with real-life scores between image-based scoring of the trunk region by the $\text{CNN}_{\text{CORAL}}$ and physicians. The performance of the $\text{CNN}_{\text{CORAL}}$ and physicians on image-based scoring is depicted for each subscore (erythema, desquamation, induration, area) of the trunk region. The test set of the trunk region included 109 concatenated imaging series for each subscore. The performance was evaluated using the intraclass correlation coefficient (ICC) with 95% confidence intervals (CI), reflecting the agreement with real-life scores. The agreement was defined as < 0.50, poor agreement; 0.51–0.75, moderate; 0.75–0.90, good; 0.91–1, excellent. *Overall agreement of image-based scoring by physicians and real-life scores shown by the mean ICC of the five physicians. **ICC (95% CI) of the $\text{CNN}_{\text{CORAL}}$ and real-life scores.

practice. However, current results were reached with explorative and relatively small data sets with class imbalance. Therefore, expanding the training set could substantially increase the performance of the $\text{CNN}_{\text{CORAL}}$ for automated PASI scoring. With the increasing interest in tele-dermatology and at-home hospital care,[28] image-based automated PASI scoring by CNNs may not only provide objective PASI scoring in clinical practice and research, but may facilitate reliable psoriasis severity assessments for remote consultations in the future by incorporating the algorithm into a mobile application that could be used at home.

In conclusion, CNNs have the potential to automatically perform image-based PASI scoring at an anatomical region level. For the erythema, desquamation and induration scoring, CNNs performed similar to physicians, while for the area scoring CNNs outperformed physicians on image-based PASI scoring. In the future, automated PASI scoring could enable objective and efficient PASI scoring in (remote) clinical practice and clinical research.

Acknowledgements

We would like to thank Elke ter Haar, Lara van der Schoot and Marloes van Muijen for their contribution to the reader study. We thank Hans Groenewoud for his advice on the statistical analyses. Furthermore, we are grateful to Jeffrey Dokter, Lex Dingemans, Anouk van der Schot and Tom Loonen for their contribution during the preparation of the study.

References

1 Parisi R, Symmons DP, Griffiths CE *et al.* Global epidemiology of psoriasis: a systematic review of incidence and prevalence. *J Invest Dermatol* 2013; **133**: 377–385.

2 Fredriksson T, Pettersson U. Severe psoriasis–oral therapy with a new retinoid. *Dermatologica* 1978; **157**: 238–244.

3 Cabrera S, Chinniah N, Lock N, Cains GD, Woods J. Inter-observer reliability of the PASI in a clinical setting. *Australas J Dermatol* 2015; **56**: 100–102.

4 Fink C, Alt C, Uhlmann L, Klose C, Enk A, Haenssle HA. Intra- and interobserver variability of image-based PASI assessments in 120 patients suffering from plaque-type psoriasis. *J Eur Acad Dermatol Venereol* 2018; **32**: 1314–1319.

5 Ihtatho D, Fadzil MH, Affandi AM, Hussein SH. Area assessment of psoriasis lesion for PASI scoring. Conference Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Cité Internationale, Lyon, France, August 23–26, 2007; **2007**: 3446–3449.

6 Ahmad Fadzil MH, Ihtatho D, Mohd Affandi A, Hussein SH. Objective assessment of psoriasis erythema for PASI scoring. *J Med Eng Technol* 2009; **33**: 516–524.

7 Iyatomi H, Oka H, Hagiwara M *et al.* Computerized quantification of psoriasis lesions with colour calibration: preliminary results. *Clin Exp Dermatol* 2009; **34**: 830–833.

8 Fink C, Fuchs T, Enk A, Haenssle HA. Design of an algorithm for automated, computer-guided PASI measurements by digital image analysis. *J Med Syst* 2018; **42**: 248.

9 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.

10 Kislal EE, Halasz CL. Software for quantifying psoriasis and vitiligo from digital clinical photographs. *J Dermatolog Treat* 2013; **24**: 107–111.

11 Lu J, Kazmiercazk E, Manton JH, Sinclair R, eds. Automatic Scoring of Erythema and Scaling Severity in Psoriasis Diagnosis. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

12 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.

13 Lu J, Kazmierczak E, Manton JH, Sinclair R. A quantitative technique for assessing the change in severity over time in psoriatic lesions using computer aided image analysis. Conference Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Cité Internationale, Osaka, Japan, 3-7 July, 2013; **2007**: 2380–2383.

14 Lu J, Kazmierczak E, Manton JH, Sinclair R. Machine learning methods for segmenting psoriatic lesions from 2D images. *Front Med Imaging* 2014; 121–149.

15 Banu S, Toacse G, Danciu G. Objective erythema assessment of psoriasis lesions for Psoriasis Area and Severity Index (PASI) evaluation. International Conference and Exposition on Electrical and Power Engineering (EPE 2014). 16–18 October, lasi, Romania, 2014; **2014**: 52–56.

16 George YM, Aldeen M, Garnavi R. Automatic scale severity assessment method in psoriasis skin images using local descriptors. *IEEE J Biomed Health Inform* 2020; **24**: 577–585.

17 Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev* 2020; **53**: 5455–5516.

18 Pal A, Chaturvedi A, Garain U, Chandra A, Chatterjee R. Severity grading of psoriatic plaques using deep CNN based multi-task learning. 23rd International Conference on Pattern Recognition (ICPR) Cancun Center, Cancun, Mexico, December 4–8, 2016: 1478–1483.

19 Pal A, Chaturvedi A, Garain U, Chandra A, Chatterjee R, Senapati S, editors. Severity Assessment of Psoriatic Plaques Using Deep CNN Based Ordinal Classification. Springer International Publishing, Cham; 2018.

20 Cao W, Mirjalili V, Raschka S. Consistent rank logits for ordinal regression with convolutional neural networks. *Pattern Recognit Lett* 2019; **140**: 325–331. https://doi.org/10.1016/j.patrec.2020.11.008

21 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *IEEE Conf Comput Vision Pattern Recognit* 2016; **2016**: 770–778.

22 Russakovsky O, Deng J, Su H *et al.* ImageNet large scale visual recognition challenge. *Int J Comput vis* 2015; **115**: 211–252.

23 Kingma DP, Adam BJ. A method for stochastic optimization. *CoRR* 2015. https://arxiv.org/abs/1412.6980

24 Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* 2019.

25 Meakin J, Zeeland H, Koek M, Gerke PK, de Dobbelaer B, Pinckaers H *et al*. Grand Challenge.org. Version v2020.12 ed. Zenodo 2020.

26 Niu Z, Zhou M, Wang L, Gao X, Hua G. Ordinal regression with multiple output CNN for age estimation. *IEEE Conf Comput Vision Pattern Recognit* 2016; **2016**: 4920–4928.

27 Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966; **19**: 3–11.

28 Beer J, Hadeler E, Calume A, Gitlow H, Nouri K. Teledermatology: current indications and considerations for future use. *Arch Dermatol Res* 2021; **313**: 11–15.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Standardized imaging series representing each anatomical region before image pre-processing.

Figure S2. Confusion matrices of $\text{CNN}_{\text{CORAL}}$ for (a) erythema, (b) desquamation, (c) induration and (d) area scoring of the arms region.

Figure S3. Confusion matrices of $\text{CNN}_{\text{CORAL}}$ for (a) erythema, (b) desquamation, (c) induration and (d) area scoring of the legs region.

Figure S4. Input (a) and output (b) of the pre-processing algorithm for the arms region.

Table S1. Distribution of the PASI subscores in the image series of three anatomical regions.

Table S2. Performance of Convolutional Neural Network (CNN) architectures for each subscore by anatomical region.

Table S3. Image-based average predictions of physicians and the $\text{CNN}_{\text{CORAL}}$ compared to the ground truth.

Appendix S1. Methods.