Detecting Eczema Areas in Digital Images: An Impossible Task?

Guillem Hurault[1], Kevin Pan[1], Ricardo Mokhtari[1], Bayanne Olabi[2], Eleanor Earp[3], Lloyd Steele[4], Hywel C. Williams[2,5] and Reiko J. Tanaka[1]

Assessing the severity of atopic dermatitis (AD, or eczema) traditionally relies on a face-to-face assessment by healthcare professionals and may suffer from inter- and intra-rater variability. With the expanding role of telemedicine, several machine learning algorithms have been proposed to automatically assess AD severity from digital images. Those algorithms usually detect and then delineate (segment) AD lesions before assessing lesional severity and are trained using the data of AD areas detected by healthcare professionals. To evaluate the reliability of such data, we estimated the inter-rater reliability of AD segmentation in digital images. Four dermatologists independently segmented AD lesions in 80 digital images collected in a published clinical trial. We estimated the inter-rater reliability of the AD segmentation using the intraclass correlation coefficient at the pixel and the area levels for different resolutions of the images. The average intraclass correlation coefficient was 0.45 (standard error = 0.04) corresponding to a poor agreement between raters, whereas the degree of agreement for AD segmentation varied from image to image. The AD segmentation in digital images is highly rater dependent even among dermatologists. Such limitations need to be taken into consideration when AD segmentation data are used to train machine learning algorithms that assess eczema severity.

JID Innovations (2022);**2:**100133 doi:10.1016/j.xjidi.2022.100133

INTRODUCTION

Atopic dermatitis (AD) (also called eczema) is one of the most common chronic skin diseases (Langan et al., 2020). Many clinical trials on AD treatment include the assessment of AD severity that changes dynamically over time. The assessment of AD severity usually consists of the visual inspection of eczema lesions by healthcare professionals who grade the intensity of several disease signs (such as dryness, redness, and excoriations) and estimate the area (extent) covered by eczema.

The recent development of machine learning (ML) methods, together with the need for telemedicine, resulted in an increasing interest in developing computer vision algorithms for automatic evaluation of AD severity from digital images (Alam et al., 2016; Bang et al., 2021; Junayed et al., 2020; Pan et al., 2020). Those algorithms generally consist of two steps: (i) identifying areas covered by eczema in each image—so that the images are segmented—either manually as part of the data preprocessing (human in the loop) or automatically by an algorithm and then (ii) predicting the severity of eczema features in the segmented areas. Therefore, reliable detection of eczema lesions is a prerequisite for assessing the severity of these lesions.

The lack of high-quality segmentation labels is one of the main obstacles to developing ML methods in medical applications (Ching et al., 2018; Karimi et al., 2020). If the eczema segmentation data provided by dermatologists are of low quality (noisy labels), the algorithms for automatic detection of AD lesions trained with such data may learn the biases contained in the data. For example, models trained with noisy labels to segment brain lesions required an order of magnitude more data than those trained with accurate labels to achieve similar segmentation performance (Karimi et al., 2020). Inaccurate eczema segmentation can also have effects on assessing severity in eczema images because the segmentation may only include severe lesions or specific disease signs. For example, if the areas of dryness are never segmented, the assessments of dryness from the segmented eczema images are likely to be inaccurate. The classification accuracy of cancerous prostate tissue images by ML algorithms was found to be decreased by 10% when trained with data that contained incorrectly labeled images (Karimi et al., 2020).

However, to the best of our knowledge, it is still unclear whether high-quality eczema segmentation data can be obtained from dermatologists consistently. Trying to measure the eczema area accurately is challenging in real life due to the ill-defined nature of AD. Charman et al. (1999) showed a very poor inter-rater reliability (IRR) with a kappa statistic of 0.09 for in-person scoring of the extent of eczema by six experts on six patients. IRR refers to the degree of agreement between raters, that is, to what extent the labels (the extent of eczema in the case of

[1]*Department of Bioengineering, Imperial College London, London, United Kingdom;* [2]*Biosciences Institute, Faculty of Medical Sciences, Newcastle University, United Kingdom;* [3]*Department of Dermatology, Lauriston Building, Edinburgh, United Kingdom;* [4]*Department of Dermatology, The Royal London Hospital, Barts Health NHS Trust, London, United Kingdom; and* [5]*Centre of Evidence-Based Dermatology, University of Nottingham, Nottingham, United Kingdom*

Correspondence: Reiko J. Tanaka, Department of Bioengineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom. E-mail: r.tanaka@imperial.ac.uk

Abbreviations: AD, atopic dermatitis; ICC, intraclass correlation coefficient; IRR, inter-rater reliability; KA, Krippendorff's alpha; ML, machine learning

Received 19 October 2021; revised 28 April 2022; accepted 2 May 2022; accepted manuscript published online XXX; corrected proof published online XXX

Cite this article as: JID Innovations 2022;2:100133

© 2022 The Authors. Published by Elsevier, Inc. on behalf of the Society for Investigative Dermatology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Charman et al. [1999]) are independent of a particular rater. If IRR is high, "raters can be used interchangeably without the researcher having to worry about the categorization being affected by a significant rater factor" (Gwet, 2010).

This study quantified the IRR of eczema area segmentation by four dermatologists in 80 digital images of varying quality from pediatric patients with AD collected in a published clinical study (Thomas et al., 2011).

RESULTS

Quality of images for eczema area segmentation

The 80 images we dealt with were from different representative AD sites: 31 images were for legs, 19 were for hands, 14 were for arms, 12 were for feet, and 4 were for the head and neck area (Table 1). The 80 images were selected by random sampling and did not include any images of the trunk/back because those sites were not included abundantly enough in the original study. The probability of selecting 80 images from the original 1,345 images without including images of trunk/back (48 images) is approximately 5%, as described by hypergeometric distribution. Each image was annotated with the intensity scores (0–3) of the six signs of Six Area, Six Sign Atopic Dermatitis severity score (erythema, exudation, excoriation, dryness, cracking, and lichenification) at the corresponding representative AD site, and the average sum of the signs was 6.3 (where the maximum is $18 = 3 \times 6$ signs, SD = 3.2).

The four raters fully agreed on the image quality for 23 of the 80 images. Four images were assessed to have a contradicting image quality of high and low by at least two raters, and the remaining 53 images had a combination of normal/high or normal/low. Visual inspection of the four images with contradictory image quality, together with the reasons for the low quality provided, revealed that the textures and features of the eczema area were not well-preserved or easily identifiable in those images, without obvious technical issues. A total of 43 of the 80 images were deemed of poor quality by at least one rater. Among those 43 images, 35 were deemed out of focus (Figure 1a), 15 were deemed overexposed, and 14 had "other reasons" (details not given).

We computed the image quality score for each image by averaging the quality assessed by the four raters, with "low," "normal," and "high" being coded as −1, 0, and 1, respectively. Over the 80 images, the image quality score had a mean of −0.081 (SD = 0.45), which is close to normal image quality (0). We investigated whether the image quality score could be confounded by the body regions or the severity score (the sum of the intensity scores for the six disease signs) in a linear model, but no coefficients appeared as significant (Figure 1b).

Table 1. Characteristics of the Original Dataset and the Selected Images

Characteristics	SWET Dataset	Selected Images
Number of patients with images (% female)	287 (43%)	71 (44%)
Patients of (declared) white ethnicity, n (%)	223 (78%)	54 (76%)
Mean age in years (SD)	5.6 (4.1)	5.1 (4.0)
Number of images	1,345	80
Images of legs, n (%)	534 (40%)	31 (39%)
Images of hands, n (%)	372 (28%)	19 (24%)
Images of arms, n (%)	190 (14%)	14 (17%)
Images of feet, n (%)	148 (11%)	12 (15%)
Images of the head and neck area, n (%)	53 (4%)	4 (5%)
Images of the trunk or back, n (%)	48 (3%)	0 (0%)
Mean regional SASSAD (SD) (maximum = 18)	6.3 (3.2)	6.3 (3.2)
Mean TISS (SD) (maximum = 9)	3.0 (1.8)	3.2 (1.8)

Abbreviations: SASSAD, Six Area, Six Sign Atopic Dermatitis; SWET, Softened Water Eczema Trial; TISS, Three Item Severity Score.

IRR of eczema segmentation

We assessed the IRR of eczema segmentation for each image using intraclass correlation coefficient (ICC), calculated at the pixel level. The pixel-level ICC showed a large image-to-image variation: 56% of the images (40 of 71) had an ICC < 0.5 (considered as a poor agreement), and 11% (8 of 71) had an ICC > 0.9 (considered as an excellent agreement), leaving the remaining 33% with a moderate level of pixel-level agreement (example segmentation masks in Figure 2). The average pixel-level ICC was 0.45 (standard error = 0.04) (Figure 3), corresponding to a poor average agreement between raters. Similarly, the pixel-level Krippendorff's alphas (KAs) (another IRR metric) were strongly correlated with the pixel-level ICC (Figure 4) (Pearson correlation of 0.879, 95% confidence interval = 0.812–0.923).

We also assessed the ICC at the area level to investigate whether there was a consensus in identifying larger regions of eczema beyond the pixel-level precise segmentation. The area-level ICC for different resolutions of the images were strongly correlated with each other and with pixel-level ICC (Figure 4). The average area-level ICC was not significantly different from the average pixel-level ICC and corresponded to a poor average agreement between raters (Figure 3). Beyond eczema segmentation, we calculated the extent ICC, that is, the ICC for the proportion of eczema in the images. The extent ICC was 0.440 (95% confidence interval = 0.313–0.555), confirming that the raters could not agree on the proportion of eczema in the images.

We investigated whether the IRR was confounded by body regions, severity scores, average labeling time by the four raters, and image quality scores using a linear model with IRR metrics as the dependent variables (Figure 5a). The only significant effect was found for the image quality scores, with a higher quality associated with a lower ICC. This counter-intuitive result may be explained by the fact that the raters did not attempt a precise segmentation on lower-quality images resulting in a higher agreement.

We conducted a sensitivity analysis to assess whether the IRR estimates were driven by the segmentation of particular raters, but we did not find significant differences in pixel-level ICC estimates when one of the raters was removed in turn from the analysis (Figure 5b). Our results are not driven by the segmentation labels of a particular rater, including the most experienced rater (rater 1).

Performance for eczema area segmentation

To evaluate how much of the segmentation errors can be attributed to the IRR, we calculated an average rater's

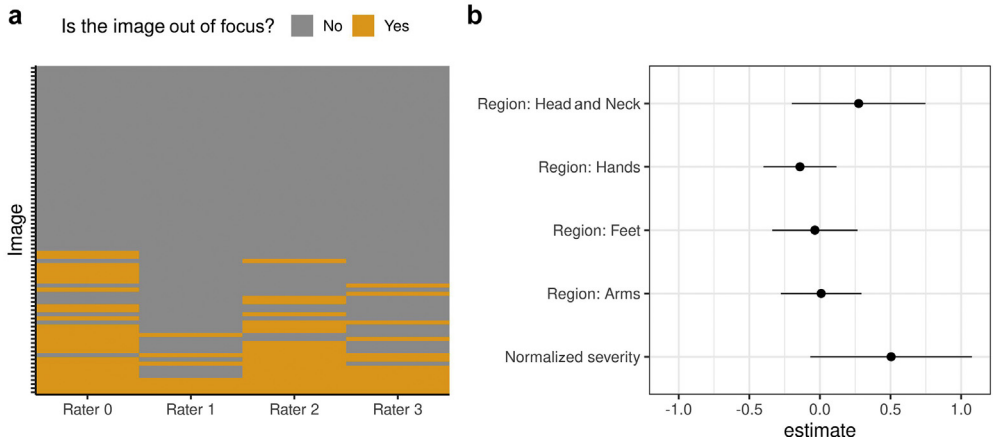

Figure 1. Segmentation quality assessment. (**a**) Distribution of out-of-focus assessments (orange) by the four raters (x-axis) for each image (y-axis). A total of 17 images were deemed out of focus by only one rater, 8 images were deemed out of focus by two raters, 6 images were deemed out of focus by three raters, and 4 images were deemed out of focus by all the four raters. (**b**) Estimated coefficients (and 95% confidence interval) for variables in a linear model that predicts the mean image quality score across raters. The coefficients for the regions quantify the difference in the intercept from that of the default region (legs).

segmentation performance, quantifying the difference between the segmentation of an average rater and the consensus segmentation of their peers. We compared this average rater's segmentation performance with a naïve segmentation performance achieved by a naïve rater who segments all skin regions as eczema (Figure 6). The performance of the average rater's segmentation was always better than that of a naïve segmentation, except for the true positive rate metric, which does not penalize false positives that the naïve segmentation produces by many. None of the metrics for the average rater's performance was close to a perfect score of 1, which would have been expected if the IRR were excellent and if any raters' segmentation could be used interchangeably. Our estimation suggested that an average rater can segment eczema with an accuracy of 80.6 ± 1.5 % (averaged across images) compared with segmentation by the consensus of their peers.

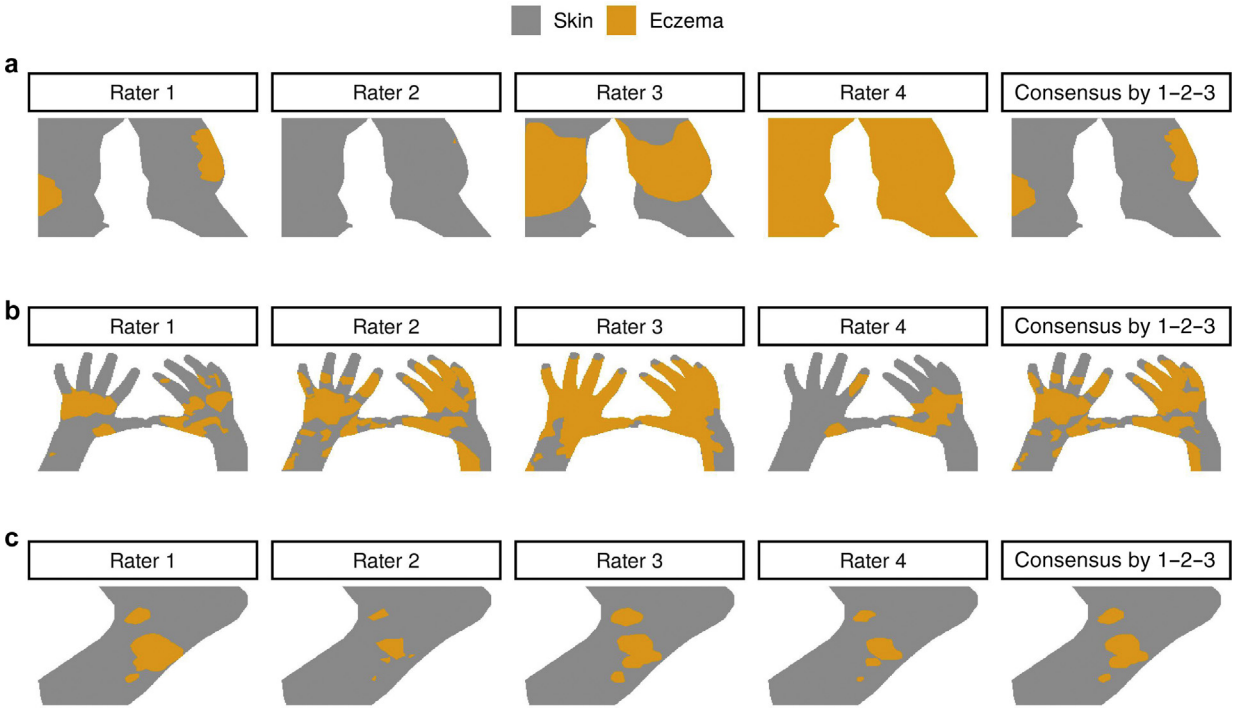

Figure 2. Example eczema (orange) and skin (gray) masks segmented by four raters for three images. The images correspond to (**a**) leg, (**b**) hands, and (**c**) foot. The last column illustrates a consensus segmentation by raters 1, 2, and 3. The consensus segmentation was used as a ground truth when evaluating the segmentation performance of rater 4. As an example, the three images represent (**a**) a very poor IRR ($ICC = 0.026$, $KA = -0.19$), (**b**) an average (i.e., poor in this dataset) IRR ($ICC = 0.41$, $KA = 0.19$), and (**c**) an excellent IRR ($ICC = 0.976$, $KA = 0.63$). ICC, intraclass correlation coefficient; IRR, inter-rater reliability; KA, Krippendorff's alpha.

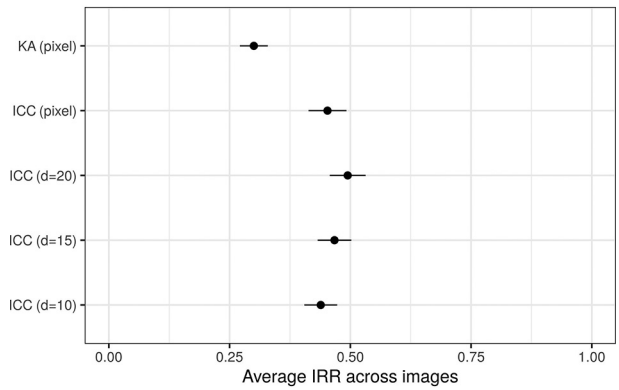

Figure 3. Average (± *SE*) pixel- and area-level IRR estimates (the higher the better). *KA* (*pixel*) and *ICC* (*pixel*) correspond to pixel-level IRR metrics, and other *ICCs* correspond to area-level IRR metrics, where *d* describes the image resolution. A larger *d* corresponds to a smaller area of interest. ICC, intraclass correlation coefficient; IRR, inter-rater reliability; KA, Krippendorff's alpha; SE, standard error.

The difference between the naïve and the average raters' performance varies across metrics, with a greater difference in accuracy, positive predicted value (also known as precision), and true positive rate than for the F1 score and intersection over union (Figure 6). These results suggest that not all metrics may be appropriate to monitor improvement in the segmentation performance of an ML algorithm and discourage the use of F1 score and intersection over union because they show a smaller difference between the naïve and the average raters' performance. A narrow difference makes it difficult to detect the improvement in the segmentation performance of an ML algorithm that is likely to be above the naïve performance and below the average rater's performance.

DISCUSSION

In this study, we investigated the IRR of eczema segmentation from digital images. Four dermatologists (raters) segmented eczema lesions in 80 images collected in a previously published clinical trial (Figure 2). The IRR of eczema segmentation varied from image to image, with a poor agreement between the raters on average (Figure 3). We also estimated the segmentation performance for an average rater and a naïve rater who segments all skin regions as eczema (Figure 6). Those segmentation performances could be used to benchmark eczema segmentation algorithms and to choose the most appropriate metric to monitor the performance of eczema segmentation algorithms.

Our results highlight the difficulty of detecting eczema lesions from digital images consistently, raising questions on the validity of ML models to automatically assess eczema severity if they rely on eczema area segmentation without accounting for raters' potential biases. The results are perhaps not surprising given that AD is "ill-defined erythema with surface change" by definition (Williams et al., 1995) and that clinical assessment of affected areas is challenging (Charman et al., 1999).

The problem of the poor IRR in the segmentation data for training ML models could be addressed in several ways. For example, we can design end-to-end ML models for automatic assessment of eczema severity without relying on eczema segmentation masks provided by dermatologists and instead work with original images or images with background

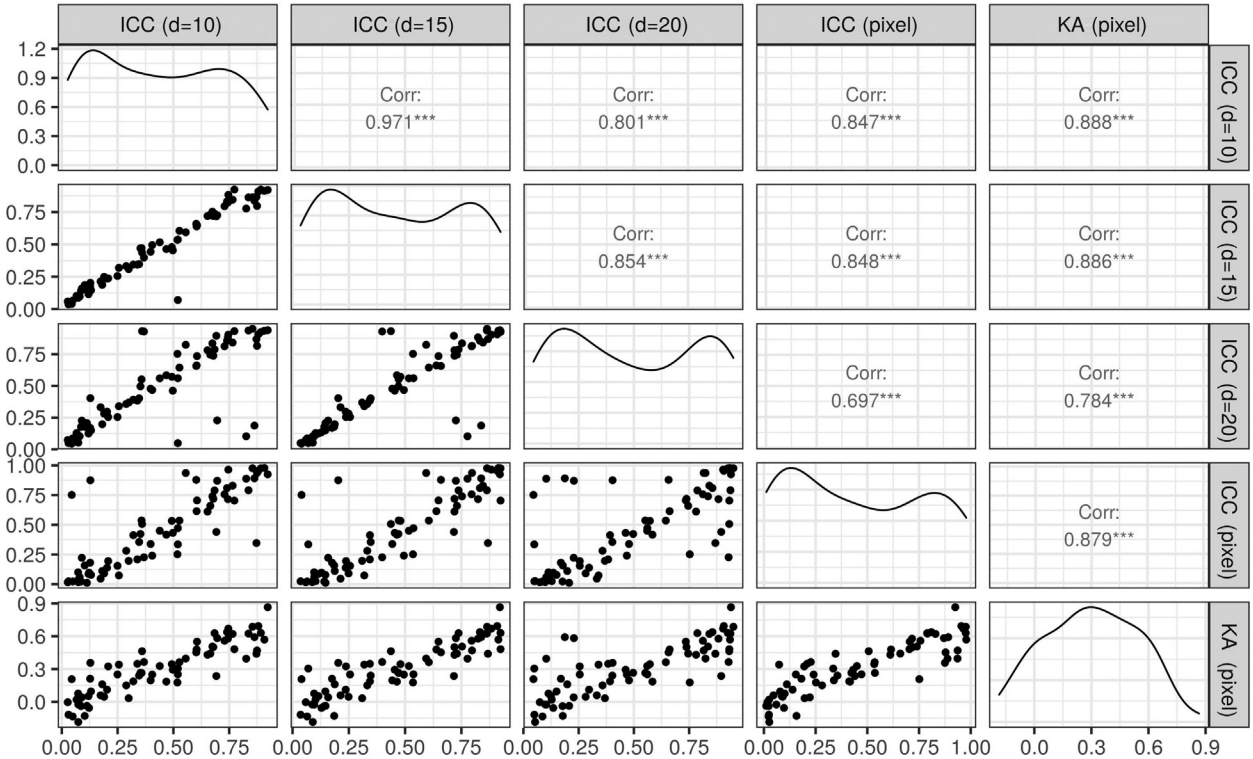

Figure 4. Comparison between the IRR metrics considered in this study (scatter plots, density plots, and Pearson correlations). Corr, correlation coefficient; IRR, inter-rater reliability; KA, Krippendorff's alpha.

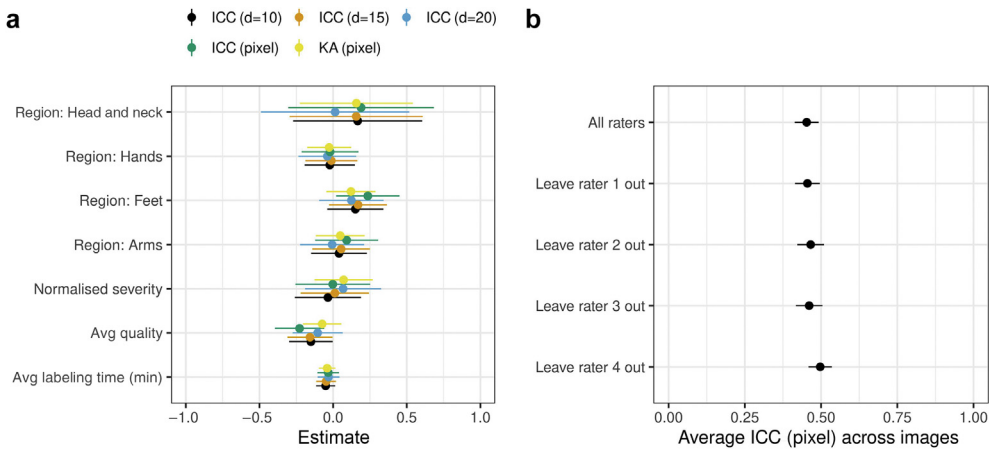

Figure 5. IRR confounding and sensitivity. (**a**) Estimated coefficients (and 95% confidence interval) for the variables in a linear model that predicts IRR metrics. The variables were normalized to a sensible scale for a fair interpretation of the effect sizes. The coefficients for the regions quantify the difference in intercept from that of the default region (legs). (**b**) Leave-one-rater out sensitivity analysis of the average pixel-level ICC measure. Avg, average; ICC, intraclass correlation coefficient; IRR, inter-rater reliability; KA, Krippendorff's alpha; min, minute.

removed (using skin segmentation masks) and let the algorithm identify eczema regions by itself. Other possibilities include using eczema segmentation algorithms that can be trained on noisy segmentation labels (Karimi et al., 2020) or trying to improve eczema segmentation labels. Improving the quality of eczema segmentation could be achieved by better training of raters, such as providing feedback on a reference set of training images until a certain level of ICC is achieved. Averaging the segmentation from multiple independent raters may also help. For example, the ICC of an average (consensus) segmentation by nine independent raters will be 0.9 (excellent) if we assume that the ICC of an individual segmentation is 0.5 (poor) (Nakagawa and Schielzeth, 2010). It may also be possible to ask fewer raters or even nonexperts to segment eczema and take advantage of the raters' systematic biases using crowdsourcing models such as a Dawid–Skene model.

There were several limitations in this study. First, our randomly selected 80 images from a published clinical trial mostly contained images of white skin tones. Further research is needed to investigate the quality of eczema segmentation for darker skin tones. For example, erythema is more likely to appear violaceous or dark brown in darker skin tones, making the delineating of the inflamed border potentially more challenging even if the lesion is not completely missed (Kaufman et al., 2018). Collecting images of eczema on darker skin tones is also relevant for developing ML algorithms that tend to be more accurate on the skin types they were trained on (Groh et al., 2021). Second, the IRR metrics used in this study did not consider the spatial structure of the image, whereas neighboring pixel labels are unlikely to be independent. This could be addressed with additional preprocessing such as computing local extent values using a kernel smoother (e.g., Gaussian blurring), which would also avoid compressing the masks when computing the area-level IRR. Our results were nonetheless consistent between pixel-level, area-level, and extent IRRs, highlighting the robustness of our conclusions. Finally, it would be valuable to estimate the intra-rater reliability of AD segmentation in digital images, that is, to what extent raters identify the same lesions on different occasions consistently. Poor intra-rater reliability may also be detrimental to the development of ML models relying on AD segmentation data.

In conclusion, this study showed that the AD segmentation in digital images is highly rater dependent even among dermatologists. Such limitations need to be taken into consideration when the AD segmentation data are used to train ML algorithms that assess eczema severity.

MATERIALS AND METHODS

Data

Our data originate from the Softened Water Eczema Trial (Thomas et al., 2011), a randomized controlled trial that investigated the effects of the use of ion-exchange water softeners on the control of AD symptoms. The trial included 310 children aged from 6 months to 16 years with moderate-to-severe AD, from whom digital images of representative AD sites were taken, and the Six Area, Six Sign Atopic Dermatitis severity score (Berth-Jones, 1996) was assessed at multiple clinical visits. The images can be considered realistic because they were taken using different devices, contain significant areas of background, and vary in resolution and subjective quality, such as focus, lighting, and blur. Each image was annotated with the intensity scores (0–3) of the six signs for Six Area, Six Sign Atopic Dermatitis severity score (erythema, exudation, excoriation, dryness, cracking, and lichenification) at the corresponding representative AD site. We represented the severity score for each image by the sum of the six intensity scores.

From a total of 1,345 eczema images available, we used a random number generator computer program to select 80 images at random without replacement (Table 1). We asked four dermatologists (raters) to segment AD lesions within each image (fully crossed design). One rater (HCW, rater 1) is a dermatologist with over 30 years of experience in assessing eczema. The other three raters (BO, EE, and LS) are trainee dermatologists nearing the end of their training who received feedback from HCW on eczema area segmentation beforehand to minimize variation. The raters also assessed the image quality ("low," "normal," or "high") for segmenting eczema areas and optionally reported quality issues ("out of focus," "overexposed," or "other reasons"). We believe that this group of raters with different levels of clinical experience is realistic and represents a normal practice for this type of time-consuming task. The segmentation of skin versus background was performed by a non-clinician (RM) because this task did not require clinical expertise.

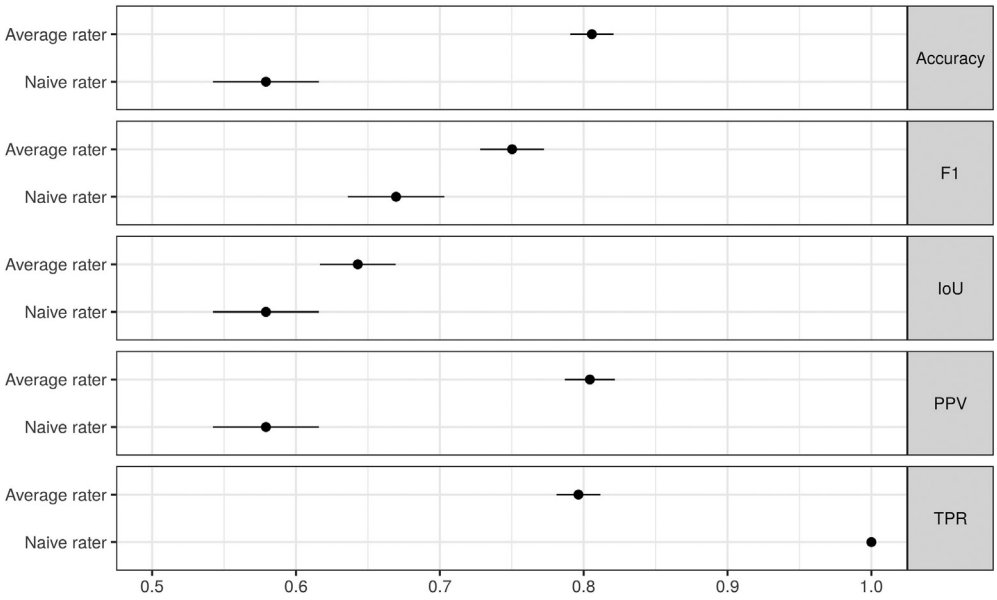

Figure 6. The segmentation performance of the average rater and the naïve rater (mean ± *SE* across images, the higher the better). IoU, intersection over union; PPV, positive predicted value (precision); SE, standard error; TPR, true positive rate.

Labelbox (https://labelbox.com) was used as the image segmentation tool because of its ease of use and accessibility. Examples of segmented image masks are shown in Figure 2.

Metrics for IRR

To quantify the IRR, we computed the ICC, which quantifies the association between categorical or continuous ratings assigned to the same rating unit (e.g., the pixel for the pixel-level eczema segmentation). The ICC is defined as the proportion of variance that can be attributed to between-unit variance (e.g., between-pixel variance) to the total variance (e.g., the sum of between-pixel variation and between-rater variance) (Hallgren, 2012; Nakagawa and Schielzeth, 2010). It takes values between 0 and 1, with higher ICC values corresponding to better consensus between raters. Although the interpretation of ICC values is still debated (Koo and Li, 2016), a consensus is that ICC $<$ 0.5 and ICC $>$ 0.9 indicate a poor and an excellent agreement, respectively.

We evaluated the ICC at the pixel and area levels for each image. The area-level ICC was calculated to investigate whether our IRR estimates were sensitive to the resolution of the images and whether there was a consensus in identifying larger regions of eczema (or the local extent of eczema) beyond the pixel-level precise segmentation. To estimate the area-level ICC, we compressed the original images to d $\times$ d cells for d$\in$(10, 15, 20) and counted the number of eczema pixels in each cell (Figure 7). We also calculated the extent ICC, which is the ICC for the total extent of eczema (proportion of eczema against skin pixels) that can be seen as the limit of the area-level ICC when d$\rightarrow$1.

We also reported KA at the pixel level. KA is a generalization of Kappa coefficients (e.g., Cohen's Kappa, Fleiss' Kappa) and is

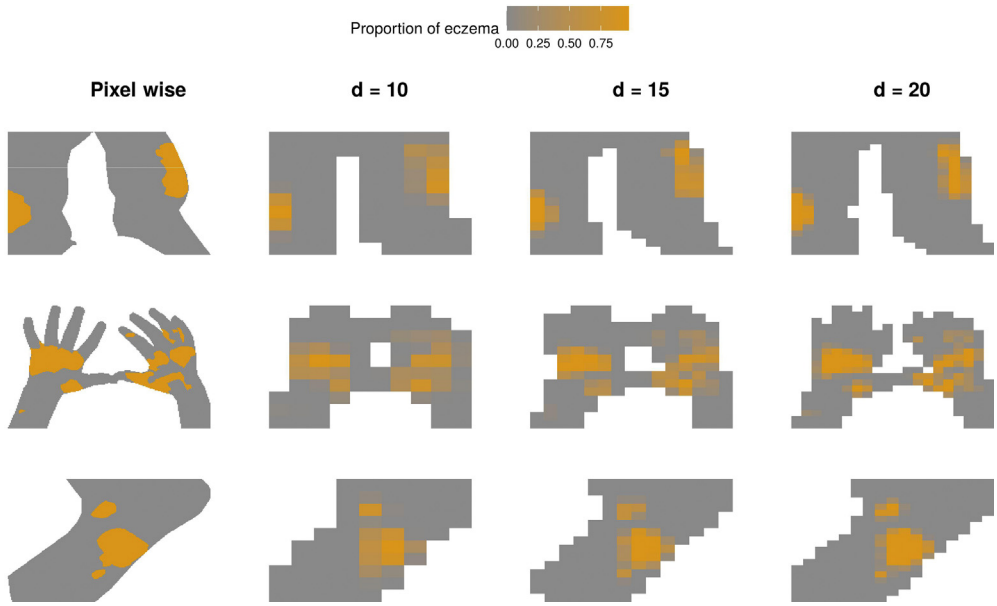

Figure 7. Illustration of area-level segmentation compared with pixel-level segmentation. The images shown are the same as in Figure 2 for segmentation by rater 1. Rows correspond to images, and columns correspond to pixel- and area-level segmentation for different image resolutions (*d*).

suitable for categorical ratings with more than two raters and with a fully crossed design, as in this study (Hallgren, 2012). The maximum KA is 1, corresponding to the perfect consensus between raters, and KA = 0 corresponds to a chance-level agreement. We did not consider the intersection over union (also known as Jaccard index), a similarity metric often used with images, because it does not control for chance agreement between raters and is limited to two raters.

Estimating the IRR metrics

In estimating the IRR metrics, we considered only skin pixels so that the chance agreement does not include background pixels. We analyzed 71 images because we excluded nine images in which the extent of eczema involvement was >95% for three or four raters because it implies a high-chance agreement for which an agreement measure cannot be estimated reliably. No images had an extent of eczema <5% for three or four raters.

We investigated whether the IRR estimates were driven by the segmentation of one rater, notably to see whether excluding the most experienced rater (rater 1) influenced the IRR, by conducting a leave-one-rater-out sensitivity analysis.

The ICC was estimated using the rptR package in R (Nakagawa and Schielzeth, 2010). For both the pixel- and the area-level ICCs, we used a mixed-effects logistic regression (binary outcomes of skin/AD) with random effects on raters and pixels/areas (two-way random-effects ICC), and the ICC was calculated on the latent (logit) scale. For the area-level ICC, the model used proportion data as input. For the extent ICC, we computed a two-way random-effects ICC (with an image as the grouping unit) using a mixed-effects linear regression model on the logit of the extent, and confidence intervals were estimated using bootstrap with 1,000 resampling. KA was computed using the irr package in R.

Metrics for segmentation performance

To evaluate how much of the segmentation errors can be attributed to the IRR, we compared the pixel-level eczema segmentation provided by a specific rater with a consensus segmentation and calculated the segmentation performance of the rater compared with those of their peers. The consensus segmentation was obtained from the other three raters by majority voting, that is, each pixel was labeled as eczema if at least two of the three raters labeled the pixel as eczema (Figure 2, last column).

The segmentation performance was measured using standard metrics for computer vision classification, such as intersection over union, accuracy, true positive rate (also known as sensitivity or recall), positive predictive value (also known as precision), and F1 score, where the consensus segmentation was treated as true labels. We only considered skin pixels and did not exclude any images when computing the segmentation performance.

We calculated the segmentation performance for each of the four raters in turn and averaged the performance over the four raters to derive the segmentation performance of an average rater, for each image. The average rater's segmentation performance was compared with a naïve segmentation performance for a naïve rater who would predict all skin regions to be eczema. The naïve segmentation performance corresponds to the lower bound of the segmentation performance that we would expect any segmentation algorithm to achieve.

Data availability statement

All the codes are available at https://github.com/ghurault/IRR-eczema-images.

ORCIDs

Guillem Hurault: http://orcid.org/0000-0002-1052-3564
Kevin Pan: http://orcid.org/0000-0002-2834-605X
Ricardo Mokhtari: http://orcid.org/0000-0002-7940-6489
Bayanne Olabi: http://orcid.org/0000-0002-4786-7838
Eleanor Earp: http://orcid.org/0000-0002-8316-0223
Lloyd Steele: http://orcid.org/0000-0003-4745-1338
Hywel C. Williams: http://orcid.org/0000-0002-5646-3093
Reiko J. Tanaka: http://orcid.org/0000-0002-0769-9382

AUTHOR CONTRIBUTIONS

Conceptualization: GH, HCW, RJT; Data Curation: KP; Formal Analysis: GH, KP; Funding Acquisition: RJT; Investigation: GH, KP; Methodology: GH, KP; Resources: RM, BO, EE, LS, HCW; Software: GH; Supervision: RJT, HCW; Validation: GH, RJT; Visualization: GH, KP; Writing — Original Draft Preparation: GH, KP; Writing — Review and Editing: RJT, HCW

ACKNOWLEDGMENTS

We thank Kim S. Thomas and the Softened Water Eczema Trial team for sharing the dataset. The Softened Water Eczema Trial trial was funded by the National Institute for Health and Care Research Health Technology Assessment Programme. This study was funded by the British Skin Foundation.

CONFLICT OF INTEREST

The authors state no conflict of interest.

REFERENCES

Alam MN, Munia TTK, Tavakolian K, Vasefi F, Mackinnon N, Fazel-Rezai R. Automatic detection and severity measurement of eczema using image processing. Annu Int Conf IEEE Eng Med Biol Soc 2016;2016:1365–8.

Bang CH, Yoon JW, Ryu JY, Chun JH, Han JH, Lee YB, et al. Automated severity scoring of atopic dermatitis patients by a deep neural network [published correction appears in Sci Rep 2021;11:15640] Sci Rep 2021;11:6049.

Berth-Jones J. Six Area, six Sign Atopic Dermatitis (SASSAD) seventy score: a simple system for monitoring disease activity in atopic dermatitis. Br J Dermatol Suppl 1996;135:25–30.

Charman CR, Venn AJ, Williams HC. Measurement of body surface area involvement in atopic eczema: an impossible task? Br J Dermatol 1999;140:109–11.

Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15:20170387.

Groh M, Harris C, Soenksen L, Lau F, Han R, Kim A, et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2021. p. 1820–8.

Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Gaithersburg, MD: STATAXIS Publishing Company. Advanced Analytics, LLC; 2010.

Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012;8:23–34.

Junayed MS, Sakib ANM, Anjum N, Islam MB, Jeny AA. EczemaNet: A deep CNN-based eczema diseases classification. Paper presented at: Fourth IEEE International Conference on Image Processing, Applications and Systems (IPAS 2020). 2020; Genova, Italy.

Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal 2020;65:101759.

Kaufman BP, Guttman-Yassky E, Alexis AF. Atopic dermatitis in diverse racial and ethnic groups-variations in epidemiology, genetics, clinical presentation and treatment. Exp Dermatol 2018;27:340–57.

Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research [published correction appears in J Chiropr Med 2017;16:346] J Chiropr Med 2016;15:155–63.

Langan SM, Irvine AD, Weidinger S. Atopic dermatitis [published correction appears in Lancet 2020;396:758] Lancet 2020;396:345–60.

Nakagawa S, Schielzeth H. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev Camb Philos Soc 2010;85:935–56.

Pan K, Hurault G, Arulkumaran K, Williams HC, Tanaka RJ. EczemaNet: automating detection and severity assessment of atopic dermatitis. In: Machine Learning in Medical Imaging. MLMI 2020;12436:220–230.

Thomas KS, Dean T, O'Leary C, Sach TH, Koller K, Frost A, et al. A randomised controlled trial of ion-exchange water softeners for the treatment of eczema in children. PLoS Med 2011;8:e1000395.

Williams HC, Forsdyke H, Boodoo G, Hay RJ, Burney PGJ. A protocol for recording the sign of flexural dermatitis in children. Br J Dermatol 1995;133:941–9.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/