HHS Public Access

Author manuscript

Skin Res Technol. Author manuscript; available in PMC 2020 July 01.

Published in final edited form as:

Skin Res Technol. 2019 July ; 25(4): 572–577. doi:10.1111/srt.12688.

Crowdsourcing to delineate skin affected by chronic graft-vs-host disease

Eric R. Tkaczyk[1,2,3], **Joseph R. Coco**[4], **Jianing Wang**[5], **Fuyao Chen**[1,2,3], **Cheng Ye**[5], **Madan H. Jagasia**[6], **Benoit M. Dawant**[3,5], **Daniel Fabbri**[4,5]

[1]Department of Veterans Affairs, Tennessee Valley Health System, Nashville, Tennessee

[2]Department of Dermatology, Vanderbilt Cutaneous Imaging Clinic, Vanderbilt University Medical Center, Nashville, Tennessee

[3]Department of Biomedical Engineering, Vanderbilt University, Nashville, Tennessee

[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee

[5]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee

[6]Vanderbilt-Ingram Cancer Center, Nashville, Tennessee

Abstract

Background: Estimating the extent of affected skin is an important unmet clinical need both for research and practical management in many diseases. In particular, cutaneous burden of chronic graft-vs-host disease (cGVHD) is a primary outcome in many trials. Despite advances in artificial intelligence and 3D photography, progress toward reliable automated techniques is hindered by limited expert time to delineate cGVHD patient images. Crowdsourcing may have potential to provide the requisite expert-level data.

Materials and methods: Forty-one three-dimensional photographs of three cutaneous cGVHD patients were delineated by a board-certified dermatologist. 410 two-dimensional projections of the raw photos were each annotated by seven crowd workers, whose consensus performance was compared to the expert.

Results: The consensus delineation by four of seven crowd workers achieved the highest agreement with the expert, measured by a median Dice index of 0.7551 across all 410 images, outperforming even the best worker from the crowd (Dice index 0.7216). For their internal agreement, crowd workers achieved a median Fleiss's kappa of 0.4140 across the images. The time a worker spent marking an image had only weak correlation with the surface area marked, and very low correlation with accuracy. Percent of pixels selected by the consensus exhibited good correlation (Pearson $R = 0.81$) with the patient's affected surface area.

Correspondence: Eric Tkaczyk, Department of Veterans Affairs, Tennessee Valley Healthcare System, Nashville, TN. eric.tkaczyk@vanderbilt.edu.

Conclusion: Crowdsourcing may be an efficient method for obtaining demarcations of affected skin, on par with expert performance. Crowdsourced data generally agreed with the current clinical standard of percent body surface area to assess cGVHD severity in the skin.

Keywords

body surface area; crowdsourcing; graft-vs-host disease; machine learning; photography; stem cell transplantation; three-dimensional imaging

1 | INTRODUCTION

A major impediment to deriving clinical value from massive existing sets of medical photographs is the need for trained physician time to evaluate the images. At the same time, quantifying extent of affected skin is an important unmet clinical need both for research and practical management of patients suffering from burns, cutaneous lymphomas, drug-induced hypersensitivity syndrome, Stevens-Johnson syndrome, atopic dermatitis, erythrodermas, and many more diseases. To leverage the ongoing revolution of machine learning into the care of such patients, it is critical to reliably demarcate diseased areas in corresponding clinical photos. Herein, we explore the important example of chronic graft-vs-host disease (cGVHD) as a model for the potential of crowdsourced photo demarcation to overcome the fundamental bottleneck of limited expert time.

Chronic GVHD is the leading cause of nonrelapse long-term morbidity and mortality in hematopoietic stem cell transplantation (HCT) patients.[1] Skin is the most commonly affected organ in cGVHD,[2] resulting in distressing patient morbidity. Cutaneous disease burden is strongly associated with survival,[3] and so is often used as the primary endpoint metric in large multicenter trials.[4] Existing scales that assess clinical skin response in cGVHD are subjective.[5–8] Poor agreement and reproducibility of these scales have been a central issue impeding progress in cGVHD.[9] Despite many trials of promising potential therapies, still no specific treatments are approved based on randomized controlled trials. Photographs of skin are commonly used in the medical documentation of cGVHD. However, due to the lack of expert time these photographs do not have a defined role in patient care and have not been leveraged in objective, or even subjective, clinical studies of disease progression.

Crowdsourcing refers to distribution, through an online platform, of a task for simultaneous processing by multiple individuals (crowd workers). Crowdsourcing allows nonexpert workers to solve independent tasks for which there are no existing optimal computational algorithms. Tasks are designed to be simple, such that extensive training is unnecessary, and the necessary information to complete the task is presented alongside it. To accommodate for the lower accuracy of nonexpert workers, efforts are reduplicated by multiple workers. Repetitive tasks like data annotation can thus enjoy a markedly accelerated completion through the added value of many individuals working in parallel. Crowdsourced annotations of electronic health records and images have previously met with success in radiology.[10] Inspired by these achievements, we hypothesize that the collective opinion of inexperienced individuals with minimal training (ie, medical student and nursing staff crowd workers), as

harnessed by crowdsourcing, might enable reasonably accurate demarcations of cGVHD photos as compared to an expert. This would set the stage for machine learning applications.

2 | MATERIALS AND METHODS

2.1 | Dataset creation

With local IRB approval, the stereoscopic camera Vectra H1 (Canfield Scientific) was used to obtain 41 three-dimensional photographs of three cutaneous cGVHD patients. Through calibration of distance (via ranging lights), light intensity, and color (via a standardized flash), the camera is able to render 3D photos of body regions in a 27 cm × 16.5 cm × 10 cm capture volume with submillimeter resolution. The ground truth of affected areas was determined clinically by a board-certified dermatologist (ET) with specific interest in cGVHD, who highlighted these areas in 3D with the commercial Vectra software. An automatic script exported ten two-dimensional projections of each of the 41 three-dimensional photos, for a total dataset of 410 two-dimensional images, each with a corresponding ground truth annotation. Note that different projections of the same 3D photo could result in different crowd worker demarcations (Figure 1). However, the clinically determined ground truth is conserved across all 10 two-dimensional projections because it is marked on the original 3D photo.

2.2 | Crowdsourcing

The unlabeled images were loaded into a custom secure web-based demarcation tool (Figure 1A) for crowdsource annotation.[11] Seven workers without specific domain knowledge in either dermatology or oncology were recruited with compensation. Each worker was directed to complete the task of annotating all 410 two-dimensional images by highlighting all affected areas of skin. The only instructions provided were (a) a 12-slide Microsoft PowerPoint presentation with photos of GVHD and (b) five of the 410 images with the corresponding expert-determined ground truth. Using a digital paintbrush, each user was asked to demarcate the affected area in all images. Pixels that are not painted are considered unaffected.

2.3 | Timing analysis

The system recorded the exact time the user started and stopped annotating each image. From this, we were able to compare all results to the time an individual worker spent demarcating images.

2.4 | K-consensus

For each image, the k-consensus demarcation is defined as the set of pixels that k or more of the 7 crowd workers selected in their individual demarcations. Special cases are the 1-consensus (blue regions in Figure 1B), which comprises all regions that *any* worker thought might be affected; and the 7-consensus, which is simply the intersection of all crowd demarcations, that is, regions that *all* workers thought are affected (red regions in Figure 1B).

2.5 | Evaluation of performance

Performance of the crowd was evaluated based on the extent of agreement with the expert ground truth over the set of demarcated images. For any given demarcation of an image, agreement with the expert ground truth was determined by the Dice index.[12] Specifically, let X be the set of pixels assigned in the demarcation of the affected class and Y be the set of ground truth pixels in the affected class, and $X \cap Y$ be the intersection of these sets. Then, the Dice index is calculated as $\frac{2|X \cap Y|}{|X| + |Y|}$, in which the absolute value signs | | refer to the cardinality of each set.

2.6 | Surface area measurements

While photo demarcation as of yet has no role in clinical care, affected percent body surface area (%BSA) has been integrated into the NIH cGVHD skin score[13] as an important accepted clinical standard measure of disease severity. Therefore, we explored agreement of the crowd 2D demarcation with the ground truth affected surface area. As a %BSA surrogate, we employ the metric of pixel percent, defined as the quotient (in percent) of demarcated pixels divided by total skin area pixels in the photo. We also analyzed if the size of the affected region impacted crowd performance and user consistency. Here, we are assisted by the combination of a calibrated source 3D photo providing absolute affected surface area in cm^2 corresponding to 10 different 2D projections. Thus, for each unique 3D photo, it is possible to determine the median and standard deviation of the Dice index of the corresponding crowd demarcation.

3 | RESULTS

3.1 | Crowd performance

The performance of the k-consensus is plotted in Figure 2A (red bars) and the performance of individual users is shown in Figure 2B. As a function of k, the crowd performance is an approximate inverse parabolic function with a relative plateau from 3 to 5 workers. Performance is worst at the extremes (1-consensus and 7-consensus), where only 1 or all workers are required to select a pixel. The highest agreement (median Dice = 0.7551) was reached with 4 of 7 workers, that is, 4-consensus. As one of the users (user 16) performed almost as well as the 4-consensus (median Dice = 0.7216), we re-analyzed with the omission of her data (Figure 2A, blue bars). In this case, the optimum remained 4 users, albeit with a median Dice index (0.7167) slightly worse than the top user. Effectively, the best user from the crowd performed on par with the collection of the remaining users. However, having input from the entire crowd is better than simply relying on the top user. Agreement between the individual users across all 2D images exhibited a median Fleiss's kappa of 0.4140 (standard deviation 0.2001).

3.2 | Timing analysis

The time spent marking an image was only weakly correlated with the number of pixels a worker selected (overall linear Pearson correlation coefficient $R = 0.37$, range over individual users −0.06 to 0.71) and even more weakly correlated with his or her accuracy (Dice index with the expert for a particular image) ($R = 0.15$, range of individual users 0.09

to 0.59). Therefore, there appears to be little advantage in biasing or selecting for crowd workers who spend more time in their annotations.

3.3 | Accuracy of crowdsourced surface area measurement

The crowd-determined pixel percent (4-consensus pixel percent) exhibited good correlation ($R = 0.81$) with the expert-determined affected %BSA (Figure 3A). Thus, crowdsourcing results could be directly translated into an NIH skin score, with reasonable agreement with an expert clinical exam, raising the possibility of eventual direct clinical utility.

3.4 | Effect of surface area on crowd performance and consistency

Across the 41 different 3D photos, the crowd performance (measured by median Dice of the 4-consensus with the expert ground truth) had a moderate correlation (Pearson linear correlation coefficient $R = 0.44$) with the percent affected skin (percent of photographed surface area that was marked as affected in the 3D ground truth) (Figure 3B). Similarly, the standard deviation itself has a moderate negative correlation ($R = -0.47$) with the ground truth %BSA. Thus, both the accuracy and precision of the consensus are higher for regions that have a larger fraction of affected area.

4 | DISCUSSION

Our results suggest that crowdsourcing can be an efficient method for obtaining demarcations of cGVHD affected areas that are on par with expert demarcation, raising a possible practical approach to generating machine learning datasets. These could be specific to an institution and protocol (both physical device and photography procedure). The highest median Dice index (0.76) was obtained through the 4-consensus of the 7 workers, which was superior to even the top individual user's performance. Further research is needed to determine if a 4-consensus is best with different photographs and different workers. Generally, the crowd tended to perform better both in precision (reproducibility) and accuracy (agreement with the expert) in demarcating more severely affected patients (larger %BSA). Notably, as previously described, the Dice index is an imperfect metric whose limitations include a bias against small areas of affected skin.[14] In our results, we observed large qualitative variation in demarcations of different 2D projections of a single 3D photo, even when Dice indices relative to the truly affected region were very similar. Importantly, we observed good agreement of pixel percent derived from crowd demarcations with the true %BSA. This suggests that even without further technological development, crowdsourcing could have potential application in the current clinically accepted BSA-based scoring of cGVHD.

The study has several limitations which must be acknowledged. Firstly, the 410 images marked by the seven workers only came from three unique patients. Secondly, the ground truth was determined by a single clinical opinion. However, variation in assessing extent of cutaneous cGVHD is well documented,[15] and rating 3D photos likely has even greater disagreement.[16] Thirdly, the entire affected area was marked by the crowd, but current practice in cGVHD assessment has recently fallen away from including visual post-inflammatory changes (eg, poikiloderma and other pigmentary changes) in the assessment of

affected BSA.[13] However, opinions continue to vary on this point, and we have recently observed median kappas of 0.3544 among top GVHD experts in demarcating only active areas of disease in 2D photos (unpublished data). Thus, crowdsourcing will likely be less successful in discriminating active disease from post-inflammatory changes. Notably, the Vienna Total Skin Score still includes post-inflammatory and pigmentary changes[7] and is a validated primary end point in multicenter trials.[4] Therefore, despite all of the above limitations, there is potential for crowdsourced measurements of total affected cGVHD area to become a practical and much-needed tool both in cGVHD clinical practice and research, including machine learning image analysis applications such as deep learning. More broadly, this work demonstrates the feasibility of crowdsourcing complex medical imaging phenotypes to obtain results that are of acceptable similarity to expert evaluation.

5 | CONCLUSION

A significant impediment for patient care and research in cGVHD and many other diseases affecting the skin is the limited availability of expert time to annotate images or to even perform reliable skin surface area assessments in clinical day-to-day practice.[17] Consensus opinion derived from crowd workers can provide reasonably accurate demarcations of affected skin in photos of cGVHD patients, and the pooled opinion of four of seven workers provided the most reliable demarcation in this study. The crowdsourced data from 410 images generally agreed with the current clinical standard of %BSA to assess cGVHD severity. This suggests that further crowdsourcing studies could potentially be used to train artificial intelligence approaches to measuring cGVHD and result in applications of direct clinical utility in diagnosing and staging individual patients with complex and severe diseases.

ACKNOWLEDGEMENTS

The authors are grateful for the patients who participated in this research as well as to Professor Bradley Malin for his suggestions on the manuscript.

This work was supported by Career Development Award Number IK2 CX001785 from the United States (U.S.) Department of Veterans Affairs Clinical Sciences R&D (CSRD) Service (Dr. Eric Tkaczyk), NCI K12 grant number CA0906525 from NIH (Dr. Eric Tkaczyk), a Discovery Grant from Vanderbilt University (Dr. Benoit Dawant), and NIH grant number 1 UH2 CA203708–01 (Dr. Daniel Fabbri).

Funding information

National Cancer Institute, Grant/Award Number: CA0906525; U.S. Department of Veterans Affairs, Grant/Award Number: IK2 CX001785; National Institutes of Health, Grant/Award Number: 1 UH2 CA203708–01; Vanderbilt University, Grant/Award Number: Discovery Grant

REFERENCES

1. Socié G, Ritz J. Current issues in chronic graft-versus-host disease. Blood. 2014;124(3):374–384. [PubMed: 24914139]
2. Rodgers CJ, Burge S, Scarisbrick J, et al. More than skin deep? Emerging therapies for chronic cutaneous GVHD. Bone Marrow Transplant. 2013;48(3):323–337. [PubMed: 22863725]
3. Curtis LM, Grkovic L, Mitchell SA, et al. NIH response criteria measures are associated with important parameters of disease severity in patients with chronic GVHD. Bone Marrow Transplant. 2014;49(12):1513–1520. [PubMed: 25153693]

4. Flowers ME, Apperley JF, Van Besien K, et al. A multicenter prospective phase 2 randomized study of extracorporeal photopheresis for treatment of chronic graft-versus-host disease. Blood. 2008;112(7):2667–2674. [PubMed: 18621929]
5. Pavletic SZ, Martin P, Lee SJ, et al. Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: IV. Response criteria working group report. Biol Blood Marrow Transplant 2006;12(3):252–266. [PubMed: 16503494]
6. Filipovich AH, Weisdorf D, Pavletic S, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: I. Diagnosis and staging working group report. Biol Blood Marrow Transplant 2005;11(12):945–956. [PubMed: 16338616]
7. Greinix HT, Pohlreich D, Maalouf J, et al. A single-center pilot validation study of a new chronic GVHD skin scoring system. Biol Blood Marrow Transplant. 2007;13(6):715–723. [PubMed: 17531782]
8. Jacobsohn DA, Chen AR, Zahurak M, et al. Phase II study of pentostatin in patients with corticosteroid-refractory chronic graft-versus-host disease. J Clin Oncol. 2007;25(27):4255–4261. [PubMed: 17878478]
9. Palmer JM, Lee SJ, Chai X, et al. Poor agreement between clinician response ratings and calculated response measures in patients with chronic graft-versus-host disease. Biol Blood Marrow Transplant. 2012;18(11):1649–1655. [PubMed: 22691695]
10. Cocos A, Qian T, Callison-Burch C, et al. Crowd control: effectively utilizing unscreened crowd workers for biomedical data annotation. J Biomed Inform. 2017;69:86–92. [PubMed: 28389234]
11. Ye C, Coco J, Epishova A, et al. A crowdsourcing framework for medical data sets. AMIA Jt Summits Transl Sci Proc. 2018;2017:273–280. [PubMed: 29888085]
12. Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3): 297–302.
13. Jagasia MH, Greinix HT, Arora M, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: I. The 2014 Diagnosis and Staging Working Group report. Biol Blood Marrow Transplant 2015;21(3):389–401. [PubMed: 25529383]
14. Wang J, Chen F, Dellalana LE, et al. Segmentation of skin lesions in chronic graft versus host disease photographs with fully convolutional networks. Proc SPIE Int Soc Opt Eng. 2018;10575:10575N-1–10575N-7
15. Mitchell SA, Jacobsohn D, Powers KE, et al. A multicenter pilot evaluation of the National Institutes of Health chronic graft-versus-host disease (cGVHD) therapeutic response measures: feasibility, inter-rater reliability, and minimum detectable change. Biol Blood Marrow Transplant. 2011;17(11):1619–1629. [PubMed: 21536143]
16. Tkaczyk ER, Chen F, Wang J, et al. Overcoming human disagreement assessing erythematous lesion severity on 3D photos of chronic graft-versus-host disease. Bone Marrow Transplant. 2018;53:1356–1358. [PubMed: 29740182]
17. Carpenter PA. How I conduct a comprehensive chronic graft-versus-host disease assessment. Blood. 2011;118(10):2679–2687. [PubMed: 21719600]

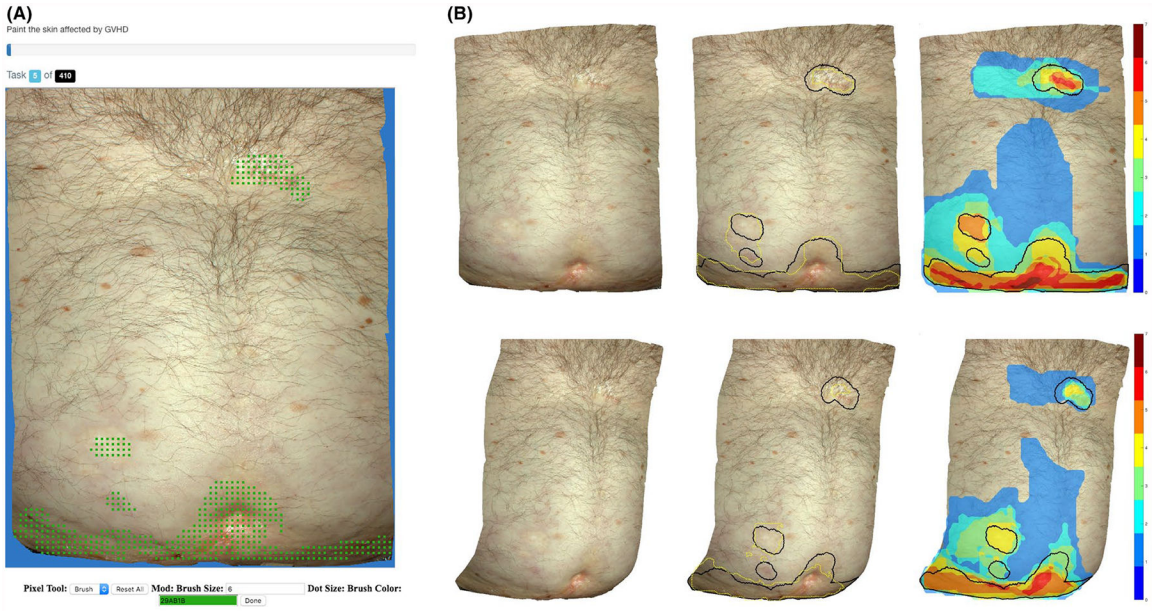

FIGURE 1.

A, Interface for web-based demarcation tool. B, Generated data for two projections (top and bottom left) from the same 3D source photo. Far right: individual worker data are accumulated to build different consensus-level images (different colors). Middle: Dice indices 0.829 (top) and 0.740 (bottom) of the 4-consensus (yellow) compared to the expert ground truth (black) in two 2D exports of the same 3D image

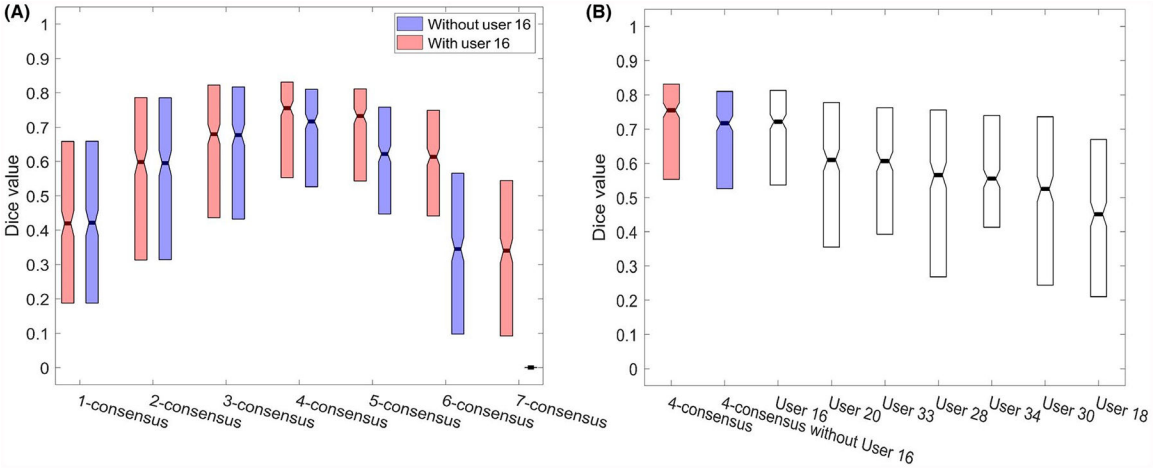

FIGURE 2.

Boxplot showing the median and interquartile range of Dice index (relative to the expert ground truth), reflecting the accuracy of crowd worker demarcations of the 410 test images. A, Performance of different consensus-level segmentations, both with (red) and without (blue) incorporation of input from the most accurate user (user 16). B, Performance of individual users as well as the 4-consensus (median Dice 0.7551 including user sixteen; 0.7167 excluding user sixteen). The most accurate user (user sixteen) had a median Dice index of 0.7216 with the expert ground truth

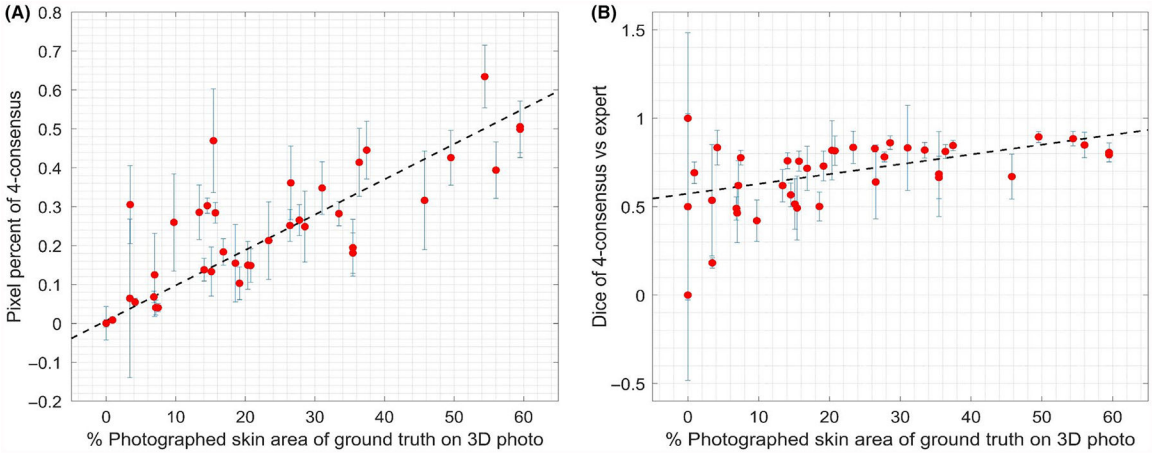

FIGURE 3.

A, Fraction of photographed skin in 2D images selected as affected by the crowd 4-consensus (in percent of pixels) vs the percent of the photographed skin surface area that was demarcated as affected by the expert dermatologist in 41 three-dimensional photos. Linear fit has a Pearson R of 0.81. B, Median agreement (Dice index) of the 4-consensus with the expert vs percent affected surface area for 41 three-dimensional photos (Pearson $R = 0.44$). Error bars represent the standard deviation over all 10 images exported from a single expert-demarcated 3D photograph