

Original Paper

# Crowdsourcing Skin Demarcations of Chronic Graft-Versus-Host Disease in Patient Photographs: Training Versus Performance Study

Andrew J McNeil<sup>1,2,3</sup>, PhD; Kelsey Parks<sup>2</sup>, MD; Xiaoqi Liu<sup>2,3</sup>, PhD; Bohan Jiang<sup>1,2,3</sup>, BS; Joseph Coco<sup>4</sup>, MS; Kira McCool<sup>5</sup>, BA; Daniel Fabbri<sup>4</sup>, PhD; Erik P Duhaime<sup>5</sup>, MPhil, PhD; Benoit M Dawant<sup>3</sup>, PhD; Eric R Tkaczyk<sup>1,2,3,4</sup>, MD, PhD

<sup>1</sup>Dermatology Service and Research Service, Department of Veterans Affairs, Tennessee Valley Healthcare System, Nashville, TN, United States

<sup>2</sup>Department of Dermatology, Vanderbilt University Medical Center, Nashville, TN, United States

<sup>3</sup>Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, United States

<sup>4</sup>Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

<sup>5</sup>Centaur Labs, Boston, MA, United States

**Corresponding Author:**

Eric R Tkaczyk, MD, PhD

Dermatology Service and Research Service

Department of Veterans Affairs

Tennessee Valley Healthcare System

1310 24th Avenue South

Nashville, TN, 37212

United States

Phone: 1 6159364633

Email: [eric.tkaczyk@vumc.org](mailto:eric.tkaczyk@vumc.org)

## Abstract

**Background:** Chronic graft-versus-host disease (cGVHD) is a significant cause of long-term morbidity and mortality in patients after allogeneic hematopoietic cell transplantation. Skin is the most commonly affected organ, and visual assessment of cGVHD can have low reliability. Crowdsourcing data from nonexpert participants has been used for numerous medical applications, including image labeling and segmentation tasks.

**Objective:** This study aimed to assess the ability of crowds of nonexpert raters—individuals without any prior training for identifying or marking cGVHD—to demarcate photos of cGVHD-affected skin. We also studied the effect of training and feedback on crowd performance.

**Methods:** Using a Canfield Vectra H1 3D camera, 360 photographs of the skin of 36 patients with cGVHD were taken. Ground truth demarcations were provided in 3D by a trained expert and reviewed by a board-certified dermatologist. In total, 3000 2D images (projections from various angles) were created for crowd demarcation through the DiagnosUs mobile app. Raters were split into high and low feedback groups. The performances of 4 different crowds of nonexperts were analyzed, including 17 raters per image for the low and high feedback groups, 32-35 raters per image for the low feedback group, and the top 5 performers for each image from the low feedback group.

**Results:** Across 8 demarcation competitions, 130 raters were recruited to the high feedback group and 161 to the low feedback group. This resulted in a total of 54,887 individual demarcations from the high feedback group and 78,967 from the low feedback group. The nonexpert crowds achieved good overall performance for segmenting cGVHD-affected skin with minimal training, achieving a median surface area error of less than 12% of skin pixels for all crowds in both the high and low feedback groups. The low feedback crowds performed slightly poorer than the high feedback crowd, even when a larger crowd was used. Tracking the 5 most reliable raters from the low feedback group for each image recovered a performance similar to that of the high feedback crowd. Higher variability between raters for a given image was not found to correlate with lower performance of the crowd consensus demarcation and cannot therefore be used as a measure of reliability. No significant learning was observed during the task as more photos and feedback were seen.

**Conclusions:** Crowds of nonexpert raters can demarcate cGVHD images with good overall performance. Tracking the top 5 most reliable raters provided optimal results, obtaining the best performance with the lowest number of expert demarcations required for adequate training. However, the agreement amongst individual nonexperts does not help predict whether the crowd has provided an accurate result. Future work should explore the performance of crowdsourcing in standard clinical photos and further methods to estimate the reliability of consensus demarcations.

(*JMIR Dermatol* 2023;6:e48589) doi: [10.2196/48589](https://doi.org/10.2196/48589)

## KEYWORDS

graft-versus-host disease; cGVHD; crowdsourcing; dermatology; labeling; segmentation; skin; medical image; imaging; feasibility; artificial intelligence

## Introduction

Chronic graft-versus-host disease (cGVHD) is the leading cause of nonrelapse long-term morbidity and mortality in patients after allogeneic hematopoietic cell transplantation [1]. Skin is the earliest and most commonly affected organ [2]. Changes in cutaneous manifestations are used to evaluate treatment efficacy and disease progression, assessed by the affected surface area involvement. The National Institutes of Health (NIH) Skin Score is the primary outcome measure in major clinical trials. It is a composite score derived from assessment of the body surface area of erythema (categorical approximation), moveable sclerosis, and nonmoveable sclerosis, combined with a functional and symptomatic assessment [3,4]. The clinical trial for the first Food and Drug Administration–approved (2017) cGVHD treatment, ibrutinib, had a minimum of 25% body surface area erythema as an inclusion criterion [5]. In practice, surface area is estimated by visual assessment using either the Wallace rule of nines, Lund and Browder chart, or palmar units [6]. However, visual assessment of cGVHD suffers from low reliability. Mitchell et al [7] found that the threshold for defining change exceeding measurement error was 19%-22% of the entire body surface area for erythema. This poses a significant barrier to improving patient care through accurate tracking of disease severity and is compounded by the low availability of expert dermatologist evaluation [8].

A recent multicenter cohort study showed that for erythema-type cGVHD, percentage body surface area involvement was a better predictor of mortality than the categorical NIH cGVHD Skin Score [8]. Automated analysis of body surface area from photographs by artificial intelligence (AI) image analysis has shown promise, with a recent study finding that 77% of AI demarcations were scored as clinically acceptable by a board-certified dermatologist across more than 300 photos [9]. Further development of such engineering solutions is greatly hampered by the cost and difficulty of collecting expert demarcations for large numbers of photographs for training and validation.

Crowdsourcing data from a large number of nonexpert participants has been widely used for many medical applications [10,11], including bioinformatics [12], histology image labelling and cell segmentation [13-15], demarcating organs and regions of disease in both 2D and 3D radiology images [16,17], and combining crowd opinions with AI models for improving the severity scoring of diabetic retinopathy [18]. Recent work has also shown expert-level crowd performance for identifying some

features of pigmented skin lesions in dermoscopic images, which comprise high magnification, narrow field of view cross-polarized photos of the skin surface [19].

Our study aimed to assess the ability of a crowd of nonexpert raters to demarcate photos of cGVHD-affected skin, which could provide a scalable solution for demarcating large numbers of patient photos for AI training. Cutaneous cGVHD often presents as complex areas of erythema and surface changes with ill-defined borders. This demarcation task typically requires significant training and is known to exhibit high variability even among experts [7]. To study the effect of training and feedback on crowd performance, we split raters into 2 groups (high and low feedback) that each received a different amount of ground truth feedback during data collection.

## Methods

### Materials

Patient characteristics are given in Table 1. The patient cohort had an age range of 21-72 (median 58, IQR 46-66) years and were photographed at 6 to 4520 days post–hematopoietic cell transplantation (median 1092, IQR 266-1724 days). Cutaneous cGVHD presented as erythema for 7 patients, sclerosis for 7 patients, and both erythema and sclerosis for 9 patients.

We took 360 3D photographs of the skin of 36 patients with cGVHD using a handheld commercial stereoscopic camera (Vectra H1, Canfield Scientific). This stereoscopic camera provides a cross-polarized flash and ranging lights to improve the consistency of photographic conditions between body sites and patients. It also enables accurate ground truth markings of affected skin areas directly on the 3D surface, which can generate multiple 2D views of the same skin from different angles to more closely emulate standard clinical photography. From each 3D photo, a set of 2D images were created from different viewing angles using Vectra Analysis Module software (Canfield) following an automated scripting protocol, as described in previous work [9,20]. Defining the original camera view as an angle of 0 degrees, we rotated the skin surface through combinations of 0 degrees, +15 degrees, and -15 degrees along the horizontal and vertical axes to create multiple views of the same skin area under identical photographic conditions. Table 2 shows how the photo set was split into photos for training and feedback (7 angles per photo) and photos for testing crowd performance (9 angles per photo). We refer to the training set as the “ground truth provided” set and the test set as the “ground truth withheld” set (Table 2). In total, the full set

consisted of 3000 2D images for demarcation by the crowd, with cGVHD-affected skin (per ground truth), with the test set consisting of 711 ground truth withheld images

**Table 1.** Race, gender, ethnicity, and Fitzpatrick skin types of patients involved in the study.

Characteristic	Patients (n=36)
<b>Race, n (%)</b>	
American Indian or Alaska Native	0 (0)
Asian	0 (0)
Black or African American	3 (8)
Native Hawaiian or Other Pacific Islander	0 (0)
White	33 (92)
<b>Gender, n (%)</b>	
Man	26 (72)
Woman	10 (28)
<b>Ethnicity, n (%)</b>	
Hispanic or Latino	0 (0)
Not Hispanic or Latino	36 (100)
<b>Fitzpatrick skin type, n (%)</b>	
I	0 (0)
II	8 (22)
III	24 (67)
IV	1 (3)
V	1 (3)
VI	2 (7)
Age (years), median (IQR)	58 (46-66)
Time since transplant (days), median (IQR)	1092 (266-1724)

**Table 2.** Distribution of photos and patients between the feedback (ground truth provided) and analysis (ground truth withheld) sets. For each stereoscopic photo, 9 images from different angles were produced.

Data category	Ground truth provided set	Ground truth withheld set
<b>cGVHD<sup>a</sup>-affected</b>		
Photos (n=179), n (%)	100 (56)	79 (44)
Patients (n=25), n (%)	19 (76)	22 (88)
<b>cGVHD-unaffected</b>		
Photos (n=181), n (%)	20 (11)	161 (90)
Patients (n=23), n (%)	18 (78)	22 (96)
<b>Total</b>		
Photos (n=360), n (%)	120 (33)	240 (67)
Patients (n=36), n (%)	31 (86)	34 (94)

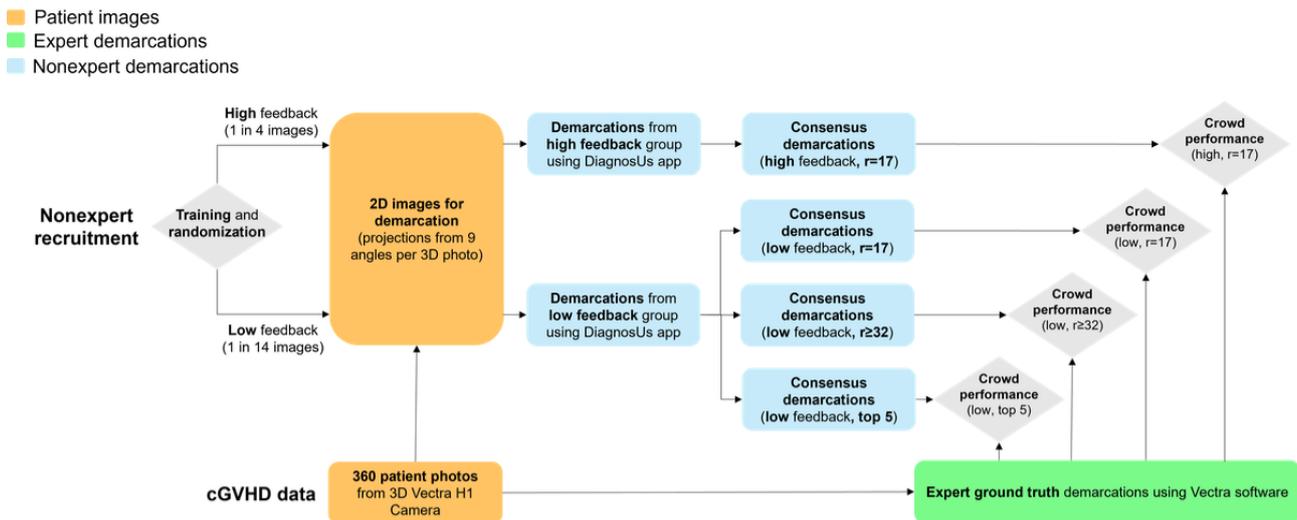
<sup>a</sup>cGVHD: chronic graft-versus-host disease.

## Study Design

A flowchart of the study design is shown in [Figure 1](#). Patient photos were first demarcated by a trained expert to provide the ground truth. A crowd of nonexperts were then recruited through the Centaur Labs' DiagnosUs app and randomized into low and

high feedback groups after training. Crowd demarcations were gathered for each image, which were combined into separate consensus demarcations for each crowd. Performance was assessed by comparing the consensus demarcations to the expert ground truth.

**Figure 1.** Flowchart of the study design. cGVHD: chronic graft-versus-host disease; r: rater.

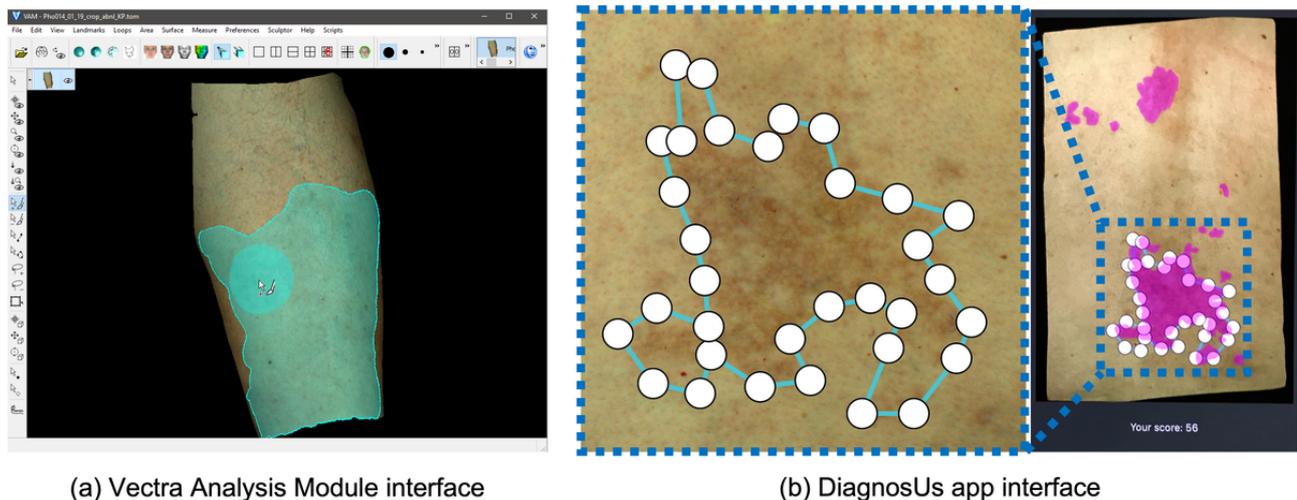


## Ground Truth Demarcations

A single expert (KP) provided ground truth demarcations of cGVHD-affected skin following an extensive training program [21], which were reviewed by a board-certified dermatologist (ERT) for accuracy. Affected skin areas were demarcated on the 3D skin surface using Vectra Analysis Module software

(Figure 2a). The skin surface could be rotated and zoomed in 3D space, with affected skin areas demarcated using a paintbrush tool. This allowed us to create the most accurate demarcations of affected skin from the visual appearance alone. These ground truth areas were used for training the crowd through visual feedback and evaluating their performance.

**Figure 2.** Annotation interfaces used for (A) ground truth demarcations using the Vectra Analysis Module and (B) crowd demarcations using the DiagnosUs app, including ground truth feedback during training.



(a) Vectra Analysis Module interface

(b) DiagnosUs app interface

## Crowd Training and Data Collection

The DiagnosUs app gamifies medical image demarcation tasks, creating time-limited competitions with leaderboards and prizes to incentivize engagement. Only 2D images are supported through the app and its mobile interface, necessitating the use of projected images of the skin surface. The interface uses the touch screen of a mobile device for demarcating areas using nodes, which outline the desired shape (Figure 2b). Multiple nontouching areas can be marked on a single image if needed, and node positions can be adjusted after placing before the final submission.

We used 8 images to train all raters before they began the demarcation task (Figure 3). For each training image, the rater

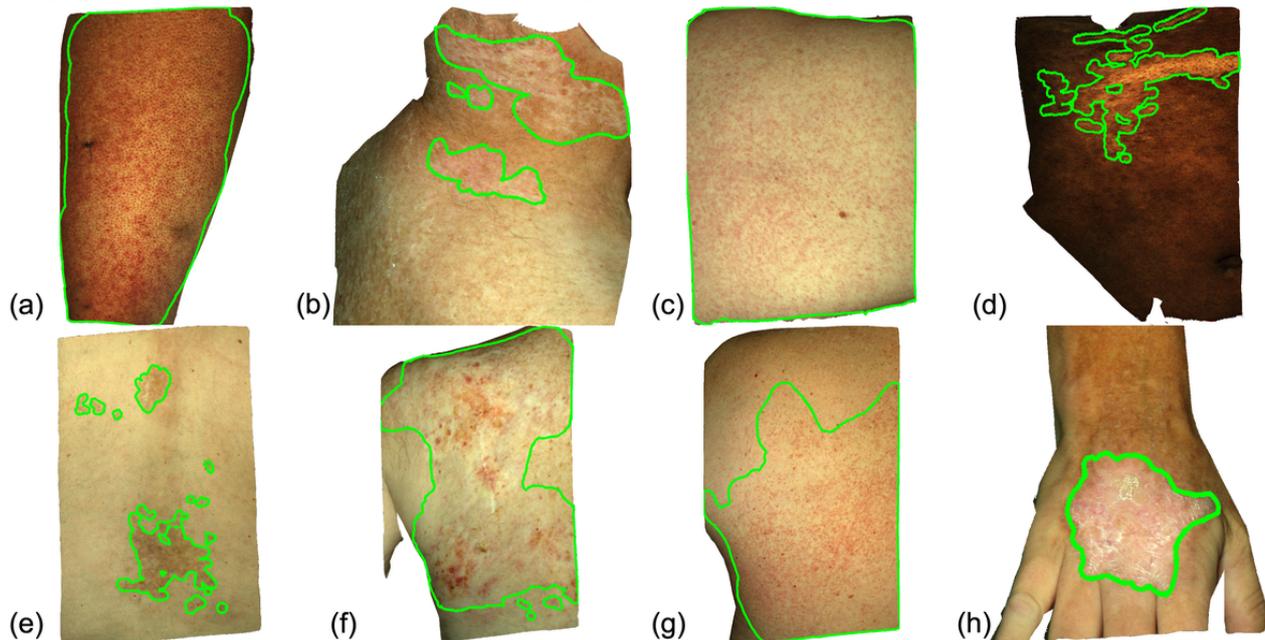
was first shown the unmarked image with corresponding text by the expert dermatologist describing what features they should look for (Table 3). Upon submitting their demarcation, the rater was then shown the expert ground truth and their accuracy score. Training was completed once all 8 training images had been adequately marked. Each rater was randomly assigned to the high or low feedback group. No knowledge of the different groups or their assignments were communicated to the raters.

We held 8 24-hour competitions for data collection, with each user asked to demarcate 200 randomly selected images from the full set of 3000. Cash prizes were offered for each competition based on performance ranking, and all 200 images needed to be demarcated to be eligible for a given competition. Images were split into those for which the ground truth may be

provided for feedback and quality assurance and those for which the correct answer was entirely withheld. Crowd performance was assessed on the 711 images of affected skin for which the ground truth was entirely withheld (never released to Centaur Labs). Raters in the high feedback group received feedback on 1 out of every 4 cases, while those in the low feedback group received feedback on 1 out of every 14 cases. For each image, we recorded the first 17 rater ( $r$ ) opinions in both the high and low feedback groups, denoted as " $r=17$ ". To test if a larger

crowd could overcome the expected performance drop from less training feedback, data collection was extended for the low feedback group up to 32-35 demarcations per image, denoted as " $r \geq 32$ ". Finally, the effect of tracking only the most reliable raters was examined. The performance of individuals was tracked on the images for which the ground truth was provided (ground truth provided set in Table 2), and the 5 best performers for each image in the ground truth withheld set were selected in the low feedback group, denoted as the "top 5" group.

**Figure 3.** The 8 images used for training the crowd during study enrollment. Ground truth demarcations of cGVHD-affected skin are shown in green. The corresponding text descriptions of each disease presentation are given in Table 3. cGVHD: chronic graft-versus-host disease.



**Table 3.** Text descriptions provided for each of the 8 training images shown in Figure 3.

Image <sup>a</sup>	Description
(a)	On the anterior left thigh, there are numerous diffuse erythematous macules and papules coalescing into patches and plaques. When large areas of the skin are affected by cGVHD <sup>b</sup> , visual comparison to healthy skin can be difficult.
(b)	On the right, posterior neck of a patient with sun damage, there is a large scaly, hypopigmented patch with erythematous borders.
(c)	On the back, there are diffuse, erythematous papules that coalesce into a plaque in the central, left portion of the image. This is an example of a morbilliform eruption, which can be caused by cGVHD. The entirety of the skin image may be affected by cGVHD, as in this example, so careful attention must be paid to all areas shown.
(d)	On the right upper quadrant of the abdomen, there is a well-defined hypopigmented patch. As skin tones vary between patients, visual comparisons should be made with unaffected regions on each image when possible.
(e)	On the back, there are several well-defined, hyperpigmented patches consistent with post-inflammatory changes that can occur in cutaneous cGVHD.
(f)	cGVHD may present as a large area of sclerosis with superimposed areas of color and texture changes.
(g)	On the back, there are numerous diffuse erythematous macules and papules. Areas of rash may be ill-defined, so care must be taken to examine the skin in detail. The annotation boundaries should be carefully placed to encompass all high certainty regions of affected skin.
(h)	On the dorsum of the right hand, there is a single well-defined, scaly erythematous plaque.

<sup>a</sup>Image letters correspond to the panels in Figure 3.

<sup>b</sup>cGVHD: chronic graft-versus-host disease.

## Crowd Consensus Demarcations

We constructed 4 different sets of consensus demarcations from the 4 constructed crowds ( $r=17$  from the high feedback group and  $r=17$ ,  $r\geq 32$ , and top 5 from the low feedback group). Each crowd's consensus demarcation for an image was calculated by simple majority vote. For a given image, each pixel was labelled with the number of raters who marked it as affected. The consensus demarcation consisted of all pixels labelled by the plurality of raters (50% or more) in the crowd being analyzed, following the standard majority vote method [22]. The final consensus mask provides a binary label for every pixel in the image, being either cGVHD-affected skin or cGVHD-unaffected skin. The consensus demarcation for each crowd was considered their best estimate of cGVHD-affected skin for the given image.

## Agreement Measures

Crowd demarcations were analyzed only for photos for which the ground truth was never provided to any of the app users. We used 2 metrics: the Dice coefficient and the surface area error. To measure spatial overlap, we used the machine vision metric of the Dice coefficient [23], which ranges from 0 for no overlap to 1 for perfect agreement. For context, a recent study of 3 nonexperts who underwent an extensive 4-month training program for demarcating cGVHD led by a board-certified dermatologist were found to achieve a median Dice of 0.75 (IQR 0.68-0.84) when compared to the expert [21]. While commonly used for comparing demarcations in many medical imaging tasks, use of the Dice metric alone has been shown to be inadequate for capturing training effects for cGVHD skin demarcations [21]. Therefore, we also calculated the surface area error, which represents the absolute difference in the percentage of skin area marked by the crowd compared to the ground truth. For example, if the ground truth demarcation covered 10% of the skin area and the crowd marked 25%, then the surface area error is 15%. This performance measure of skin area estimation parallels the scoring accuracy measures often used for in-person clinical assessment [7].

## Learning Effects

The effect of feedback and experiential learning on the performance of the crowd was examined by tracking the performances of individual raters over the first 100 images with cGVHD-affected skin as the ground truth.

## Ethical Considerations

This study was performed using nonidentifiable photographs under Vanderbilt University institutional review board exemption 191042.

## Results

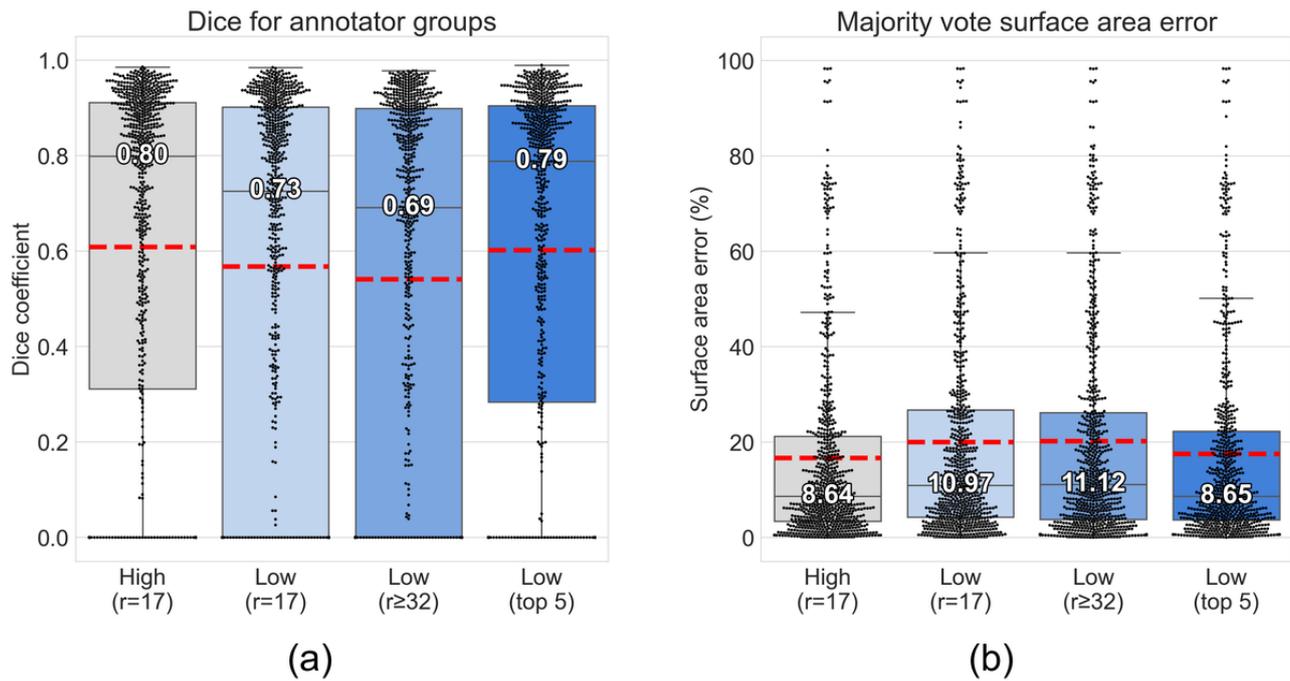
### Demarcating cGVHD-Affected Skin

Combined across all 8 competitions, a total of 291 raters were recruited, with 130 (45%) assigned to the high feedback group and 161 (55%) to the low feedback group, 111 (38%) of which contributed to the low  $r=17$  crowd. This produced a total of 133,854 individual demarcations, including 54,887 (41%) from the high feedback group and 78,967 (59%) from the low feedback group.

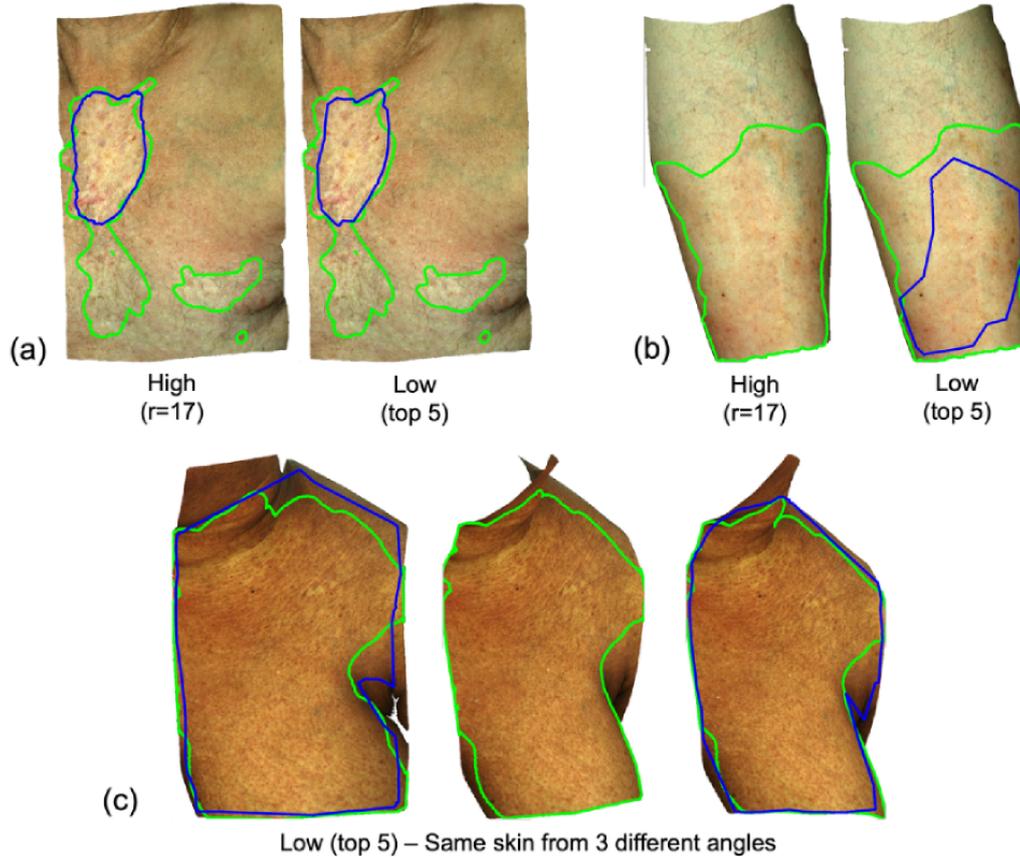
The subset of photos for which no ground truth was shown to any user included 79 photos from 22 patients with cGVHD-affected skin per the completely withheld ground truth (Table 2). To avoid the possibility that a user might have somehow seen the solutions to their delineation task for a particular patch of skin, we assessed the performance of the crowd on these 711 images (9 angles per photo) only. Figure 4a shows the crowd performance by Dice. The high feedback  $r=17$  crowd and low feedback top 5 crowd were not significantly different according to a Mann-Whitney  $U$  test ( $P=.64$ ), with a Dice coefficient of 0.8. Compared to the high feedback crowd, both the low  $r=17$  and low  $r\geq 32$  crowds were significantly different, with Dice coefficients of 0.7. Figure 4b shows the surface area error for each group. The high feedback and low feedback top 5 crowds had a surface area error of 9%, but the other 2 low feedback crowds were significantly different at 11%.

Figure 5 shows examples of inconsistencies observed across the crowds. Figure 5a shows consistent demarcation of highly affected areas by both the high and low top 5 crowds; however, there was poor identification of subtle surface changes, which were also marked in the expert ground truth. Figure 5b demonstrates an instance where the high feedback crowd failed to identify abnormal changes while the low top 5 crowd identified 75% of the abnormal skin area. Figure 5c highlights an instance of high variability between images from different angles of same skin region in the low top 5 crowd, where 2 images show good agreement with the ground truth, but the third image consensus predicted no affected skin. In all cases, we also noted the sharp edge and lower specificity of areas marked by the crowd. This is likely due to the mobile interface providing lower fidelity for marking complex shapes as compared to the Vectra Analysis Module software used for marking the ground truth (Figure 2).

**Figure 4.** Performance of crowd groups for demarcating images with cGVHD-affected skin per ground truth using the (A) Dice coefficient and (B) surface area error. Each point represents the majority vote mask for a single image (711 images in total). Whiskers indicate  $1.5 \times$  IQR. Mean values are shown indicated by the dashed red line. cGVHD: chronic graft-versus-host disease; r: rater.



**Figure 5.** Example demarcations from the crowd (blue) versus ground truth (green). (A) Consistent demarcation of highly affected areas by crowds assembled from both the high and low feedback groups, but both missed areas of subtle surface changes. (B) The high feedback crowd failed to identify abnormal changes while low feedback top 5 crowd identified 75% of abnormal skin areas. (C) High variability between images of the same skin region viewed from different angles by the low feedback top 5 crowd. cGVHD: chronic graft-versus-host disease; r: rater.



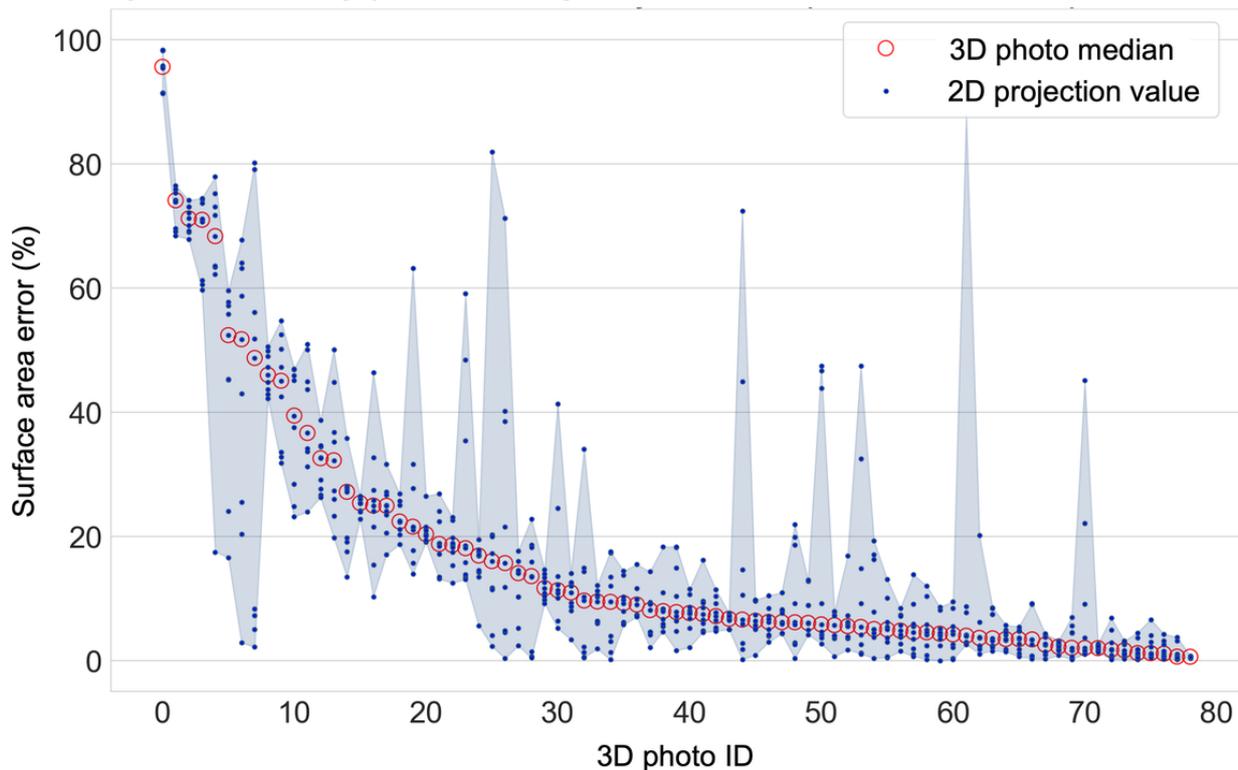
## Reliability of Demarcations

Despite good median performance of the different crowds across the full set, we observed a number of high error images in all groups (Figure 4b). Because the expert ground truth was available in 3D on our unique photo set but crowd annotation was done in 2D, we were able to test the performance of the crowd on the same area of skin under identical photographic conditions from different viewing angles. Figure 6 shows the surface area error for the low feedback top 5 crowd for each of the 9 projected images from each 3D photo, ordered by descending median error. We observe outlier images with significant errors in both high and low median error photos. The set of individual raters contributing to the consensus

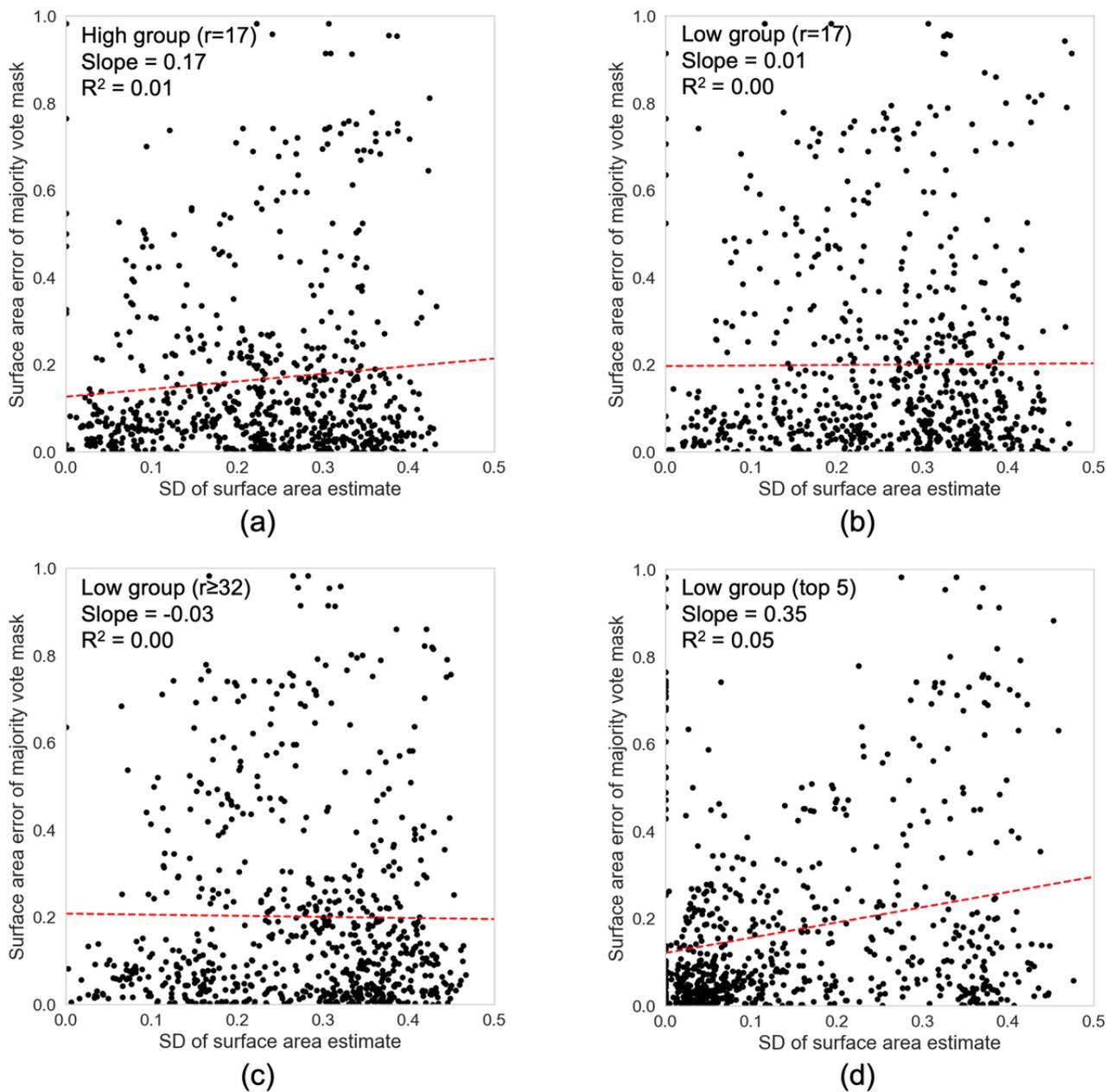
demarcation will vary between images, suggesting that interrater variability could contribute to the inconsistent reliability across images of the same skin.

To test if there was an observable association between the level of disagreement and the accuracy of the crowd demarcations, Figure 7 shows correlation plots of the surface area error of the consensus mask for a given photo against the SD of the crowd's estimate of surface area for that photo. We found no significant correlation between the variability of individual raters and the accuracy of the consensus mask, with a near-zero coefficient of determination for simple linear regression in all groups. Therefore, the level of disagreement between raters cannot be used as a measure of reliability for this task.

**Figure 6.** Per-photo surface area error for the low feedback group (top 5 raters). 3D photo IDs are ordered by decreasing median error. The shaded area shows the range of error between 2D projections for each 3D photo.



**Figure 7.** Surface area error of the majority vote mask versus the SD of surface area estimates for each photo. Slope and coefficient of determination ( $R^2$ ) for the linear regression fit (red dashed line) are also given.

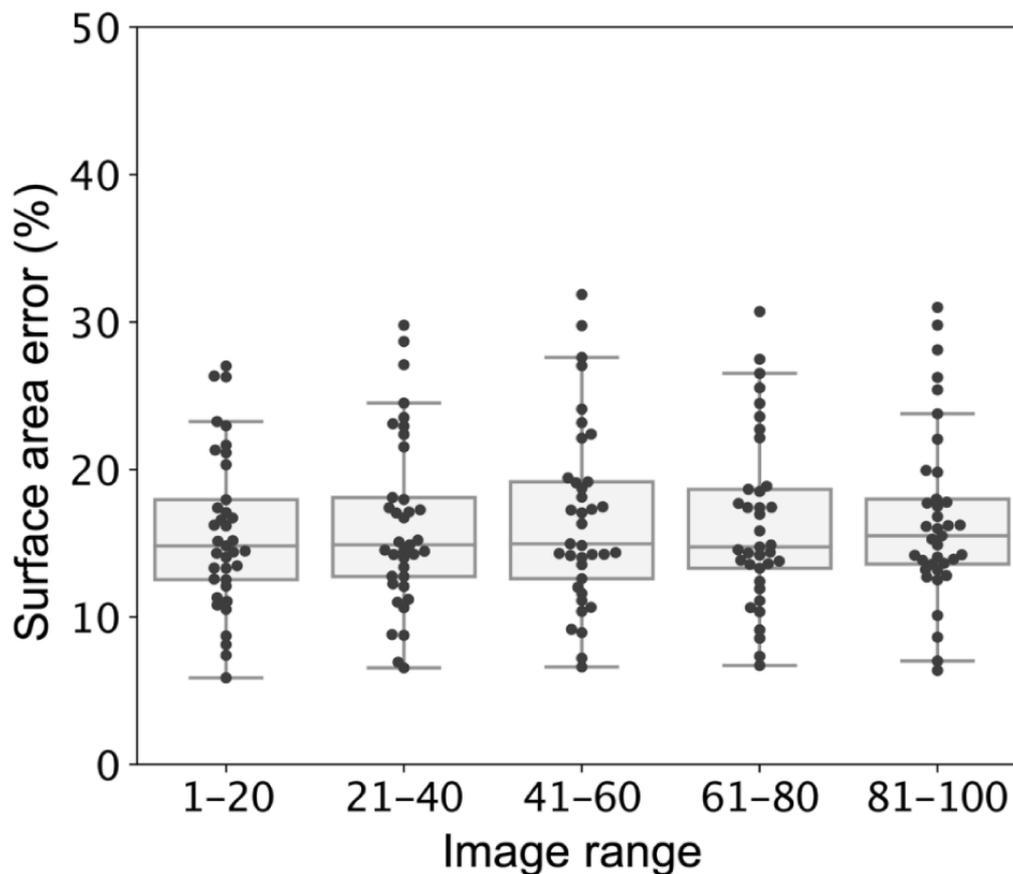


### Performance Over Time

The effect of training on performance was examined by tracking the performance of all raters over the first 100 affected skin photos that they marked for which the ground truth was never

shown. A total of 37 raters met this criterion. The mean performance over time for each eligible rater is shown in Figure 8, with no increase in performance seen over the first 100 images marked. Similar results on minimal training effects have been reported in other crowdsource studies [17].

**Figure 8.** Surface area error of individual raters in successive groups of affected images. Error is displayed for all 37 raters who marked at least 100 affected images. Each point is the mean error for a given user for the group of 20 photos in the stated image range.



## Discussion

### Principal Findings

A crowd of nonexpert raters was able to achieve good overall performance for segmenting cGVHD-affected skin with minimal training. The median surface area error was less than 12% for all crowds. The low feedback  $r=17$  crowd performed slightly poorer than the high feedback crowd, with a 2% to 3% increase in surface area error. Recruiting more raters to the low feedback group for a larger crowd ( $r \geq 32$ ) did not improve performance relative to the original low feedback crowd ( $r=17$ ). However, tracking the top 5 most reliable raters from the low feedback group for each image was able to recover almost identical performance to the high feedback crowd. We believe this is due to individuals within the crowd likely having different skill levels for the assigned task, as has been noted in similar crowdsourcing studies [22]. We therefore recommend tracking rater performance to ensure the most reliable individuals contribute to the consensus demarcations. This optimal strategy will yield the best crowd performance while lowering the required number of expert demarcations for training.

High variability between different angles of the same skin area were observed, raising concerns for the reliability of the consensus demarcation for any given image with an unknown ground truth. Higher variability between raters for a given image was not found to correlate with lower performance of the crowd consensus demarcation and cannot therefore be used as a

measure of reliability. Finally, no significant learning was observed during the task as more photos and feedback were seen.

### Limitations

A limitation of our study was the lack of diversity in skin types. Our cohort was dominated by lighter skin tones; 32 patients had Fitzpatrick skin type III or lower, while only 4 patients had skin types IV and higher. Despite the good overall performance of the crowd, further study is needed to disentangle the possible sources of disagreement that were observed and develop methods to mitigate these effects.

### Future Work

We have demonstrated the potential utility of employing a crowd of nonexpert raters for demarcating cGVHD-affected skin in patient photos. Next steps should explore if this performance is maintained when applied to standard clinical photos, where lighting conditions, imaging distances, and photo quality may be more variable than the more standardized set used here. In addition, further methods of estimating the reliability of the consensus demarcations should be explored to provide more robust quality assurance and filter out high error outliers. Future studies using more extensive training techniques and a unified interface for crowd and expert demarcations will also be important for establishing nonexperts' understanding of the complex task and methods to minimize potential sources of variability. Accurately recognizing active disease is a major

clinical concern, with even expert dermatologists commonly disagreeing on whether a skin area is erythematous or hyperpigmented [24]. New technologies, such as hyperspectral imaging, have shown promise for addressing this limitation in clinical practice [25]. The ability of the crowd to differentiate between disease types should also be explored in future studies given the good overall performance for recognizing cGVHD-affected skin.

Ultimately, the application of crowdsourcing could offer a scalable solution for labelling large sets of images for the training of automated AI algorithms. The effect of training with a larger volume of lower quality demarcations versus a smaller

number of expert demarcations, as reported previously [9], also warrants future investigation.

## Conclusion

We have shown that a crowd of nonexpert raters is capable of delineating surface areas of cGVHD-affected skin (9%-11% surface area error) better than the current clinical standard (19-22% [7]). Crowd demarcation therefore offers a practical solution for accurately demarcating large databases of patient photos, which is a crucial unmet need for training reliable AI image analysis methods. Such methods could provide a clinically meaningful improvement to the current standard given that body surface area has been shown to be a better predictor of mortality than the NIH cGVHD Skin Score [8].

## Acknowledgments

This work was supported by a Career Development Award to ERT from the United States Department of Veterans Affairs Clinical Sciences R&D Service (grant IK2 CX001785), a Discovery Research Grant from Vanderbilt University, the Vanderbilt University Medical Center Departments of Medicine and Dermatology, and the National Institutes of Health (grants K12 CA090625 and R21 AR074589). In addition, the research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (grants TL1TR002244 and UL1 TR002243). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. The authors thank Fuyao Chen for her assistance with data acquisition. We also gratefully acknowledge Canfield Scientific for their technical support.

## Conflicts of Interest

EPD is the chief executive officer of Centaur Labs and holds shares in the company. KM is an employee of Centaur Labs.

## References

1. Socié G, Ritz J. Current issues in chronic graft-versus-host disease. *Blood*. 2014 Jul 17;124(3):374-384 [FREE Full text] [doi: [10.1182/blood-2014-01-514752](https://doi.org/10.1182/blood-2014-01-514752)] [Medline: [24914139](https://pubmed.ncbi.nlm.nih.gov/24914139/)]
2. Rodgers CJ, Burge S, Scarisbrick J, Peniket A. More than skin deep? Emerging therapies for chronic cutaneous GVHD. *Bone Marrow Transplant*. 2013 Mar;48(3):323-337 [doi: [10.1038/bmt.2012.96](https://doi.org/10.1038/bmt.2012.96)] [Medline: [22863725](https://pubmed.ncbi.nlm.nih.gov/22863725/)]
3. Jagasia MH, Greinix HT, Arora M, Williams KM, Wolff D, Cowen EW, 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 Mar;21(3):389-401.e1 [FREE Full text] [doi: [10.1016/j.bbmt.2014.12.001](https://doi.org/10.1016/j.bbmt.2014.12.001)] [Medline: [25529383](https://pubmed.ncbi.nlm.nih.gov/25529383/)]
4. Lee SJ, Wolff D, Kitko C, Koreth J, Inamoto Y, Jagasia M, 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. The 2014 Response Criteria Working Group report. *Biol Blood Marrow Transplant*. 2015 Jun;21(6):984-999 [FREE Full text] [doi: [10.1016/j.bbmt.2015.02.025](https://doi.org/10.1016/j.bbmt.2015.02.025)] [Medline: [25796139](https://pubmed.ncbi.nlm.nih.gov/25796139/)]
5. Miklos D, Cutler CS, Arora M, Waller EK, Jagasia M, Pusic I, et al. Ibrutinib for chronic graft-versus-host disease after failure of prior therapy. *Blood*. 2017 Nov 23;130(21):2243-2250 [FREE Full text] [doi: [10.1182/blood-2017-07-793786](https://doi.org/10.1182/blood-2017-07-793786)] [Medline: [28924018](https://pubmed.ncbi.nlm.nih.gov/28924018/)]
6. Pham C, Collier Z, Gillenwater J. Changing the way we think about burn size estimation. *J Burn Care Res*. 2019 Jan 01;40(1):1-11 [doi: [10.1093/jbcr/iry050](https://doi.org/10.1093/jbcr/iry050)] [Medline: [30247559](https://pubmed.ncbi.nlm.nih.gov/30247559/)]
7. Mitchell SA, Jacobsohn D, Thormann Powers KE, Carpenter PA, Flowers MED, Cowen EW, et al. A multicenter pilot evaluation of the National Institutes of Health chronic graft-versus-host disease (cGVHD) therapeutic response measures: feasibility, interrater reliability, and minimum detectable change. *Biol Blood Marrow Transplant*. 2011 Nov;17(11):1619-1629 [FREE Full text] [doi: [10.1016/j.bbmt.2011.04.002](https://doi.org/10.1016/j.bbmt.2011.04.002)] [Medline: [21536143](https://pubmed.ncbi.nlm.nih.gov/21536143/)]
8. Baumrin E, Baker LX, Byrne M, Martin PJ, Flowers ME, Onstad L, et al. Prognostic value of cutaneous disease severity estimates on survival outcomes in patients with chronic graft-vs-host disease. *JAMA Dermatol*. 2023 Apr 01;159(4):393-402 [doi: [10.1001/jamadermatol.2022.6624](https://doi.org/10.1001/jamadermatol.2022.6624)] [Medline: [36884224](https://pubmed.ncbi.nlm.nih.gov/36884224/)]
9. McNeil A, Parks K, Liu X, Saknite I, Chen F, Reasat T, et al. Artificial intelligence recognition of cutaneous chronic graft-versus-host disease by a deep learning neural network. *Br J Haematol*. 2022 Jun;197(6):e69-e72 [FREE Full text] [doi: [10.1111/bjh.18141](https://doi.org/10.1111/bjh.18141)] [Medline: [35322873](https://pubmed.ncbi.nlm.nih.gov/35322873/)]
10. Ørting SN, Doyle A, Van Hilten A, Hirth M, Inel O, Madan CR, et al. A survey of crowdsourcing in medical image analysis. *Human Computation*. 2020;7:1-26 [FREE Full text] [doi: [10.15346/hc.v7i1.1](https://doi.org/10.15346/hc.v7i1.1)]

11. Petrović N, Moyà-Alcover G, Varona J, Jaume-I-Capó A. Crowdsourcing human-based computation for medical image analysis: a systematic literature review. *Health Informatics J*. 2020 Dec;26(4):2446-2469 [FREE Full text] [doi: [10.1177/1460458220907435](https://doi.org/10.1177/1460458220907435)] [Medline: [32141371](https://pubmed.ncbi.nlm.nih.gov/32141371/)]
12. Good BM, Su AI. Crowdsourcing for bioinformatics. *Bioinformatics*. 2013 Aug 15;29(16):1925-1933 [FREE Full text] [doi: [10.1093/bioinformatics/btt333](https://doi.org/10.1093/bioinformatics/btt333)] [Medline: [23782614](https://pubmed.ncbi.nlm.nih.gov/23782614/)]
13. Grote A, Schaadt NS, Forestier G, Wemmert C, Feuerhake F. Crowdsourcing of histological image labeling and object delineation by medical students. *IEEE Trans Med Imaging*. 2019 May;38(5):1284-1294 [doi: [10.1109/TMI.2018.2883237](https://doi.org/10.1109/TMI.2018.2883237)] [Medline: [30489264](https://pubmed.ncbi.nlm.nih.gov/30489264/)]
14. Law J, Paulson T, Sanchez C, Galipeau P, Jansen M, Stachler M, et al. Wisdom of the crowd for early detection in Barrett's esophagus. 2021 Presented at: IEEE 18th International Symposium on Biomedical Imaging; April 13-16, 2021; Nice [doi: [10.1109/isbi48211.2021.9433763](https://doi.org/10.1109/isbi48211.2021.9433763)]
15. Schlesinger D, Jug F, Myers G, Rother C, Kainmueller D. Crowd sourcing image segmentation with iaSTAPLE. 2017 Presented at: IEEE 14th International Symposium on Biomedical Imaging; April 18-21, 2017; Melbourne [doi: [10.1109/isbi.2017.7950547](https://doi.org/10.1109/isbi.2017.7950547)]
16. Heim E, Roß T, Seitel A, März K, Stieltjes B, Eisenmann M, et al. Large-scale medical image annotation with crowd-powered algorithms. *J Med Imaging (Bellingham)*. 2018 Jul;5(3):034002 [FREE Full text] [doi: [10.1117/1.JMI.5.3.034002](https://doi.org/10.1117/1.JMI.5.3.034002)] [Medline: [30840724](https://pubmed.ncbi.nlm.nih.gov/30840724/)]
17. O'Neil A, Murchison J, van BE, Goatman K. Crowdsourcing labels for pathological patterns in CT lung scans: can non-experts contribute expert-quality ground truth? In: *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Cham. Springer International Publishing; 2017 Presented at: LABELS 2017, STENT 2017, CVII 2017; September 10-14; Quebec City p. 96-105 [doi: [10.1007/978-3-319-67534-3\\_11](https://doi.org/10.1007/978-3-319-67534-3_11)]
18. Bhattar P, Frisch E, Duhaime E, Jain A, Fischetti C. Diabetic retinopathy detection using collective intelligence. *J Sci Innov Med*. 2020;3(1):1 [doi: [10.29024/jsim.47](https://doi.org/10.29024/jsim.47)]
19. Kentley J, Weber J, Liopyris K, Braun RP, Marghoob AA, Quigley EA, et al. Agreement between experts and an untrained crowd for identifying dermoscopic features using a gamified app: reader feasibility study. *JMIR Med Inform*. 2023 Jan 18;11:e38412 [FREE Full text] [doi: [10.2196/38412](https://doi.org/10.2196/38412)] [Medline: [36652282](https://pubmed.ncbi.nlm.nih.gov/36652282/)]
20. Wang J, Chen F, Dellalana L, Jagasia M, Tkaczyk E, Dawant B. Segmentation of skin lesions in chronic graft versus host disease photographs with fully convolutional networks. *Medical Imaging 2018. Computer-Aided Diagnosis International Society for Optics and Photonics*; 2018 Presented at: SPIE Medical Imaging; February 10-15, 2018; Houston p. 160-166 URL: <https://doi.org/10.1117/12.2293334> [doi: [10.1117/12.2293334](https://doi.org/10.1117/12.2293334)]
21. Parks K, Liu X, Reasat T, Khera Z, Baker LX, Chen H, et al. Non-expert markings of active chronic graft-versus-host disease photographs: optimal metrics of training effects. *J Digit Imaging*. 2023 Mar;36(1):373-378 [doi: [10.1007/s10278-022-00730-8](https://doi.org/10.1007/s10278-022-00730-8)] [Medline: [36344635](https://pubmed.ncbi.nlm.nih.gov/36344635/)]
22. Duhaime E, Jin M, Moulton T, Weber J, Kurtansky N, Halpern A, et al. Nonexpert crowds outperform expert individuals in diagnostic accuracy on a skin lesion diagnosis task. 2023 Presented at: IEEE 20th International Symposium on Biomedical Imaging; April 18-21, 2023; Cartagena de Indias [doi: [10.1109/isbi53787.2023.10230646](https://doi.org/10.1109/isbi53787.2023.10230646)]
23. Zijdenbos A, Dawant B, Margolin R, Palmer A. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging*. 1994;13(4):716-724 [doi: [10.1109/42.363096](https://doi.org/10.1109/42.363096)] [Medline: [18218550](https://pubmed.ncbi.nlm.nih.gov/18218550/)]
24. Tkaczyk ER, Coco JR, Wang J, Chen F, Ye C, Jagasia MH, et al. Crowdsourcing to delineate skin affected by chronic graft-vs-host disease. *Skin Res Technol*. 2019 Jul;25(4):572-577 [FREE Full text] [doi: [10.1111/srt.12688](https://doi.org/10.1111/srt.12688)] [Medline: [30786065](https://pubmed.ncbi.nlm.nih.gov/30786065/)]
25. Saknite I, Kwun S, Zhang K, Hood A, Chen F, Kangas L, et al. Hyperspectral imaging to accurately segment skin erythema and hyperpigmentation in cutaneous chronic graft-versus-host disease. *J Biophotonics*. 2023 Jul;16(7):e202300009 [doi: [10.1002/jbio.202300009](https://doi.org/10.1002/jbio.202300009)] [Medline: [36942511](https://pubmed.ncbi.nlm.nih.gov/36942511/)]

---

## Abbreviations

- AI:** artificial intelligence  
**cGVHD:** chronic graft-versus-host disease  
**NIH:** National Institutes of Health  
**r:** rater
-

*Edited by J Lipoff; submitted 01.05.23; peer-reviewed by E Baumrin, H Mondal, R Wei; comments to author 07.08.23; revised version received 02.10.23; accepted 24.10.23; published 26.12.23*

*Please cite as:*

*McNeil AJ, Parks K, Liu X, Jiang B, Coco J, McCool K, Fabbri D, Duhaime EP, Dawant BM, Tkaczyk ER*

*Crowdsourcing Skin Demarcations of Chronic Graft-Versus-Host Disease in Patient Photographs: Training Versus Performance Study*

*JMIR Dermatol 2023;6:e48589*

*URL: <https://derma.jmir.org/2023/1/e48589>*

*doi: [10.2196/48589](https://doi.org/10.2196/48589)*

*PMID: [38147369](https://pubmed.ncbi.nlm.nih.gov/38147369/)*

©Andrew J McNeil, Kelsey Parks, Xiaoqi Liu, Bohan Jiang, Joseph Coco, Kira McCool, Daniel Fabbri, Erik P Duhaime, Benoit M Dawant, Eric R Tkaczyk. Originally published in JMIR Dermatology (<http://derma.jmir.org>), 26.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Dermatology, is properly cited. The complete bibliographic information, a link to the original publication on <http://derma.jmir.org>, as well as this copyright and license information must be included.