Published in final edited form as:
Bone Marrow Transplant. 2018 October ; 53(10): 1356–1358. doi:10.1038/s41409-018-0211-8.

Overcoming human disagreement assessing erythematous lesion severity on 3D photos of chronic graft-versus-host disease

Eric R. Tkaczyk, MD, PhD[1,2,3], **Fuyao Chen, BE**[1,2], **Jianing Wang, MS**[4], **Jocelyn S. Gandelman, BS**[5], **Inga Saknite, PhD**[1], **Laura E. Dellalana, BS**[1], **Madan H. Jagasia, MD, MS, MMHC**[5], **Benoit M. Dawant, PhD**[2,4]

[1.]Department of Dermatology, Vanderbilt University Medical Center, Nashville, TN, USA

[2.]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, USA

[3.]Dermatology Service, Department of Veterans Affairs Tennessee Valley Healthcare System, Nashville, TN, USA

[4.]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA

[5.]Division of Hematology and Oncology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Accurate estimation of extent and severity of erythema is essential for assessing disease severity and treatment response in many diseases, with particular impact on patients with chronic-graft-versus host disease (cGVHD). In cGVHD, reversal of erythema is associated with improved survival[1]. Estimation of erythema surface area and conversion to a NIH 2014 skin score is the current standard of care in cGVHD[2]. However, trained clinicians can detect only a minimum change of 19% to 22% body surface area,[3] an obstacle in proving treatment efficacy necessary for the approval of novel therapeutic agents in clinical trials and daily clinical care of patients with cGVHD[4].

Erythema has been successfully assessed using 2D photography in the case of psoriasis[5], however a key difference between psoriasis and cutaneous cGVHD is that the erythema of psoriasis is well defined. Kohli *et al.*[6] demonstrated 3D-photography as a reproducible method to measure vitiligo. 3D photography has the strength that 3D images can more accurately measure body surface area as 2D images may underestimate involvement of skin areas that fall beyond a contour in the image[6,7]. In this report, we explore 3D-photography to monitor cGVHD erythema.

Corresponding author: Eric Tkaczyk, MD, PhD, Department of Dermatology, Vanderbilt University Medical Center, 719 Thompson Lane, One Hundred Oaks Suite 26300, Nashville, TN 37204 (eric.tkaczyk@vanderbilt.edu, 615-936-4633).

Conflict of interest disclosures: None

[4.]Competing Interests
There is no competing financial interests in relation to the work described.

We selected a complex case of cGVHD erythema with areas of superimposed sclerosis to test whether 3D photography could be used to track skin severity in cGVHD. The patient was a middle-aged man who developed cGVHD (skin, lung, eyes, mouth) five years after stem cell transplant from a matched unrelated donor for acute myelogenous leukemia. He was maintained on prednisone, tacrolimus, and methotrexate, and had failed treatment with ibrutinib, ruxolitinib, and extracorporeal photopheresis. Examination revealed widespread erythematous patches and sclerotic plaques with focal ulcerations over his trunk, extremities, head and neck. Informed consent was obtained for an IRB-approved imaging study, and a handheld stereoscopic camera (Vectra H1, Canfield Scientific, NJ, USA) captured cross-polarized 3D images of his upper back under standardized distance, lighting during clinic visits. After two months, methotrexate was temporarily held after a minor surgery, and he rapidly worsened. His condition stabilized with pentostatin, which was discontinued due to GI side effects, followed by dramatic worsening and transition to comfort care.

Fully-rotatable 3D photos were cropped and registered to identical stereoscopic views and independently analyzed in unlabeled, random order by the director (ET) and five members of the Vanderbilt Cutaneous Imaging Clinic. All were trained to recognize erythematous lesions by ET, a board-certified dermatologist with interest in cGVHD. The raters were instructed to only count redness and to ignore any hyperpigmentation, hypopigmentation or skin changes due to sclerosis without overlying erythema. Teaching took place using demarcations of erythema on other patients with ET checking these initial demarcations and discussing if he agreed or disagreed with the demarcation. Additionally, clinic members observed how ET demarcated erythema on independent patients using the software or in clinic. Both experience with cGVHD and the amount of time with ET varied between observers.

Erythematous skin areas were demarcated with the selection tools available in the Vectra 3D Visualization, Analysis, Measurement software. Then the software automatically calculated each demarcation's surface area (cm^2) and average red/green coordinate (a*) in the CIELAB color space (Fig. 1). Examples of 3D images of the quantitative photos produced by raters' demarcation of skin redness are shown in 2D space in Figure 1 (Rater 2 and 3). Total redness was calculated as the product of a* and area. Additionally, a single en face 2D image was exported from each 3D photo for processing in Matlab. Each image was segmented into approximately 1200 superpixels of similar size (20×20 pixels). The a* value of each superpixel in the entire image set was plotted as a histogram representing the total patient-specific a* distribution. From this distribution, 8 evenly spaced thresholds were picked to uniformly span the 60th to 95th percentiles of a*. Correspondingly, 8 different algorithmic demarcations of lesional skin were selected for each image, defined as the set of all pixels contained within any superpixels with a* value above the preset threshold.

Intraclass correlation coefficient (ICC), calculated from two-ways ANOVA on log scale, was used to quantify agreement between raters. Difference between human and algorithm image processing were compared by Wilcoxon signed rank test.

Six trained raters significantly disagreed in selection of erythematous area (Fig. 2A). The ICC (0.09) fell within the range of reproducibility previously reported for cGVHD erythema surface area estimation[2]. While humans at times disagreed on whether surface area was increasing or decreasing (Fig. 2A), total redness calculated from their selections revealed a consistent trend (Fig. 2B), with high interrater reliability (ICC=0.85). Total redness further remained stable during treatment and worsened off treatment (Fig. 2C). Notably, human estimation of total redness did not differ from image processing derived total redness (p=0.81, Fig. 2C).

Our results contrast with prior findings in vitiligo and psoriasis[5,6] and underscore the difficulty of achieving consensus on extent of ill-defined lesions and scoring cutaneous cGVHD. Poor agreement on erythema stems not only from human inability to estimate surface areas, but from differing perception of borders. Humans disagree not simply because they measure poorly, but because they *see* differently. Differences in demarcations by the human raters reflects their personal thresholds of intensity of redness that constitutes erythema (Fig 1). For instance, rater 3 selected only areas with a high a* (average 14.23 on the final day of follow-up), whereas rater 2 delineated wider and less intensely red areas, resulting in a lower average a* (7.02 on the final day). Accordingly, when the intensity of erythema is taken into account, observers' disagreement can be overcome. By the total redness metric presented, threshold-based image processing independently recapitulated human evaluation of erythema (Fig 2C).

This is a single case report presenting proof of concept of 3D photography and total redness as a potential tool for monitoring cGVHD erythema. Because our patient was Caucasian, a notable limitation of our findings is potential difficulty applying this method to track erythema in darker skin types, whose unaffected skin will have higher a* values. Ultimately, multiple techniques such as the algorithm development to find subtle skin changes, skin biomechanical assessment, and cellular level imaging may help with comprehensive assessment of cutaneous disease activity. Another important future direction is to study how superimposed sclerosis, scale and ulceration affects the machine-based interpretation and human rating of erythema. Further study in additional patients is warranted to determine whether automated, threshold-based image processing for total redness could be developed as a practical tool to quantitatively track clinical course of cGVHD.

Acknowledgements

Dr. Tkaczyk is grateful for support from NIH K12 CA090625.

Funding/Support: NIH K12 CA090625

5. References

1. Curtis LM, Grkovic L, Mitchell SA, Steinberg SM, Cowen EW, Datiles MB, et al. NIH response criteria measures are associated with important parameters of disease severity in patients with chronic GVHD. Bone Marrow Transplantation. 2014;49(12):1513–20. [PubMed: 25153693]
2. 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. Biology of Blood and Marrow Transplantation. 2015;21(6):984–99. [PubMed: 25796139]

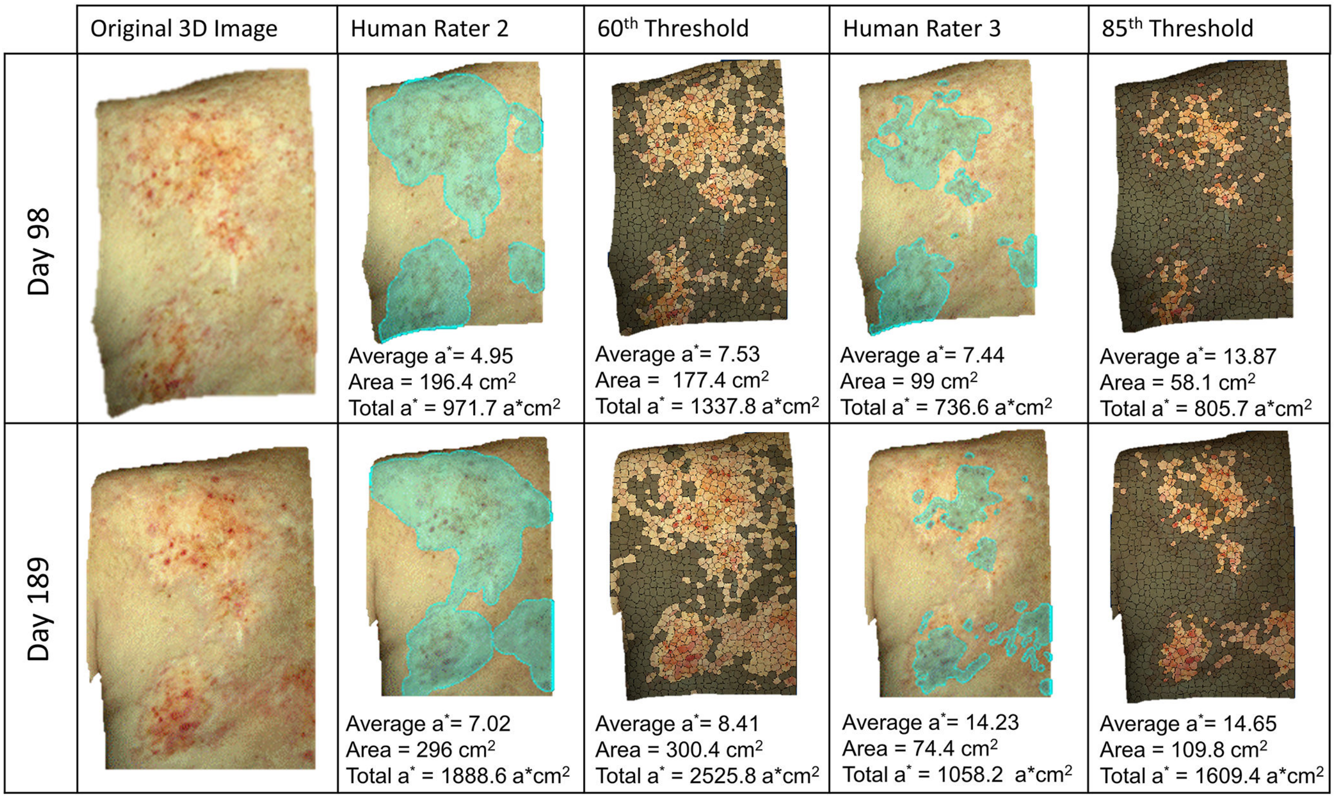
3. Mitchell SA, Jacobsohn D, Powers KET, Carpenter PA, Flowers ME, Cowen EW, et al. A Multicenter Pilot Evaluation of the National Institutes of Health Chronic Graft-versus-Host Disease (cGVHD) Therapeutic Response Measures: Feasibility, Interrater Reliability, and Minimum Detectable Change. Biology of Blood and Marrow Transplantation. 2011;17(11):1619–29. [PubMed: 21536143]

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. Biology of Blood and Marrow Transplantation. 2015;21(6):984–99. [PubMed: 25796139]

5. Raina A, Hennessy R, Rains M, Allred J, Diven D, Markey MK. Objective measurement of erythema in psoriasis using digital color photography with color calibration. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2014;

6. Kohli I, Isedeh P, Al-Jamal M, DaSilva D, Baston A, Canfield D, et al. Three-dimensional imaging of vitiligo. Experimental Dermatology. 2015;24(11):879–880. [PubMed: 26119511]

7. Van Geel N, Vander Haeghen Y, Ongenae K, Naeyaert JM. A new digital image analysis system useful for surface assessment of vitiligo lesions in transplantation studies. European Journal of Dermatology. 2004;14(3):150–155. [PubMed: 15246939]



	Original 3D Image	Human Rater 2	60th Threshold	Human Rater 3	85th Threshold
Day 98		Average a*= 4.95 Area = 196.4 cm^2 Total a* = 971.7 a*cm^2	Average a*= 7.53 Area = 177.4 cm^2 Total a* = 1337.8 a*cm^2	Average a*= 7.44 Area = 99 cm^2 Total a* = 736.6 a*cm^2	Average a*= 13.87 Area = 58.1 cm^2 Total a* = 805.7 a*cm^2
Day 189		Average a*= 7.02 Area = 296 cm^2 Total a* = 1888.6 a*cm^2	Average a*= 8.41 Area = 300.4 cm^2 Total a* = 2525.8 a*cm^2	Average a*= 14.23 Area = 74.4 cm^2 Total a* = 1058.2 a*cm^2	Average a*= 14.65 Area = 109.8 cm^2 Total a* = 1609.4 a*cm^2

Figure 1.

Example of 3D images, rater selections of erythematous area, and estimation of lesion produced by image processing, with corresponding average a*, area, and total redness (total a*).

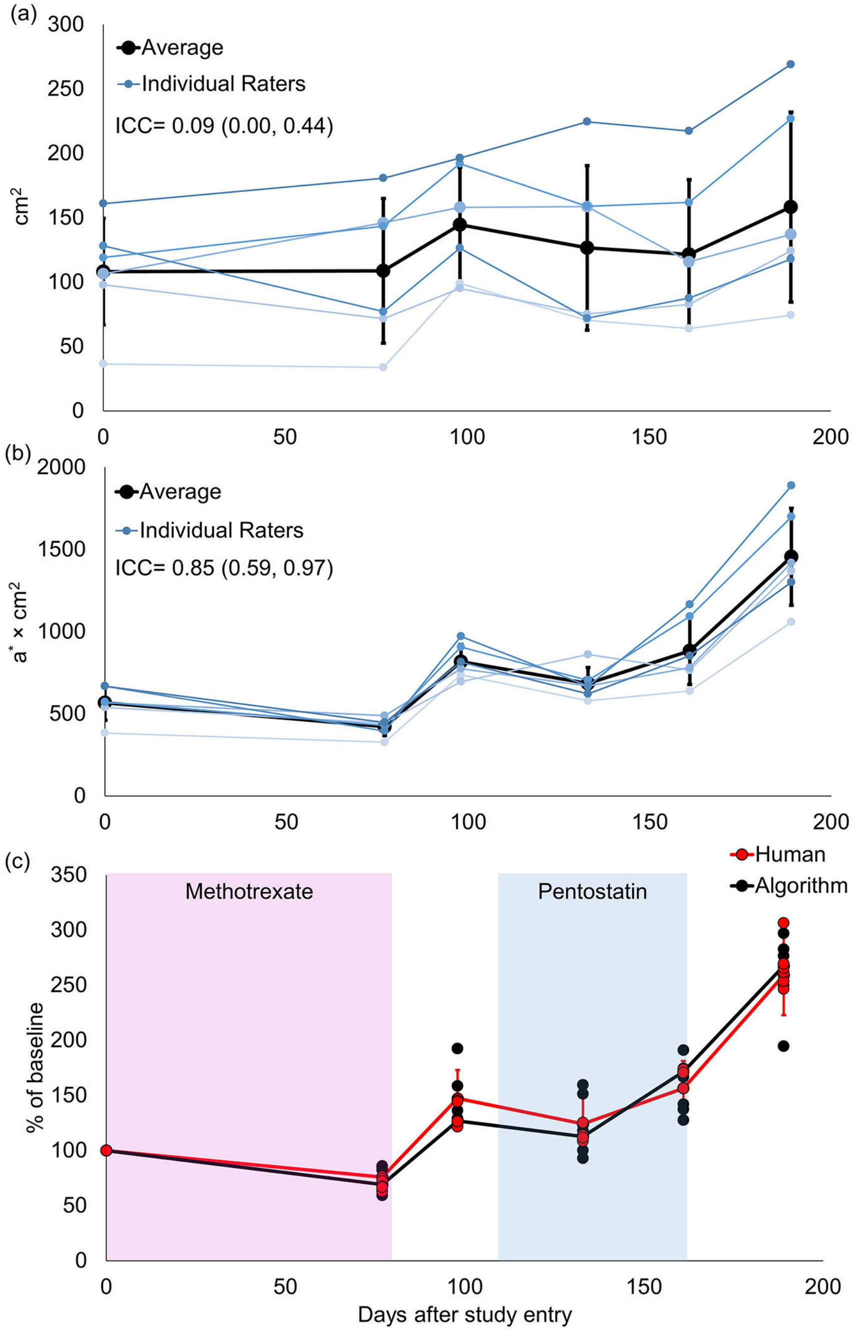


Figure 2.

a) Time evolution of size of erythematous area selected by six humans exhibited poor reproducibility (ICC=0.09, range: 0.00, 0.44). b) Total redness from the selections exhibited high reproducibility (ICC=0.85, range: 0.59, 0.97). c) Total redness normalized to day 0 (baseline) is shown for both individual humans (red dots) and algorithm with different thresholds (black dots – a* thresholds of 2.0, 2.9, 3.9, 5.0, 6.3, 8.0, 10.4, 14.6, corresponding to the 60th, 65th, 70th, 75th, 80th, 85th, 90th, 95th percentiles, respectively). No statistically significant difference is present between the average human (red line) and algorithm (black line) data (p=0.81). Treatments are shown by the colored boxes. Error bars in all panels represent standard deviations.