Comparison of Autophotomontage Software Programs in Eyes with CMV Retinitis

Jenny Chen,[1,2] Somsanguan Ausayakhun,[3] Sakarin Ausayakhun,[3] Choeng Jirawison,[4] Claire M. Khouri,[1] Travis C. Porco,[1,2] David Heiden,[5] Jeremy D. Keenan,[1,2] and Todd P. Margolis[1,2]

PURPOSE. Automated mosaic software programs are used to stitch together overlapping retinal fundus photographs. The performance of these programs in eyes with retinal diseases has not been independently evaluated. This study compares the quality of the mosaic products of three autophotomontage software programs, using digital fundus photographs of eyes with cytomegalovirus (CMV) retinitis.

METHODS. Photographs of 99 eyes with CMV retinitis of 94 patients with HIV were taken at Maharaj Nakorn Chiang Mai Hospital in Chiang Mai, Thailand. Automated mosaic images were created for each of the 99 eyes by three different commercially available programs: IMAGEnet (Topcon, Oakland, NJ), i2k Retina (DualAlign LLC, Clifton Park, NY), and AutoMontage (OIS, Sacramento, CA). Three masked graders ranked each set of mosaics for each eye. The graders also assessed the overall image quality and documented mosaic artifacts in each image.

RESULTS. i2k Retina was ranked as the best program (70%–88%) more often than AutoMontage (10%–33%, $P < 0.001$) or IMAGEnet (0%–4%, $P < 0.001$) for creating automontages from digital fundus photographs of eyes with CMV retinitis. Acceptable quality mosaic images were reported most commonly for i2k Retina (93%–94%) and AutoMontage (91%–95%), followed by IMAGEnet (27%–56%, $P < 0.001$). IMAGEnet had a significantly higher percentage of mosaic errors than did either i2k Retina or AutoMontage ($P < 0.001$).

CONCLUSIONS. In eyes with CMV retinitis, both the i2k Retina and AutoMontage software packages appear to create higher quality mosaics than does IMAGEnet. Automated retinal mosaic imaging may be valuable in diagnosing CMV retinitis and observing disease progression. (*Invest Ophthalmol Vis Sci.* 2011; 52:9339–9344) DOI:10.1167/iovs.11-8322

Automated mosaic software programs stitch together multiple fundus photographs for the evaluation of retinal diseases. These programs use multiple overlapping digital fundus photographs to construct a single composite image, easing the review of these photographs.[1,2] Programs with automated mosaic functions may use landmarks in the normal retina, such as the optic disc and the vasculature, to assist in the creation of a photomontage.[3,4] Although many mosaic programs are in use, their performance in images with retinal pathology has not been independently evaluated.

CMV retinitis is an opportunistic infection affecting severely immunocompromised patients, including many patients with HIV/AIDS.[5] Although the incidence of CMV retinitis has declined in industrialized countries due to the widespread availability of antiretroviral medications,[6] it remains a significant cause of blindness in many regions of the world affected by HIV/AIDS.[7–9] In Southeast Asia, CMV retinitis is still present in up to one third of HIV/AIDS patients with CD4 counts of less than 50 cells/μL.[7,10] CMV retinitis has a characteristic appearance[7] and is a new and important target for telemedicine screening in developing nations.[11]

An accurate photomontage of eyes with CMV retinitis can be helpful in determining the location, extent, and progression of disease, and may have a role in the remote diagnosis and treatment of CMV. We recently performed a study on telemedicine diagnosis of CMV retinitis at an ocular infectious diseases clinic in northern Thailand. As an ancillary study, we evaluated three commercially available autophotomontage software programs using digital fundus photographs of eyes with CMV retinitis.

METHODS

The study protocol was in compliance with the Declaration of Helsinki, and institutional review board approval was prospectively obtained at the University of California, San Francisco, and Chiang Mai University.

This study was conducted using fundus photographs acquired from 94 patients with HIV at the Ocular Infectious Diseases Clinic at Maharaj Nakorn Chiang Mai Hospital in Chiang Mai, Thailand from August 2008 through April 2009. All patients presenting for their initial visit or diagnosed with CMV retinitis in the previous month were offered enrollment in the study. The image capture protocol was analogous to the fundus photography protocol used in the Longitudinal Study of the Ocular Complications of AIDS.[12] A trained ophthalmic photographer

used a digital fundus camera (TRC-NW 6S; Topcon, Oakland, NJ) to capture overlapping 45° fundus photographs in each eye: one central image of the optic disc and macula, with eight or more surrounding midperipheral images. Each eye was photographed only once. A mean of nine fundus photographs were taken for each eye (range, 8–15).

In this ancillary study, we included 89 eyes with a definite clinical diagnosis of CMV retinitis, 5 eyes with an uncertain diagnosis of CMV, and 5 eyes with a diagnosis from photographic review only. One of the authors (JC) loaded all photographs for each study eye onto each of three commercially available automontage programs: OIS AutoMontage (OIS, Sacramento, CA), i2k Retina (DualAlign LLC, Clifton Park, NY), and IMAGEnet Professional (Topcon, Oakland, NJ). The authors used the factory default settings for each program, without specialized image registration settings. Although we did not use any of the advanced options in any of the programs, it should be noted that all programs have the capability to accommodate as many individual photographs as necessary; all have the option to view images as blended or unblended mosaics, with advanced blending options available in i2k Retina and OIS AutoMontage, and IMAGEnet allows for manual construction of a mosaic. One of the authors (JC) created all the mosaics, documenting whether at least two individual photographs were stitched together and also whether all the individual photographs in the set had been incorporated into the mosaic. The completed mosaics were presented to three expert graders (DH, TPM, and JDK) for two rounds of grading. First, the mosaics were presented as randomly sorted sets of three mosaics per study eye (one mosaic from each software program), to facilitate a side-by-side comparison. Second, the mosaics were presented in random order (without stratifying by study eye), with each mosaic paired with the individual central photograph used in constructing the mosaic. The graders were masked to the identity of the software.

In the first round of grading (i.e., side-by-side comparison of the three fundus montages from the same study eye), the three graders assigned each image a rank based on personal preference (1 , best; 3, worst). Images that could not be stitched into a mosaic by one program, but could be by another, were automatically given the lowest ranking. In the second round of grading, the same three graders assessed whether each mosaic was gradable or ungradable. Images that were considered to be ungradable were not assessed further. For gradable mosaics, images were assessed for the presence or absence of the following five photographic features: duplicated vessels (duplication of the same blood vessel from a common origin), breakage in vessel continuity (sudden sharp disruption of a single, nonduplicated blood vessel that is not due to image misplacement), misplaced images (placement of one or more of the individual photographs in an incorrect sector of the montage), relative blurriness of the mosaic image

(compared to the individual central photograph), and acceptable versus unacceptable overall quality of the mosaic (acceptable mosaics were defined as the presence of few or no mosaic errors and the preference for using the mosaic as opposed to the individual images for clinical care). The graders were also invited to make comments about individual mosaics. They repeated this exercise on a set of 50 randomly selected duplicate mosaics approximately 2 weeks after the initial grading, to assess intrarater reliability. All graders viewed images for the study on a glossy monitor manufactured by Apple, Inc. (Cupertino, CA).

Statistics

We performed descriptive statistics for each software program separately for each grader. The proportion of mosaic images assigned a rank of 1 and the proportion assigned a rank of 3 were computed for each program. To assess whether the ranks differed between the three software programs (using all three raters for each eye), we computed a simple rank centroid statistic, as described in Appendix A. We further examined whether the differences in rank were statistically significant in pairwise comparisons of the three programs.

Intrarater reliability was assessed using Cohen's κ on two separate measurements from each rater (with 95% CI computed using the bootstrap percentile method),[13] except when the estimate was 1, in which case we used the exact method to compute a lower, 97.5%, boundary.[14] Interrater reliability between all three raters was assessed with the Fleiss κ for multiple raters (with 95% CI computed using the bootstrap percentile method).[13] The κ coefficients were interpreted according to the following categories: $\kappa < 0.4$ corresponds to poor agreement, κ from 0.4 to 0.75 corresponds to fair to good agreement, and $\kappa > 0.75$ corresponds to excellent agreement.[15]

We used clustered logistic regression to model the probability of the presence of each feature, accounting for the correlation between images captured from the same person and also the correlation between images graded by the same grader. We used a likelihood ratio to assess whether there was a significant difference in the frequency of each photographic feature between the three programs. When this overall hypothesis test was correct, we further examined pairwise contrasts.

We then assessed whether the presence or absence of each of the photographic features was associated with the programs' ranking. Since each grader also determined the presence or absence of each feature of interest, we calculated the correlation coefficient for each eye–grader combination. Each estimated correlation coefficient summarizes the degree of association between the presence or absence of each feature and the rankings for the three software programs (for that eye and grader). In the absence of any association, the expected
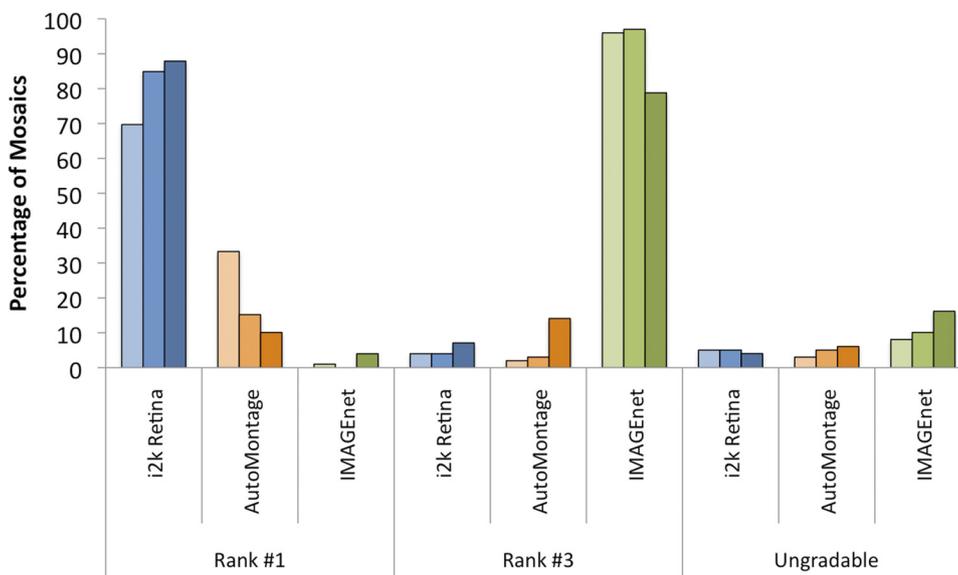

FIGURE 1. Percentage of mosaics ranked 1 and 3 and those deemed ungradable. Results for each grader are shown separately, grouped by program (*left* to *right*, in each dataset the bars represent graders A, B, and C).

TABLE 1. Intra- and Interrater Agreement on the Presence of Mosaic Errors

Photographic Feature	Intrarater Grader A	Intrarater Grader B	Intrarater Grader C	Interrater A, B, & C	Interrater A & B
Duplicated vessels	0.91 (0.78–1)	0.53 (0.24–0.81)	0†	0.29 (0.23–0.36)	0.65 (0.55–0.75)
Vessel breaks	0.78 (0.57–0.98)	0.31 (0–0.65)	0†	0.10 (0.03–0.17)	0.20 (0.07–0.33)
Misplaced images	1 (0.998–1)*	0.81 (0.56–1.0)	0.63 (0.25–1)	0.79 (0.71–0.87)	0.85 (0.77–0.93)
Blurriness	0.91 (0.80–1)	0.75 (0.53–0.98)	0†	0.23 (0.16–0.31)	0.47 (0.34–0.59)
Unacceptable overall quality	0.78 (0.48–1)	0.67 (0.41–0.93)	0.37 (0–0.93)	0.58 (0.45–0.69)	0.56 (0.43–0.68)

Data are the κ statistic (95% CI).

* Conditional exact bottom boundary.

† κ is formally 0 because the grader assigned the same grade to all repeat images; the uncorrected percentage agreements were 98% (duplicated vessels), 98% (vessel breaks), and 96% (blurriness).

average of these correlations is 0. We tested the hypothesis that these were 0 by estimating the intercept in a linear mixed effects regression model with eye and grader as random effects.

To determine whether missing individual images (i.e., the full set of individual images was not incorporated into the mosaic) was predictive of the ranks by each grader, we computed the rank centroid for images with the missing indicator and the rank centroid for those without and computed the distance between these two centroids (see Appendix A for details). Under the null hypothesis, this difference of distances should be 0 on average; the sampling distribution of this difference was computed using the permutation method. We then computed Spearman's rank correlation coefficient between missing individual images and the mean rank of each photo. Statistical analysis was performed with R 2.12 (University of Auckland, Auckland, New Zealand). We accounted for multiple comparisons by setting the significance level at $\alpha = 0.002$, which assumed a Bonferroni correction with a family-wise error rate of $\alpha = 0.05$ and 25 comparisons.

RESULTS

All 99 sets of retinal photographs were processed by each of the three auto-photomontage programs. The programs differed in their ability to create mosaics from the study eyes and also in their ability to incorporate all the individual photographs into the final mosaic. The i2k Retina program incorporated at least two individual images in its mosaics for 97% of the study eyes, AutoMontage for 100% of the study eyes, and IMAGEnet for 92% of the study eyes ($P = 0.01$, χ^2). i2k Retina incorpo-

rated all the individual photographs for 82.8% of the study eyes, AutoMontage for 77.8% of the study eyes, and IMAGEnet for 32.3% of study eyes ($P < 0.001$, χ^2).

As summarized in Figure 1, graders most frequently assigned the number one ranking to photographic mosaics formed by i2k Retina (70%–88%), followed by AutoMontage (10%–33%), and then IMAGEnet (0%–4%). IMAGEnet received the lowest ranking for 79% to 97% of the mosaics. Both i2k Retina (4%–7%) and AutoMontage (2%–14%) were rarely ranked 3. Differences in assigned rank between programs were statistically significant ($P < 0.001$, permutation test, rank centroid).

Photographic mosaics were graded for duplicated vessels, vessel breaks, misplaced images, relative blurriness, and overall quality. Intrarater reliability was higher for graders A and B than for grader C (Table 1). Although grader C had an intrarater κ of 0 for duplicated vessels, breaks in vessel continuity, and relative blurriness, the uncorrected agreement for each feature was ≥96%. The low κ occurred because grader C assigned the same grade to all 50 duplicate images used to assess intrarater reliability. Graders A and B each had good to excellent intrarater reliability for all photographic features except for vessel breaks ($\kappa = 0.31$ for grader B). Interrater reliability among all three graders was poor for duplicated vessels, breaks, and blurriness, largely because grader C had low agreement with the other two graders ($\kappa < 0.2$ for duplicated vessels, vessel breaks, and relative blurriness for reliability between A and C and B and C).
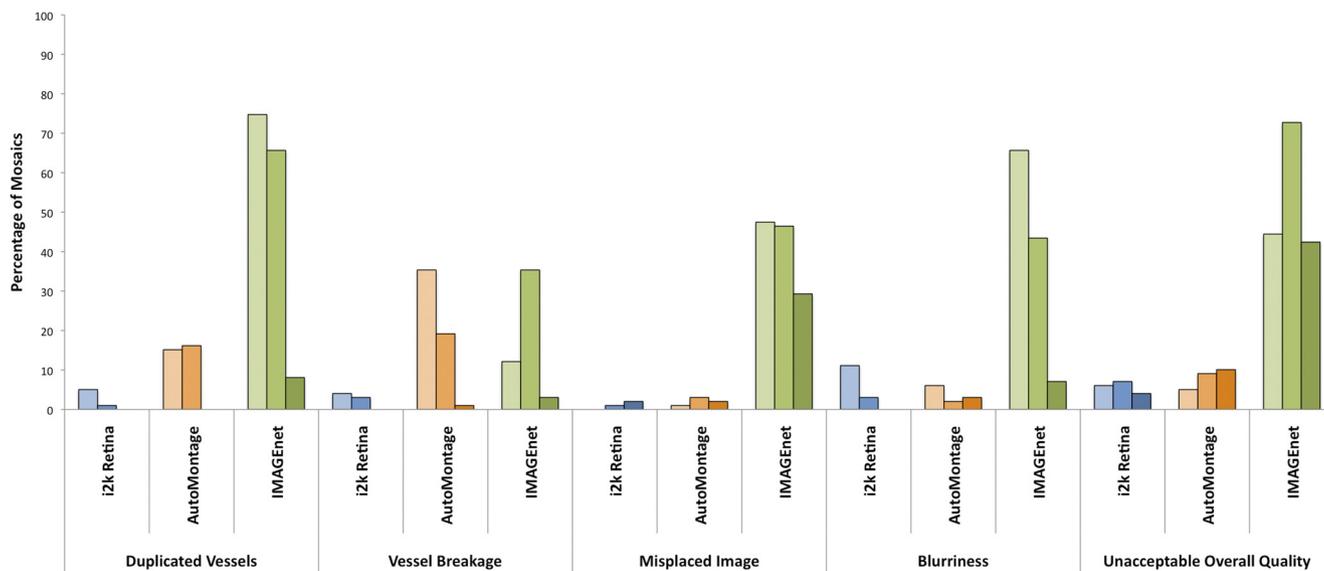

FIGURE 2. Percentage of mosaics with each photographic feature. Results for each grader are shown separately, grouped by program (*left* to *right*, in each dataset the bars represent graders A, B, and C).

There were significant differences between the three software programs for each of the five photographic features that were graded (overall $P < 0.001$ for each feature). Figure 2 summarizes the presence of the five photographic features in each program. Figures 3 and 4 show examples of the photographic features. IMAGEnet produced mosaics that were more likely to have duplicated vessels, misplaced images, relative blurriness, and unacceptable overall quality than those produced by either AutoMontage or i2k Retina ($P < 0.001$ for each comparison). Mosaics by IMAGEnet were more likely to have breaks in blood vessel continuity than those by i2k Retina ($P < 0.001$), but not more likely than those by AutoMontage ($P = 0.95$). The i2k Retina and AutoMontage programs differed in only two of the photographic features, with duplicated vessels ($P = 0.001$) and breaks in blood vessel continuity ($P < 0.001$) observed more frequently in the AutoMontage images. There was no significant difference between the i2k Retina and AutoMontage programs for misplaced images ($P = 0.37$), relative blurriness ($P = 0.75$), or unacceptable overall quality ($P = 0.15$).

The overall rank assigned to a photographic mosaic correlated significantly with the presence of each of the five photographic features ($P < 0.001$ for each feature; Table 2). The magnitude of association was relatively strong for misplaced images, image blurriness, duplicated vessels, and overall qual-

ity. A much lower correlation coefficient was observed for vessel breaks. We also found that the ability of a program to incorporate all individual images into the mosaic was associated with its overall ranking, with a correlation coefficient of 0.43 ($P < 0.001$, permutation test, rank centroid difference).

All graders commented on two types of image distortions that were not elicited in the grading form but were present in some mosaics: (1) visible, overlapping borders of the individual images at the suture lines of the peripheral mosaic and (2) linear color changes in the mosaic, usually present in the center of the image (Figs. 3, 4), that were not present in the individual central photograph. These abnormalities were systematically quantified by grader A, who documented visible overlapping borders in 23% of the 99 AutoMontage-created mosaics, 1% of the 96 i2k Retina-created mosaics, and none of the 91 IMAGEnet-created mosaics. He also noted abrupt color changes in 22% of the 96 mosaics made by i2k Retina, 4% of the 99 mosaics made by AutoMontage, and none of the 91 mosaics made by IMAGEnet.

DISCUSSION

The i2k Retina and AutoMontage software packages appear to create high-quality photographic mosaics from digital photo-

FIGURE 3. Side-by-side comparisons of mosaics made with the three software programs, showing the completed mosaics, as well as the presence or absence of duplicated blood vessels, relative blurriness of the optic nerve, and overlapping borders from individual images.
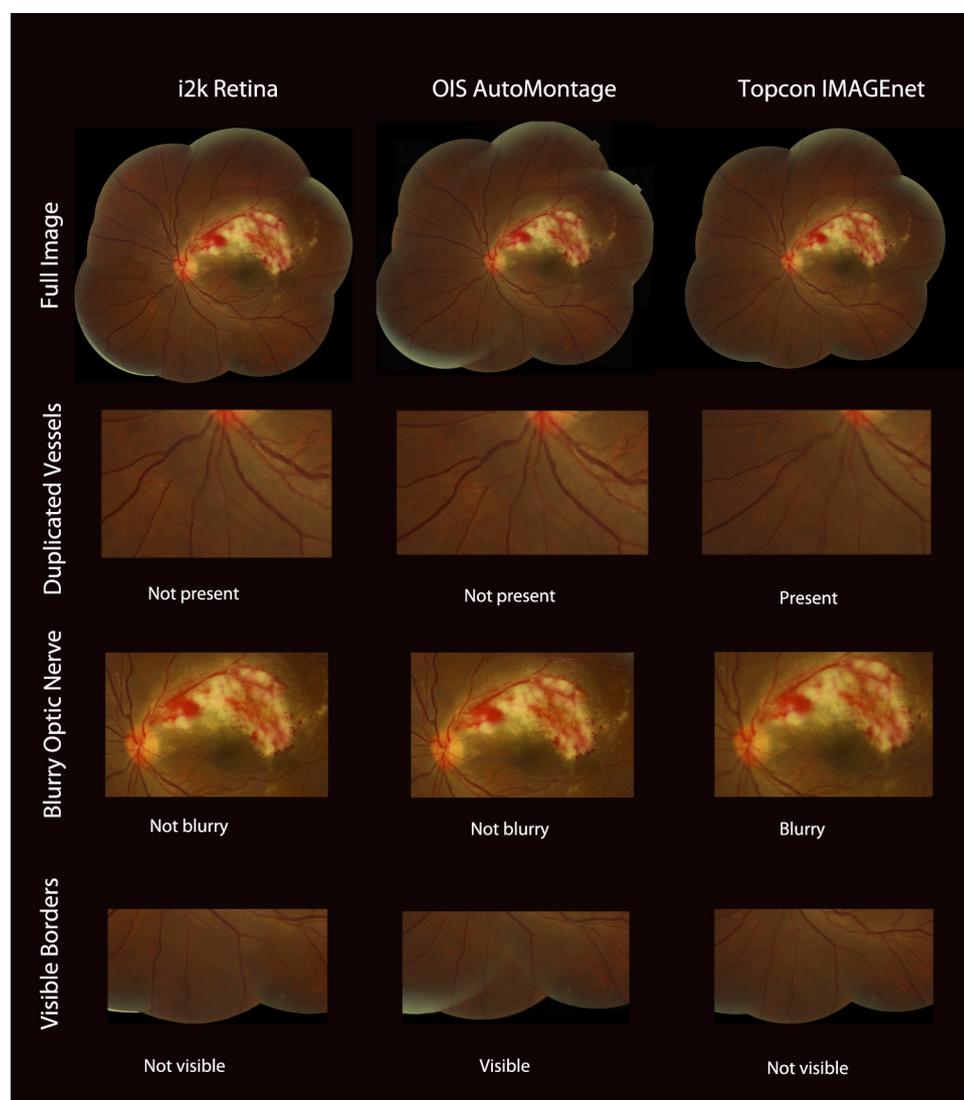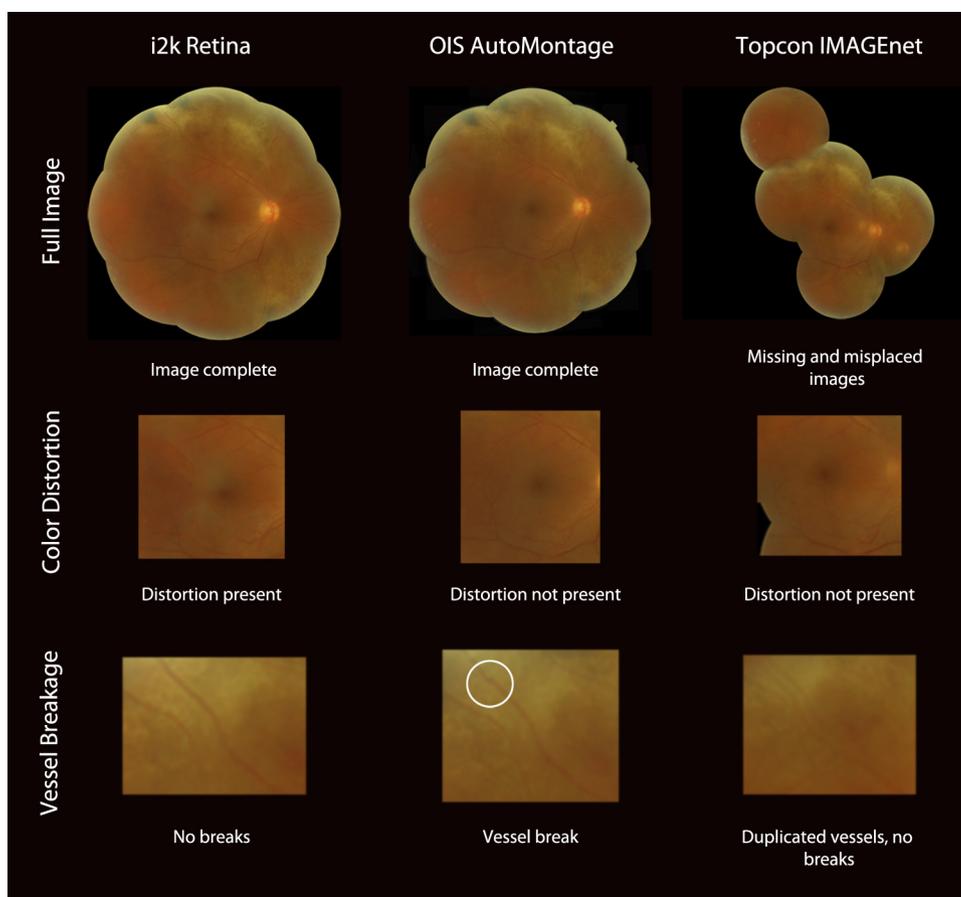
FIGURE 4. Side-by-side comparisons of mosaics made with the three software programs, showing the completed mosaics, as well as the presence or absence of missing and misplaced images, linear color distortion, and vessel breaks.

graphs of retinas with CMV retinitis. These programs created mosaics of higher quality and with fewer missing images or other errors than in those created by IMAGEnet. Although the i2k Retina and AutoMontage programs scored similarly in many aspects, all three graders ranked i2k Retina as the best program more frequently. Each of the photographic features documented by the three graders (duplicated vessels, vessel breaks, misplaced images, relative blurriness, and overall quality) was significantly associated with a program being ranked as inferior, with vessel breaks having a weaker association than the other errors.

In this study, we evaluated the specific features that play a role in determining the quality of a mosaic retinal image. Several studies have compared different imaging modalities and their clinical usefulness in retinal diseases, including comparisons between 35-mm film and digital imaging in diabetic retinopathy.[16-18] However, we are unaware of any previous reports comparing the clinician-assessed quality of different retinal imaging software products that perform the same imaging function. Because ophthalmic imaging technology is rapidly advancing, objective clinical evaluation of each new facet of technology is important. By assessing the automontage process, this study provides a framework for assessing other imaging programs based on clinician judgment.

Within the field of CMV research and telemedicine, the ability to create consistent, high-quality mosaics can be clinically useful. Although high-quality mosaics may not be necessary for the diagnosis of CMV retinitis, they allow for easier comparison with previous photographs, which could simplify the identification of new lesions and aid in the assessment of CMV retinitis border advancement. High-quality mosaics could therefore improve clinical practice, and they could also be used as an outcome measure in clinical and epidemiologic research.

We used a competitive ranking system in this study, which did not allow ties, and therefore we cannot provide information on the magnitude of difference between the three programs. However, we addressed this by also collecting information on five photographic features. We observed that although the i2k Retina software received the highest ranking most of the time, the quality of AutoMontage mosaics was similar to that of i2k Retina for three of the five features (misplaced images, relative blurriness, and overall quality), suggesting that there was probably not a large margin that separated the rankings of these two programs.

Vessel breakage in the retinal montages had some of the lowest intra- and interrater reliability values and also displayed the weakest degree of correlation with a program's ranking. In general, vessel breaks were often subtle findings that occurred at the mosaic periphery (Fig. 4). Perhaps in this location the

TABLE 2. Association of Mosaic Errors and Decreasing Rank as Expressed by Correlation Coefficients

Photographic Features	Correlation between Photographic Features and Decreasing Rank* (P)
Duplicated vessels	0.74 (<0.001)
Vessel breaks	0.38 (<0.001)
Misplaced Images	0.80 (<0.001)
Blurriness	0.77 (<0.001)
Unacceptable overall quality	0.73 (<0.001)

* Value represents the average correlation coefficient across all eyes.

breaks were not visually distracting and did not interfere with assessment of pathology. This finding suggests that vessel breaks may not be an appropriate measure of software performance. Programmers of mosaic software may want to focus on reducing the presence of other mosaic errors before addressing vessel breaks.

There were several limitations to this study. First, we did not examine any of the cartographic or mathematical problems involved in constructing an accurate mosaic (such as the coordinate system, the model for image transformation, and the construction of a flat image from a curved retinal surface[1]), but instead chose to focus on the final mosaic image itself. Second, we did not assess differences between the programs regarding measurement of lesions. Because of the cartographic distortions inherent in the mosaic construction process, linear and area measurements of the retina are artificially inflated, especially in the periphery of the mosaic—a problem outside the scope of this study. Third, we focused on photographic features that were relatively easy to quantify, but did not include some features that are most likely very important to the viewing experience but more difficult to assess, such as image distortion induced by the montaging procedure. Fourth, the retinal camera stored all individual photographs as highly compressed images (roughly 320 kb with dimensions of 3872×2592 pixels). It is possible that the mosaic programs would have performed differently had higher quality, uncompressed images been uploaded. However, results should not have been biased by the use of compressed images, since all three mosaic programs were loaded with the same sets of photographs. Finally, we did not standardize monitor size and resolution.

We found no significant difference in quality between i2k Retina and AutoMontage, both of which create high-quality images with fewer errors when compared to IMAGEnet. The i2k Retina program was most often ranked first compared to the other two programs. The accurate montages that these programs create may be valuable in future CMV retinitis research and telemedicine. Further investigation is necessary to determine whether our findings can be generalized to retinal diseases other than CMV retinitis.

References

1. Can ACS, Roysam B, Tanenbaum H. A feature-based algorithm for joint, linear estimation of high-order image-to-mosaic transformations: mosaicing the curved human retina. *IEEE Trans Pattern Anal Mach Intell.* 2002;24(3):412–419.

2. Mahurkar A, Vivino M, Trus B, Kuehl E, Datiles M, Kaiser-Kupfer M. Constructing retinal fundus photomontages. *Invest Ophthalmol Vis Sci.* 1996;37:1675–1683.

3. Tsai CL, Madore B, Leotta M, et al. Automated retinal image analysis over the Internet. *IEEE Trans Inform Technol Biomed.* 2008; 12(4):480–487.

4. Berger JW, Leventon M, Hata N, Wells W, Kikinis R. Design considerations for a computer-vision-enabled ophthalmic augmented reality environment. *Lecture Notes Comput Sci.* 1997;1205:399–408.

5. Wiegand TW, Young LH. Cytomegalovirus retinitis. *Int Ophthalmol Clin.* 2006;46(2):91–110.

6. Jacobson MA, Stanley H, Holtzer C, Margolis TP, Cunningham ET. Natural history and outcome of new AIDS-related cytomegalovirus retinitis diagnosed in the era of highly active antiretroviral therapy. *Clin Infect Dis.* 2000;30(1):231–233.

7. Heiden D, Ford N, Wilson D, et al. Cytomegalovirus retinitis: the neglected disease of the AIDS pandemic. *PLoS Med.* 2007;4(12): e334.

8. Pathanapitoon K, Ausayakhun S, Kunavisarut P, et al. Blindness and low vision in a tertiary ophthalmologic center in Thailand: the importance of cytomegalovirus retinitis. *Retina.* 2007;27(5):635–640.

9. Biswas J, Madhavan HN, George AE, Kumarasamy N, Solomon S. Ocular lesions associated with HIV infection in India: a series of 100 consecutive patients evaluated at a referral center. *Am J Ophthalmol.* 2000;129(1):9–15.

10. Ausayakhun S, Watananikorn S, Ittipunkul N, Chaidaroon W, Patikulsila P, Patikulsila D. Epidemiology of the ocular complications of HIV infection in Chiang Mai. *J Med Assoc Thai.* 2003;86(5):399–406.

11. Ausayakhun S, Skalet A, Jirawison C, et al. Accuracy and reliability of telemedicine for diagnosis of cytomegalovirus retinitis. *Am J Ophthalmol.* Published online August 19, 2011.

12. Weinberg DV, Holbrook JT, Hubbard LD, Davis MD, Jabs DA, Holland GN. Clinician versus reading center assessment of cytomegalovirus retinitis lesion size. *Ophthalmology.* 2005;112(4): 559–566.

13. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* Boca Raton, FL: Chapman and Hall/CRC; 1993;168–199.

14. Nam J. Interval estimation of the kappa coefficient with binary classification and an equal marginal probability model. *Biometrics.* 2000;56:583–585.

15. Fleiss J. 1981 statistical methods for rates and proportions. *Epidemiology.* 1996;7:561.

16. Bursell S-E, Cavallerano JD, Cavallerano AA, et al. Stereo nonmydriatic digital-video color retinal imaging compared with Early Treatment Diabetic Retinopathy Study seven standard field 35-mm stereo color photos for determining level of diabetic retinopathy. *Ophthalmology.* 2001;108(3):572–585.

17. Prasad SBP, Clearkin LG, Phillips RP. Digital fundus imaging: a quality and cost comparison with 35-mm film. *Acta Ophthalmol Scand.* 1999;77(1):79–82.

18. Lim JI, LaBree L, Nichols T, Cardenas I. A comparison of digital nonmydriatic fundus imaging with standard 35-millimeter slides for diabetic retinopathy. *Ophthalmology.* 2000;107(5):866–870.

APPENDIX A

To assess whether the ranks of the three software programs differed statistically, we proceeded as follows. For each grader, we obtained the rank for software programs 1, for 2, and 3 (note that these dependent variables always sum to 6). For each eye, we transformed these ranks to the coordinates of a point with a distance to the nearest side of an equilateral triangle that is proportional to the rank:

$$x_i = \frac{R_i^{(1)}}{12} + \frac{R_i^{(2)}}{6}$$

and

$$y_i = \frac{R_i^{(1)}\sqrt{3}}{12}.$$

Under the null hypothesis that the ranks are independent of the software program, the centroid of these points lies at the center of the triangle. We computed the following multivariate statistic (combining all three ranks for each grader on each eye). Specifically, we computed the distance of the rank centroid from the center of the triangle,

$$r = \sqrt{\left(\bar{x} - \frac{1}{2}\right)^2 + \left(\bar{y} - \frac{\sqrt{3}}{6}\right)^2}.$$

The sampling distribution of this statistic under the null hypothesis was computed using the permutation method with 10,000 replicates. All three graders were weighted equally.