Interexpert Agreement in the Identification of Macular Location in Infants at Risk for Retinopathy of Prematurity

Michael F. Chiang, MD; Preeti J. Thyparampil, MD; Daniel Rabinowitz, PhD

Objective: To characterize variability in the identification of the macular center among retinopathy of prematurity (ROP) experts.

Methods: A printed set of 25 wide-angle retinal images was compiled from infants at risk for ROP using a commercially available camera. Ten recognized ROP experts were asked to mark the macular center on each image. For each image, we measured the distance from the optic disc center to the marked macular center. Distances were standardized by normalizing the horizontal optic disc diameter in each image to 0.93 mm. In images with visible peripheral disease, interexpert agreement on the presence of zone I disease was also determined.

Results: For the image with the least variability among experts, mean (SD) distance from the optic disc to the macular center was 3.69 (0.21) mm (range, 3.13-3.81 mm). For the image with the greatest variability among experts, distance from the optic disc to the macular center was 4.32 (1.19) mm (range, 3.21-7.19 mm). In 7 of 21 images (33%) with visible peripheral disease, there would have been disagreement among experts in the diagnosis of zone I disease based on identification of the macular center. Among the 10 experts, in 17 of 25 images (68%), 1 expert identified the distance between the optic disc and macular center to be greater than 1 SD from the mean.

Conclusion: Significant variability exists among experts in identification of the macular center from wide-angle images, which raises concerns about the reliability of zone I ROP diagnosis.

Arch Ophthalmol. 2010;128(9):1153-1159

Retinopathy of prematurity (ROP) is a vasoproliferative disease affecting low-birth-weight infants and is a leading cause of childhood blindness worldwide. Major advances in diagnosis and treatment have occurred. For example, the International Classification of ROP established a standard system for describing examination findings, thereby creating a foundation for clinical care and research.1,2 The Cryotherapy for ROP3 (CRYO-ROP) and Early Treatment for ROP4 (ETROP) trials identified criteria for which treatment with cryotherapy or laser photocoagulation was shown to improve visual outcomes.

These studies have shown that identification of zone I ROP is critical for decisions regarding treatment and prognosis. Zone I of the retina is defined as a circle, the radius of which extends from the optic disc center to twice the distance from the optic disc center to the macular center. In the CRYO-ROP study, poor structural outcomes occurred in 78% of zone I eyes treated for threshold disease, whereas poor structural outcomes occurred in only 26% of zone II eyes that were treated.3 The ETROP trial subsequently showed improved outcomes in zone I eyes randomized to early treatment.4 On the basis of these studies, treatment is recommended for type 1 ROP, which is defined as zone I ROP of any stage in the presence of plus disease; zone I, stage 3 disease whether or not plus disease is present; and zone II ROP of stage 2 or 3 in the presence of plus disease. Careful follow-up is recommended for type 2 ROP, which is defined as zone I ROP of stage 1 or 2 without plus disease, or zone II, stage 3 disease without plus disease. Therefore, zone I disease must be detected accurately and reproducibly.

However, zone I diagnosis may be subjective because it requires identification of the macular center with delineation of the appropriate circular radius. Although clinical examination is considered the criterion standard for ROP care, to our knowledge agreement among multiple examiners in recognizing zone I disease has not been systematically studied. This study was designed to characterize variability in the identification of the macular center among 10 recognized experts using a set of wide-angle retinal images from infants at risk for ROP.

Author Affiliations: Departments of Ophthalmology (Drs Chiang and Thyparampil) and Biomedical Informatics (Dr Chiang), College of Physicians and Surgeons, and Department of Statistics (Dr Rabinowitz), Columbia University, New York, New York.
METHODS

This study was approved by the institutional review board at Columbia University. Written informed consent was obtained from all expert participants, and waiver of consent was obtained for the use of deidentified retinal images.

STUDY PARTICIPANTS

A set of 25 wide-angle retinal images was obtained from premature infants during routine ROP care using a commercially available camera (RetCam-I; Clarity Medical Systems, Pleasanton, California). Images were taken of infants between 32 and 38 weeks' postmenstrual age (PMA) and were not annotated with any descriptive information. No images were repeated.

Eligible experts for this study were defined as practicing pediatric ophthalmologists or retina specialists who met at least 1 of the following criteria: having been a principal investigator or study investigator in the CRYO-ROP or ETROP study, or having published at least 2 peer-reviewed ROP articles. Experts were asked whether their institution had a RetCam and about their experience in interpreting these images (experienced, somewhat experienced, or not experienced). Finally, experts were asked to rate each image based on (1) adequacy of photographic quality (adequate, possibly adequate, or inadequate) and (2) confidence in identifying the macular center (confident, possibly confident, or not confident).

CHARACTERISTICS OF EXPERT PARTICIPANTS AND RETINAL IMAGES

On the basis of the study definition of expertise, 11 individuals were invited, of whom 10 (91%) consented to participate. Among these 10 experts, 5 (50%) had served as principal investigators or study investigators in the CRYO-ROP or ETROP trials, and all (100%) had count-
thored at least 2 peer-reviewed ROP articles. All experts worked at institutions with a RetCam, and all reported that they were experienced in RetCam image review. Four experts (40%) were pediatric ophthalmologists and 6 (60%) were retina specialists.

The Table summarizes characteristics of the 25 wide-angle retinal images. Five images (20%) were taken at 32 to 33 weeks' PMA, 3 (12%) were taken at 34 to 35 weeks' PMA, and 17 (68%) were taken at 36 to 38 weeks' PMA. Twenty-one images (84%) had visible peripheral disease; in all of these images, the nearest distance to visible peripheral disease was greater than twice the mean of experts’ marked distances to the macular center.

### IMAGE QUALITY AND EXPERT CONFIDENCE RATINGS

Each expert independently reviewed all 25 images, for a total of 250 macular center, image quality, and confidence-rating responses. Image quality was scored as “adequate” for identification of the macular center in 161 of 250 responses (64%), “possibly adequate” in 68 responses (27%), and “inadequate” in 21 responses (8%). Confidence in identifying the macular center was scored as “confident” in 109 of 250 responses (44%), “somewhat confident” in 103 responses (41%), and “not confident” in 38 responses (15%).

### IDENTIFICATION OF MACULAR CENTER

Figure 1 shows examples of expert responses, indicating different patterns of discrepancy. Distances from the optic disc center to the marked macular center are displayed graphically in Figure 2. The mean (SD) distance from the optic disc to the macular center was 3.69 (0.21) mm (range, 3.13-3.81 mm) for the image with the best interexpert agreement (image 17 in the Table and Figure 2) and 4.32 (1.19) mm (range, 3.21-7.19 mm) for the image with the worst agreement (image 22 in the Table and Figure 2). When images in which experts were “not confident with marking” were excluded, the subject-specific and residual variance component estimates were smaller. However, there was an insufficient sample size to conclude that restricting attention to images in which experts were confident would result in decreased variability of macular markings.

Figure 3 displays variations in distances from the optic disc center to the marked macular center among the...
10 study experts, scaled as number of standard deviations from the mean distance for each image. For the expert with the least mean deviation from zero (expert 3 in Figure 3), the mean deviation was 0.53 SDs from the mean for each image. For the expert with the highest deviation (expert 2 in Figure 3), the mean deviation was 1.45 SDs from the mean for each image. Linear regression analysis found no statistically significant relationship between PMA and variation in marked distances to the macular center (ie, number of standard deviations from the mean) \((P = .21)\).

The 2-way mixed-effects ANOVA found the ratio of rater-specific variability to total variability around the expected distances to be 0.40 \((P = .02)\). Therefore, significant evidence indicates that different raters have different propensities, with some tending to systematically provide higher measurements than expected and others tending to provide lower ones. Because of the possibility of nonnormality, the analysis was repeated using a non-parametric rank-based ANOVA with no qualitative changes in result.

**DIAGNOSIS OF ZONE I DISEASE**

In 7 of 21 images with visible peripheral disease (33%), there would have been disagreement among some experts in the diagnosis of zone I disease based on the identification of the macular center and the location of visible peripheral disease. Among the 10 experts, in 17 of 25 images (68%), 1 expert identified the distance between the optic disc and macular center to be greater than 1 SD from the mean (Figure 3).

Using a 2-way mixed-effects ANOVA, the estimates (standard errors) of variances were 0.15 \((P = .07)\) for expert-specific variability and 0.22 \((P = .02)\) for residual variability. With the assumptions and variance estimates from the ANOVA model, the probability of diagnosing zone I disease among the experts varied.
disease depends on the location of peripheral disease according to the following function:

\[ 1 - \Phi \left( \frac{x}{2\times\sqrt{0.15 + 0.22}} \right) \]

where \( \Phi(t) \) is the probability that a standard normal variable is less than \( t \), and \( x \) is the \([\text{distance to peripheral disease} - (2\times\text{distance to the true macular center})]\).

This estimate was compared with the data by treating the mean of the experts’ measurements as the true distance and by determining the proportion of raters who would have diagnosed zone I disease as a function of the peripheral disease location (Figure 4). If all experts would have diagnosed zone I disease when (distance to peripheral disease) – \((2\times\text{distance to the true macular center})\) less than 0, then this curve would show that y-axis values equal 1.0 for all x-axis values less than or equal to 0. If all experts would have diagnosed zone I disease only when (distance to peripheral disease) – \((2\times\text{distance to the true macular center})\) less than 0, then this graph would show that y-axis values equal 0 for all x-axis values greater than or equal to greater than or equal to 0. Rather, the curve suggests that there is imperfection in diagnosis of zone I ROP as the location of peripheral disease approaches twice the distance to the macular center.

**COMMENT**

The key findings from this study are as follows: (1) There is significant variability among experts in the identification of the macular center from wide-angle images. (2) Interexpert agreement in the diagnosis of zone I ROP is likely to be imperfect. The CRYO-ROP and ETROP trials demonstrated that zone I disease is a critical feature of clinically significant ROP and that there are important benefits to early treatment of these eyes.3,4 Interestingly, there was a greater frequency of zone I disease in the ETROP study compared with the CRYO-ROP study, even when infants were controlled for birth weight and gestational age. At the same time, the incidence of favorable structural outcome in eyes treated for zone I disease was better in the ETROP cohort than in the CRYO-ROP cohort.7 One possible explanation for these differences may be that unidentified changes in the care of premature infants are causing an increased incidence of zone I disease. Alternatively, the differences might result from discrepancies in ascertainment, so that more zone I eyes in the CRYO-ROP study may have been underdiagnosed or more zone II eyes in the ETROP trial may have been overdiagnosed. We note that accurate and reproducible disease classification is essential for clinical research studies.

Retinal zones are defined using landmarks such as the optic disc, macular center, and ora serrata.1,2 Although this system assigns discrete values to each zone, we find interexpert variability in perceived location of the macular center (Table). Some images displayed a bimodal-type response distribution (Figure 1B), suggesting that 2 specific points had features that experts associated with the macular center. Other images displayed a continuous response distribution (Figure 1C and D), suggest-
ing lack of consensus regarding the macular center. Overall, 7 of 21 images with visible peripheral disease (33%) showed discrepancies in zone I diagnosis according to the location of marked macular centers. A variance model showed that the experts would have diagnosed zone I disease based on macular marking with a probability that varied according to the location of peripheral disease (Figure 4). For example, for an image in which the distance to peripheral disease was 1 mm greater than twice the distance to the true macular center (defined as the mean of marked distances), approximately 12% of experts would be expected to overdia gnose this as zone I ROP. For an image in which the distance to peripheral disease was 1 mm less than twice the distance to the true macular center, approximately 12% of experts would be expected to underdiagnose this as not zone I ROP. These findings suggest that there is likely to be imperfect agreement among experts for zone I diagnosis.

Our findings are consistent with previous studies examining interexpert agreement in ROP and other medical conditions. Plus disease diagnosis using retinal photographs was variable among pediatric ophthalmologists and retinal specialists. In the CRYO-ROP study, 12% of eyes diagnosed as having threshold disease after ophthalmoscopy by a certified examiner were diagnosed at a less-than-threshold level during confirmatory ophthalmoscopy by a different certified examiner. It is conceivable that a study design in which experts were masked to the findings of other experts would have shown even higher levels of disagreement. In an article examining the use of fluorescein angiography for the evaluation of photodynamic therapy eligibility in patients with neovascular age-related macular degeneration, the κ statistic for interphysician reliability was 0.37 to 0.40. In the domain of emergency medicine, the κ statistic for reliability among physicians, residents, and surgeons for abdominal examination findings in adolescent patients was −0.04 to 0.38. For these reasons, we believe that training and education in ROP are crucial to ensure the highest level of patient care and to allow findings from major studies such as the CRYO-ROP and ETROP trials to be applied properly and consistently.

The macular center may be difficult to identify precisely because foveal morphologic characteristics evolve significantly between 22 and 36 weeks’ PMA and may not reach full maturity until the child is 15 to 45 months of age. One study identified a sequence of visible changes associated with macular development in the premature infant, ranging from early pigmentation of the macular area at 34 weeks to subsequent development of the macular annular ring reflex and foveolar reflex. Lack of consensus among experts on these features related to macular architecture may have contributed to variability in our study. Recognition of these features might increase consistency and decrease ambiguity in clinical diagnosis. For example, darker eyes may have pigmented areas that are distinct from the actual macula (Figure 1D), and the retinal vascular arcades might be expected to branch lateral to the fovea (Figure 1B). Alternative methods for defining zone I disease, based on retinal morphologic features or differences between underlying vasculogenic and angiogenic processes, might be developed in the future. Clinically, it would be most useful to know the agreement of zone I diagnosis among multiple experts performing serial indirect ophthalmoscopy on the same infant. However, that type of study may be impractical because of safety concerns.

To simulate a real-world situation in this study, we used RetCam images obtained from infants routinely examined for ROP during a 1-year period. This wide-angle contact camera is the most well-known instrument for pediatric retinal imaging, with a field of view of more than 100°. In contrast, indirect ophthalmoscopy provides a 40° to 50° field of view, with better visualization of 3-dimensional retinal contours and potential for using this field to estimate the boundary of zone 1. This difference in perspective could affect the interpretation of study findings, particularly if some experts had less experience in interpreting RetCam images. It is also conceivable that pressure on the eye from the camera could alter the appearance of retinal vessels. However, all study experts reported a high level of experience interpreting RetCam images, and there were no statistically significant differences in mean marked distances to the macular center after excluding images in which experts were not confident. In fact, wide-angle imaging may provide important advantages over ophthalmoscopy with regard to the visualization of macular reflexes and vascular architecture. Furthermore, all experts were asked to review the exact same images, which might have decreased the variability compared with serial ophthalmoscopy because examination quality may vary based on infant stability or cooperation.

Several other limitations should be noted: (1) Images were taken of infants ranging from 32 to 38 weeks’ PMA. Because of progressive macular development during this period, it might be expected that interexpert agreement would be poorest in the youngest infants. Linear regression analysis found no statistically significant relationship between PMA and variation in marked distances to the macular center, although studies with larger sample sizes may be useful. (2) Images were not annotated with any clinical data. Although it is not clear that this would have affected diagnostic agreement, it may have biased the results against accurate photographic interpretation. (3) To account for potential differences in image magnification, all horizontal optic disc widths were standardized to 0.93 mm, based on previous research in postmortem samples from premature infants. The mean disc-to-macula distance among all images in this study was 3.70 mm, whereas a different investigation calculated a mean distance of 4.4 mm using analysis of retinal photographs. This discrepancy may result from shrinkage during tissue preservation methods. Although there may be uncertainty about the exact disc-to-macula distances depending on the technique used, this would not change our key study findings. (4) There were no study images in which most experts would have diagnosed zone I disease. Because the findings still demonstrate significant variability among experts, it is not clear that this would have affected the study findings. From a practical standpoint, infants with posterior ROP tend to be younger and therefore more difficult to photograph clearly. (5) Experts were invited according to academic criteria, but this might not reflect real-world clini-