Computer-Assisted Measurement of Retinal Vascular Width and Tortuosity in Retinopathy of Prematurity

Amanda E. Kiely, MD; David K. Wallace, MD, MPH; Sharon F. Freedman, MD; Zheen Zhao, PhD

Objectives: To validate the accuracy of ROPtool software in measuring retinal vascular width and tortuosity in a large image set compared with expert diagnoses.

Methods: Tortuosity and dilation indexes generated by ROPtool were compared with 3 expert consensus grades of normal, pre-plus, or plus disease for 368 quadrants in 92 RetCam (Clarity Medical Systems, Pleasanton, California) fundus images. Sensitivity and specificity of ROPtool software in diagnosing tortuosity and dilation sufficient for plus and pre-plus disease were calculated. These measures were compared with individual accuracies of 3 experienced pediatric ophthalmologists.

Results: The mean tortuosity indexes for expert-diagnosed categories of normal, pre-plus, and plus disease were 7.04, 18.73, and 34.62, respectively (P < .001), and the mean dilation indexes were 9.63, 12.05, and 13.61, respectively (P < .001). When optimal tortuosity and dilation index thresholds (from receiver operating characteristic curves) were applied, resultant sensitivity and specificity were 0.913 and 0.863, respectively, for plus tortuosity and 0.782 and 0.840, respectively, for plus dilation. These values were comparable to the performance of examiners judged against the same expert panel.

Conclusion: ROPtool version 2.1.5 accurately measures tortuosity and dilation of posterior pole blood vessels in RetCam images, corresponding well with expert diagnostic categories of normal, pre-plus, and plus disease and performing comparably to experienced examiners.

Arch Ophthalmol. 2010;128(7):847-852

Retinopathy of Prematurity (ROP) is a leading cause of childhood blindness. Prompt accurate diagnosis is necessary for appropriate management by laser ablative therapy. One of the most important prognostic indicators in ROP is plus disease, present when vascular changes are so marked that the posterior veins are enlarged and the arterioles are tortuous. Plus disease is typically diagnosed during indirect ophthalmoscopy based on an examiner’s comparison of posterior pole changes with those in a standard photograph that demonstrates the minimal abnormality necessary for plus disease. Pre-plus disease exists when vascular abnormality is insufficient for a diagnosis of plus disease but there is more arterial tortuosity and more venous dilation than normal. However, plus and pre-plus diagnoses based on review of images are subject to considerable interexaminer disagreement. In one study, Wallace et al found that 3 experts disagreed on the diagnosis of plus for 18 of 67 images (27%). In another study, Chiang et al found that 22 experts unanimously agreed on the diagnosis of plus vs not plus disease for only 7 of 34 images (21%).

To address this inconsistency in diagnosis of plus disease, several quantitative image analysis algorithms have been proposed. One of these is ROPtool, a program that quantifies tortuosity and dilation of posterior pole vessels chosen by an operator. Previous studies have demonstrated accuracy in the tortuosity measure of ROPtool, detecting plus-level abnormality at least as well as experienced examiners and accurately assessing tortuosity in high-quality images from video indirect ophthalmoscopy. Recent improvements to the dilation measure of ROPtool that minimize effects of image blur include a ridge-valley traversal algorithm, which uses the cross-sectional profile of a vessel to find its edges, and use of the distance between the optic nerve and the macula as a scaling factor to correct for image magnification. In a small pilot study,
this newly developed measure of vessel width performed well. Our primary objective in the present study was to validate the accuracy of ROPtool software (version 2.1.5) in measuring retinal vascular width and tortuosity using a large image set. This validation is an important step toward using ROPtool clinically at the bedside or as part of a telemedicine program. We also sought to assess (for the first time) diagnostic accuracy of ROPtool when tortuosity and dilation were considered together, as is done during clinical examination.

METHODS

PILOT STUDY FOR PRELIMINARY ROPtool CALIBRATION

Details of ROPtool operation and analysis have been published.11,12,20 Previously, the dilation measurements of ROPtool for 154 vessels from 20 RetCam (Clarity Medical Systems, Pleasanton, California) images representing 20 infants were compared with consensus grades by 2 of us (D.K.W. and S.F.F.).11,20 In the present study, without knowledge of prior grading, another of us (A.E.K.) used ROPtool version 2.1.5 to analyze the vessels graded by the 2 others. From these data, receiver operating characteristic (ROC) curves were constructed for the tortuosity index (TI) and dilation index (DI) of ROPtool. ROC curves plot sensitivity on the y-axis and 1 minus specificity (the false-positive rate) on the x-axis for multiple threshold values generated by a diagnostic test. From these ROC curves, plus and pre-plus disease thresholds (or TI and DI cutoff points) were chosen to optimize sensitivity and specificity.

VALIDATION STUDY OF ROPtool ACCURACY IN ASSESSING RETINAL VASCULAR TORTUOSITY AND DILATION

ROPtool Analysis

A convenience sample of 92 RetCam fundus images was used for this larger validation study. These were a subset of images described in previous publications,12,13 none of which were in the pilot study image set. All images were analyzed using ROPtool operated by one of us without knowledge of prior grading (A.E.K.), who selected 2 major vessels in each of 4 quadrants wherever possible, with a goal of identifying the major arteriole and major venule in each quadrant. Although the program has the capacity to record operator judgments of vessel identity (arteriole vs venule), this function was not used herein. ROPtool analysis required approximately 1 to 3 minutes per image and never exceeded 6 minutes. ROPtool returned TIs and DIs for each vessel and each quadrant, and the “quadrant index” equaled the largest vessel index in that quadrant. Figure 1 shows the interface of ROPtool version 2.1.5.

Accuracy of ROPtool

Using the cutoff points generated in the pilot study, ROPtool scores were then converted into designations of plus, pre-plus, or neither for tortuosity alone and for dilation alone. Next, sensitivity and specificity of ROPtool in diagnosing tortuosity and dilation sufficient for plus and pre-plus disease categories were calculated using the consensus of 3 expert examiners (“experts”) as the reference standard. All 3 were certified investigators in the Early Treatment for Retinopathy of Prematurity

Figure 1. ROPtool version 2.1.5 (Duke University, Durham, North Carolina) user interface with analyzed image visible. View is centered on the optic nerve and approximates the view with a 28-diopter (D) lens (the nerve and 28-D view are outlined in blue). Quadrants are demarcated by green lines, and red traces indicate selected vessels. Completed analysis is shown, including tortuosity index and dilation index for each quadrant and an overall assessment of tortuosity alone, dilation alone, and overall disease severity. DB indicates database; DI, dilation index; TI, tortuosity index; A, arteriole; and V, venule.
For comparison, individual accuracies of 3 additional experienced pediatric ophthalmologists (“examiners”) relative to those of the experts were calculated in the same way.

Details of Expert and Examiner Grading

The experts and examiners assessed whether each of 4 quadrants in an image (cropped to a 28-diopter indirect ophthalmoscopy view) contained retinal vascular dilation, tortuosity, or both sufficient to diagnose plus or pre-plus disease. Grade 0 corresponded to normal vasculature, 1 to pre-plus disease, and 2 to plus disease. In preparation for this process, all 6 graders were reminded of the definition of plus disease and were sent a copy of the standard photograph used in the Cryotherapy for Retinopathy of Prematurity study. The grades of the 3 experts were averaged and rounded to the nearest whole number to establish the reference standard (“truth”) for each quadrant.

ROC Curves

Using data from all 368 quadrants of 92 RetCam fundus images, ROC curves were constructed for plus and pre-plus TI and DI, plotting sensitivity vs 1 minus specificity of each potential threshold index compared with quadrant-level expert truth. Areas under the curves were then calculated. From these curves, candidate plus and pre-plus thresholds were identified for future recalibration of ROPtool.

STATISTICAL ANALYSIS

All data were collected and analyzed using commercially available software (Microsoft Office Excel 2003; Microsoft Corporation, Redmond, Washington). ROC curves were generated and analyzed using Excel 2003 and MatLab 7.7 (2008b; MathWorks, Natick, Massachusetts). For determining statistical significance of TI and DI “trends” across the 3 expert diagnostic groups, the generalized estimating equation approach was used to account for correlation within quadrants of the same eye. Because the increase over the disease groups for each outcome measure was not linear, the tests for trends were performed as tests for the difference from the normal group for each disease group (pre-plus and plus) simultaneously.

RESULTS

TRENDS FOR INCREASING ROPtool VALUES

Figure 2 shows trends for increasing ROPtool values for tortuosity and dilation (P < .001 for both) across expert-diagnosed categories of normal, pre-plus, and plus disease. The mean TIs for the 3 categories were 7.04, 18.73, and 34.62, respectively, and the median TIs were 6.27, 18.38, and 30.39, respectively. The mean DIs were 9.63, 12.05, and 13.61, respectively, and the median DIs were 9.81, 12.58, and 13.16, respectively.

ACCURACY OF ROPtool AND INDIVIDUAL EXAMINERS RELATIVE TO EXPERT CONSENSUS

The Table shows sensitivity, specificity, and overall concordance for diagnosing tortuosity and dilation sufficient for plus disease. The sensitivities of the pilot study thresholds were excellent (1.000 and 0.964 for tortuosity and dilation, respectively), but the specificities of 0.658 and 0.446, respectively, indicated many false-positive results or overcalls. Overall concordance with experts was 0.701 for tortuosity and 0.541 for dilation using these a priori pilot study thresholds. When new optimal cutoff points were chosen from the present data set, specificity and overall concordance of ROPtool improved to 0.863 and 0.870, respectively, for plus tortuosity, and to 0.840 and 0.832, respectively, for plus dilation.

Figure 3 shows ROC curves for tortuosity at the plus (A) and pre-plus (B) levels and for dilation at the plus (C) and pre-plus (D) levels. Areas under the curve are robust in all 4 cases.

INTEREXPERT AGREEMENT

In the pilot study, the ophthalmologists whose consensus comprised the reference standard agreed on a designation of plus, pre-plus, or normal dilation without adjudication for 127 of 154 (82%) vessels; agreement on the degree of tortuosity was previously reported to be 82%.
The 3 experts determining truth in the larger validation study were unanimous in their 3-level tortuosity grades for 242 of 368 quadrants (66%) and in their 3-level dilation grades for 187 of 368 quadrants (51%). In 13 (4%) and 18 (5%) quadrants, their opinions of tortuosity and dilation, respectively, diverged completely, with one giving a grade of “plus,” another “pre-plus,” and the third “normal.”

### COMMENT

ROPtool version 2.1.5 performed well in the present study, returning accurate measurements of tortuosity and dilation for high-quality RetCam fundus images. ROPtool analysis is rapid, typically requiring less than 3 minutes per image. The robust areas under the ROC curves indicate good correspondence of ROPtool with expert diagnoses of plus- and pre-plus-level retinal vessel tortuosity and dilation. We have shown herein that, given optimized plus and pre-plus thresholds, the diagnostic capabilities of ROPtool are on par with those of experienced examiners.

By contrast, ROPtool did not perform as well when the program was calibrated with thresholds determined from pilot study data. Instead, ROPtool performed well when calibrated with thresholds determined from the data set generated in the present study.

### Table. Sensitivity, Specificity, and Overall Concordance of ROPtool and Individual Examiner Diagnosis of Plus Disease Compared With Expert Consensus

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Concordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plus Disease Tortuosity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROPtool threshold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pilot study</td>
<td>1.00</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>New&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.91</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>Expert</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.94</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>2</td>
<td>0.80</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>3</td>
<td>0.63</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>Plus Disease Dilation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROPtool threshold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pilot study</td>
<td>0.96</td>
<td>0.45</td>
<td>0.54</td>
</tr>
<tr>
<td>New&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.78</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>Expert</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.94</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>2</td>
<td>0.89</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>3</td>
<td>0.49</td>
<td>0.97</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<sup>a</sup>Pilot study indicates ROPtool performance when calibrated with the numeric plus and pre-plus disease thresholds determined a priori from pilot study data.

<sup>b</sup>New indicates ROPtool performance when calibrated with optimal thresholds selected from the data set generated in the present study.

### Figure 3. Receiver operating characteristic curves assessing the diagnostic accuracy of ROPtool. Tortuosity index (A and B) and dilation index (C and D) agreement with expert diagnoses is shown for plus disease (A and C) and for pre-plus disease or worse (B and D). AUC indicates area under the curve. Arrows show optimal thresholds to maximize sensitivity while minimally compromising specificity (tortuosity index of 17.00 and dilation index of 10.50 for plus disease and tortuosity index of 10.50 and dilation index of 10.50 for pre-plus disease).
a priori from pilot study data. This disparity is likely due to disagreement between different groups of experts used for the pilot study and the present study. Chiang et al\(^6\) previously reported that a panel of 22 experts agreed on 3-level categorization of plus disease, similar to that in the present study, only 12% of the time. In their study the mean weighted $\kappa$ statistics for individual experts compared with the overall panel indicated only fair ($\kappa$ range, 0.21-0.40) to moderate ($\kappa$ range, 0.41-0.60) agreement. Therefore, it is not surprising that our reference standard shifted when determined by a panel of 3 experts instead of the 2 of us used for the pilot study. Among our examiners, we also observed a wide range of sensitivities for detecting abnormality, 0.630 to 0.935 for plus-level tortuosity and 0.491 to 0.945 for plus-level dilation. Although human variability represents a fundamental limitation in using expert consensus as diagnostic truth, human opinion remains the standard approach in clinical care for ROP. Given the capacity of ROP tool for consistent measurement, this computer program can be calibrated to match any desired thresholds for plus and pre-plus disease, even if these vary among different groups of clinical experts.

While established thresholds of disease severity are necessary for diagnosis and intervention, the spectrum of indexes in ROP tool may eventually allow these to be tailored to each clinical circumstance. For instance, in a telemedicine screening program, sensitivity might be the most important factor. In this circumstance, it is critical to avoid missing disease that threatens a child’s vision, even at the expense of increased false-positive results. The disadvantage of false-positive results is that they require the ophthalmologist to examine an infant in the nursery who does not need treatment, despite the use of photoscreening. We believe that these additional examinations are a reasonable price to pay for avoiding wholly unacceptable false-negative results. Thresholds for referral could be tailored to accommodate a primary provider’s baseline diagnostic capabilities simply by calibrating ROP tool. If desired, excellent sensitivity for eyes with plus disease requiring treatment or with pre-plus disease requiring close ophthalmologic supervision could be attained. On the other hand, if ROP tool was used to give a bedside “second opinion,” balancing sensitivity and specificity would be more important. Furthermore, new ROP therapies on the horizon such as anti-vascular endothelial growth factor pharmacotherapy might be used at a different level of disease severity than laser treatment. Optimal points for intervention could be located along a spectrum of cutoff points to produce a multipurpose diagnostic tool. In the future, we might even move away from the notion of 1 or 2 cut points in ROP management altogether, considering instead a progression of anatomical changes in retinal vessels that can be quantified.

Finally, results of this study give new insight into the human diagnostic process. For tortuosity and dilation, there were trends for increasing index values from one diagnostic category to the next. Of note, tortuosity measurements covered a much larger range than dilation measurements. This finding indicates more marked contrast between normal vessel curvature and pathologic tortuosity, which a human can appreciate visually. Although the trend for increasing tortuosity from normal to plus disease is clearly demonstrated (Figure 2A), we see only minimal difference in median vessel width from pre-plus to plus disease (Figure 2B). One possibility is that ROP tool cannot distinguish pathophysiologic changes in vessel dilation, but the marked difference between normal and greater-than-normal (plus or pre-plus) dilation indexes suggests otherwise. A second possibility is that the range of pathophysiologic change in vessel width from normal to pre-plus or plus level is narrower than that for vessel tortuosity in ROP. A third possibility is that clinicians appreciate increasing venous dilation but require a sufficiently tortuous arteriole to consider a quadrant plus disease, no matter how dilated nearby vessels appear. Therefore, expert categories of plus dilation and pre-plus dilation have considerable overlap of ROP tool values. Clinicians typically consider tortuosity and dilation together when judging plus disease, and further study is necessary to identify the best method for mathematically combining tortuosity and dilation measurements for overall diagnoses that reflect expert truth.

This study should be viewed in light of some limitations. First, our convenience sample size was limited in number and was composed entirely of RetCam fundus images, which are not available to all care providers. Second, experts graded images as plus, pre-plus, or neither; using a scale with finer gradations (eg, 0-9) may have highlighted subtler diagnostic distinctions. Third, the effects of image quality on ROP tool measurement remain to be clarified; when we isolated images on which ROP tool performed less well, there were no clear similarities between them with regard to image quality. Fourth, because our primary objective was evaluation of the accuracy of ROP tool, we did not compare consensus of experts viewing images with results of indirect ophthalmoscopy; such an analysis will someday have a role in justifying the use of telemedicine screening.

In conclusion, ROP tool version 2.1.5 accurately measures tortuosity and dilation of posterior pole blood vessels in RetCam images, corresponding well with expert diagnostic categories of normal, pre-plus, and plus disease. In addition, we observed trends in tortuosity and dilation indexes from one expert category to the next that may provide insight into human (not just computer-assisted) ROP diagnosis. Going forward, we plan to investigate the best method to mathematically combine vessel width and tortuosity, which will greatly facilitate bedside or telemedicine diagnosis of plus disease.

Submitted for Publication: August 16, 2009; final revision received November 9, 2009; accepted November 24, 2009.

Correspondence: David K. Wallace, MD, MPH, Department of Ophthalmology, Duke University Medical Center, Office 3802, Durham, NC 27710 (david.wallace@duke.edu).

Author Contributions: Drs Kiely, Wallace, and Freedman had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the analysis.

Financial Disclosure: None reported.

Funding/Sponsor: This study was supported by grant K23 EY015806 from the National Eye Institute (Dr Wallace).
Additional Contributions: Sandra Stinnett, PhD, provided statistical assistance. Michael Chiang, MD, MA, Antonio Capone, MD, and the PHOTO-ROP (Photographic Screening for Retinopathy of Prematurity) study group shared RetCam photographs. Colleagues who comprised the panels of experts and examiners were Graham Quinn, MD, MSCE, Michael Chiang, MD, MA, Terri Young, MD, and Laura Enyedi, MD.

REFERENCES