Published in final edited form as:
Retina. 2010 June ; 30(6): 958–965. doi:10.1097/IAE.0b013e3181c9696a.

Accuracy of Retinopathy of Prematurity Diagnosis by Retinal Fellows

R. V. Paul Chan, MD[*], **Steven L. Williams, BA**[†], **Yoshihiro Yonekawa, BA**[*], **David J. Weissgold, MD**[‡], **Thomas C. Lee, MD**[§], and **Michael F. Chiang, MD, MA**[†,**]

[*]Department of Ophthalmology, Weill Cornell Medical College, New York, New York

[†]Department of Ophthalmology, Columbia University College of Physicians and Surgeons, New York, New York

[**]Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons, New York, New York

[‡]Retina Center of Vermont, Burlington, Vermont

[§]Department of Ophthalmology, Children's Hospital Los Angeles, Los Angeles, California

Abstract

Purpose—To measure the accuracy of retinopathy of prematurity (ROP) diagnosis by retinal fellows.

Methods—An atlas of 804 retinal images was captured from 248 eyes of 67 premature infants with a wide-angle camera (RetCam-II; Clarity Medical Systems, Pleasanton, CA). Images were uploaded to a study website, from which an expert pediatric retinal specialist and seven retinal fellows independently provided a diagnosis (no ROP, mild ROP, type-2 ROP, or treatment-requiring ROP) for each eye. Sensitivity and specificity of each retinal fellow was calculated, compared to a reference standard of diagnosis by the expert pediatric retinal specialist.

Results—For detection of type-2 or worse ROP by fellows, mean (range) sensitivity was 0.751 (0.512–0.953) and specificity was 0.841 (0.707–0.976). For detection of treatment-requiring ROP, mean (range) sensitivity was 0.914 (0.667–1.000) and specificity was 0.871 (0.678–0.987).

Conclusions—In general, fellows demonstrated high accuracy for detecting ROP. However, 3/7 fellows achieved less than 80% sensitivity for diagnosis of type-2 or worse ROP, and 2/7 achieved less than 90% sensitivity for diagnosis of treatment-requiring ROP. This could lead to under-management and under-treatment of clinically significant disease, and raises potential concerns about the quality of ROP screening performed by less experienced examiners.

Corresponding author and reprints: Michael F. Chiang, MD, MA Columbia University College of Physicians and Surgeons 635 West 165th Street, Box 92 New York, NY 10032 Tel: 212-342-3440.

Summary Statement: Using an atlas of 804 retinal images, we compared the responses by an expert pediatric retinal specialist and seven retinal fellows in evaluating retinopathy of prematurity for each image. Fellows demonstrated high diagnostic accuracy for detecting type-2 ROP or treatment-requiring ROP. However, overall diagnostic performance among retinal fellows was inconsistent.

Keywords

Medical education; medical informatics; pediatric ophthalmology; retina; retinal imaging; retinopathy of prematurity

INTRODUCTION

Retinopathy of prematurity (ROP) is a vaso-proliferative disease of the developing retina that continues to be a leading cause of blindness in the pediatric population.[1] Of the approximately 4 million babies born in the United States each year,[2] about 12,000–14,000 infants will develop some degree of ROP, of which 400–900 will develop ROP-related blindness.[3–5] The Cryotherapy for Retinopathy of Prematurity (CRYO-ROP) and Early Treatment for Retinopathy of Prematurity (ETROP) studies provided information regarding the natural history of ROP and have validated treatment criteria for the disease.[5–6] A well-defined international classification system has also been established for ROP.[7,8] However, despite advances in our understanding of ROP, the disease continues to be a clinical challenge.[9–11]

Appropriate screening for ROP is critical to ensuring optimal disease management. A published guideline by the American Academy of Ophthalmology (AAO), American Academy of Pediatrics, and American Association for Pediatric Ophthalmology and Strabismus recommends that infants of birth weight <1500 grams or gestational age ≤30 weeks, and infants with birth weight 1500–2000 grams or gestational age >30 weeks with an unstable clinical course, should receive dilated examination by an experienced ophthalmologist.[12,13] The aim of these guidelines is to detect all cases of clinically-significant ROP (e.g. type-2 or worse), which benefit from close monitoring and/or treatment.

However, the number of infants at risk for ROP is increasing while the supply of available ophthalmologists who manage ROP is decreasing. The annual premature birth rate is growing in the United States and throughout the world as neonatal survival rates are improving.[14] Meanwhile, a 2006 AAO survey showed that only half of retinal specialists and pediatric ophthalmologists are willing to manage ROP, and that over 20% plan to stop because of concerns such as logistical difficulty and medicolegal liability.[15]

These trends have placed a greater burden on ophthalmologists who continue to screen for and manage ROP, as well as the healthcare system. To ease this burden, ophthalmologists without pediatric or retinal subspecialty training may be charged with screening and treatment of ROP.[16] Some institutions may utilize fellows or other trainees as primary screeners. These practices may be a significant clinical and public health concern. In fact, a review of medical malpractice claims regarding ROP showed that a significant issue regarding ROP care was failure of the ophthalmologist to recommend appropriate follow up.[17,18] Although it has been recognized that there may be differences in how examining ophthalmologists diagnose and manage ROP,[19] there is little published data that evaluates ROP diagnostic performance by ophthalmologists with varying training experience.

The purpose of this study is to measure the accuracy and reliability of ROP diagnosis by a group of retinal fellows, as compared to that of an experienced pediatric retinal specialist. Web-based presentations of wide-angle retinal images are used to simulate ophthalmoscopic examination and standardize data presentation to all participants.

METHODS

This study was approved by the Institutional Review Boards of Columbia University Medical Center and Weill Cornell Medical College, and included a waiver of consent for use of de-identified retinal images.

Diagnostic Evaluation of Infants

Infant ROP examinations were standardized by displaying wide-angle retinal images on a secure website for all study participants. Images were captured during routine ROP examinations during 2004–2005 at Columbia University, from consecutive infants whose parents provided informed consent for photography. The acquisition of this data set has been described previously.[20] Retinal photographs were obtained using a commercially-available camera (RetCam-II; Clarity Medical Systems, Pleasanton, CA). Posterior, temporal, and nasal images of each eye were captured, along with up to 2 additional images per eye if they were felt by the photographer to contribute additional diagnostic value.

Examination data were presented to study subjects using a web-based system that displayed all images from the right and left eyes simultaneously. Information was also provided about the infant's birth weight, gestational age at birth, and post-menstrual age at time of examination. Overall, 124 sets of bilateral retinal examinations from 67 infants were displayed. Among these 124 examinations, 21 examinations (42 eyes) were randomly selected by the system to be repeated for assessment of intra-physician reliability. Of the 67 infants, 19 (28%) were white, 14 (21%) were black non-Hispanic, 28 (42%) were Hispanic, and 6 (9%) were Asian.

Subjects

Study subjects consisted of 7 residency-trained, board-eligible ophthalmologists who were enrolled in retinal fellowship programs (1 second-year fellow, 6 first-year fellows), each of whom provided informed consent for participation. The first-year fellows enrolled in the study had minimal or no ROP screening experience during residency training. All were within the first 6 months of their fellowship training at programs where fellows performed weekly ROP examinations with a faculty member.

Fellows were oriented to the diagnostic levels of ROP used in this study with a 1-page guide, which was developed by the authors. All subjects were asked to diagnose each eye, using a four-level system: (1) No ROP; (2) Mild ROP, defined as ROP less than type-2 disease; (3) Type-2 ROP, defined as (a) zone 1, stage 1 or 2, without plus disease or (b) zone 2, stage 3, without plus disease; or (4) Treatment-requiring ROP, defined as (a) type-1 ROP (zone 1, any stage, with plus disease, or zone 1, stage 3, without plus disease, or zone 2, stage 2 or 3, with plus disease) or (b) threshold ROP (at least 5 contiguous or 8 non-contiguous clock hours of stage 3 in zone 1 or 2, with plus disease). Respondents had the option to provide a diagnosis of “Unknown”, if they were uncomfortable making a diagnosis based on the data provided.

The web-based images evaluated by the seven retinal fellows were also reviewed by an expert pediatric retinal specialist (TCL) who has over 10 years of experience in ROP examination and treatment at tertiary care centers, has served as a principal investigator in the ETROP study, and has previously published several peer-reviewed papers involving ROP.

Analysis

Web-based diagnostic evaluations by the expert pediatric retinal specialist (TCL) were used as the reference standard exam diagnosis. Sensitivity, specificity, and area under the curve (AUC) of receiver operating characteristic plots[21] were calculated for detection of mild or worse ROP, type-2 or worse ROP, and treatment-requiring ROP by each fellow. For comparison, this analysis was also performed for two experienced retinal specialists (DJW, RVPC).

Sensitivity and specificity were analyzed to determine if physician accuracy improved or worsened significantly as they performed more diagnostic examinations. Unknown diagnoses were considered to be incorrect responses during data analysis. At each cutoff value, logistic regression was used to obtain the odds ratio of having a "positive" diagnosis as more diagnoses were performed. For cases of no disease, odds ratio <1 would mean that "positive" responses were less likely as more diagnoses were performed (i.e. improved specificity). For cases of disease, odds ratio >1 would mean that "positive" responses were more likely as more diagnoses were performed (i.e. improved sensitivity). Finally, generalized estimating equations were used to determine whether there was any overall trend among all graders.

Intra-physician reliability was determined using the κ statistic for chance-adjusted agreement in diagnosis. A well-known scale was used for interpretation of results: 0 to 0.20 = slight agreement, 0.21 to 0.40 = fair agreement, 0.41 to 0.60 = moderate agreement, 0.61 to 0.80 = substantial agreement, and 0.81 to 1.00 = almost-perfect agreement.[22,23]

Analysis was performed using statistical software (Excel 2003, Microsoft, Redmond, WA; SPSS 15.0, SPSS Inc, Chicago, IL). Statistical significance was considered to be a 2-sided P value <0.05.

RESULTS

Distribution of Image-Based Physician Diagnoses

Figure 1 reports the distribution of diagnostic responses of 7 retinal fellows. For example, when responses by all fellows were considered together, the 28 eyes with a reference standard diagnosis of type-2 ROP were diagnosed as treatment-requiring ROP in 47% of responses (Figure 1C). When responses by fellows were considered individually, the 28 eyes with a reference standard diagnosis of type-2 ROP were diagnosed as treatment-requiring ROP in 3/28 to 21/28 eyes. When responses by all fellows were considered together, the 15 eyes with a reference standard diagnosis of treatment-requiring ROP, were diagnosed as treatment-requiring ROP in 91% of responses (Figure 1D). When responses by fellows were considered individually, the 15 eyes with a reference standard diagnosis of treatment-requiring ROP were diagnosed as treatment-requiring ROP in 10/15 to 15/15 eyes.

Accuracy and Reliability of Image-Based Diagnoses by Physicians

Table 1 reports the sensitivity, specificity and ROC area under the curve (AUC) of 7 retinal fellows at three diagnostic levels. For detection of type-2 or worse ROP among the 7 fellows, the mean (range) sensitivity was 0.751 (0.512–0.953), specificity was 0.841 (0.707–0.976), and AUC was 0.897 (0.785–0.889). For detection of treatment-requiring ROP among the 7 fellows, mean (range) sensitivity was 0.914 (0.667–1.000), specificity was 0.871 (0.678–0.987), and AUC was 0.930 (0.786–0.960). Figure 2 demonstrates images from eyes that were frequently misdiagnosed by fellows.

For detection of type-2 or worse ROP among the 2 additional retinal specialists (DJW, RVPC), the mean (range) sensitivity was 0.884 (0.837–0.930), specificity was 0.885 (0.839–0.932), and AUC was 0.885 (0.884–0.885). For detection of treatment-requiring ROP among the 2 additional retinal specialists, mean (range) sensitivity was 1.000 (1.000–1.000), specificity was 0.908 (0.888–0.927), and AUC was 0.954 (0.944–0.964).

Logistic regression and generalized estimation equation analysis showed the sensitivity for retinal fellows increased as more diagnoses were performed for detection of mild or worse (p<0.0002), type-2 or worse (p=0.001), and treatment-requiring ROP (p<0.0001). The specificity for retinal fellows decreased as more diagnoses were performed for detection of type-2 or worse ROP (p=0.02) and treatment-requiring ROP (p=0.01). For comparison, the sensitivity of the 2 retinal specialists (DJW, RVPC) increased as more diagnoses were performed for detection of type-2 or worse ROP (p=0.0007) and treatment-requiring ROP (p=0.04). The specificity for retinal specialists decreased as more diagnoses were performed for detection of type-2 or worse ROP (p=0.04).

Table 2 reports intra-physician agreement at three diagnostic levels of each fellow, when unknown responses are excluded. For detection of type-2 or worse ROP, 3/7 (43%) fellows had almost-perfect intra-physician agreement, and 4/7 (57%) fellows had substantial agreement. For detection of treatment-requiring ROP, 5/7 (71%) fellows had almost-perfect agreement, 1/7 (14%) fellows had substantial agreement, and 1/7 (14%) fellows had fair agreement.

DISCUSSION

This study evaluates the performance of image-based ROP diagnosis by 7 retinal fellows compared to a reference standard of diagnosis by an expert pediatric retinal specialist. The key finding is that overall diagnostic performance in diagnosing clinically-significant disease (type-2 or worse) was inconsistent, although fellows demonstrated high accuracy for detecting treatment-requiring ROP.

The variability in diagnostic accuracy among fellows in this study is important from a clinical perspective. In cases of treatment-requiring disease (i.e. type-1 or worse ROP), laser photocoagulation or other interventions should be done within 48–72 hours of detection.[6,13] In cases of type-2 prethreshold disease, infants should be monitored closely for progression to treatment-requiring disease.6,13 Fellows in this study demonstrated inconsistent accuracy for detection of type-2 ROP and treatment-requiring ROP. For diagnosis of type-2 or worse ROP, 3 of 7 fellows achieved less than 80% sensitivity (Table 1). Further analysis of type-2 prethreshold cases reveals that fellows tended to over-call disease (treatment-requiring ROP) in 47% of eyes and under-call disease (mild or no ROP) in 36% of eyes (Figure 1C). For the diagnosis of treatment-requiring ROP, 2 of 7 achieved less than 90% sensitivity (Table 1). Of the missed cases of treatment-requiring ROP, fellows diagnosed roughly 5% as type-2 prethreshold disease (Figure 1D). This raises potential concerns because over-diagnosis may be associated with unnecessary examination or treatment, while under-diagnosis may be associated with suboptimal management of potentially blinding disease.

Furthermore, there is little consensus as to who is “qualified” to manage ROP. Kemper et al. found that 11% of all ophthalmologists examine for and 6% treat ROP.[16] Of those who screened for ROP, there were an equal number of physicians who did not complete any fellowship training as compared to those who completed fellowship training in pediatric ophthalmology and strabismus. Nine percent of ophthalmologists who perform ROP examinations reported that their training did not adequately prepare them, regardless of whether or not they were fellowship-trained.[16] Examiners with less formal training than retinal specialists and pediatric ophthalmologists are presumably performing ROP surveillance in response to workforce pressures.

Two factors should be considered while interpreting the results of this study: (1) Diagnosis was assessed by interpretation of a standard set of retinal images. Therefore, the study findings may not necessarily be generalizable to indirect ophthalmoscopy at the bedside, particularly because study graders may have had different levels of expertise in reviewing wide-angle retinal images. Since images were uniformly presented to all study participants, any potential bias that this may have caused was applied equally among all graders. Conversely, retinal imaging may facilitate visualization of the peripheral retina, which might decrease diagnostic variability among graders with less experience in evaluating ROP by indirect ophthalmoscopy. In fact, because of differences in ophthalmoscopic examination technique and infant cooperation, image-based examination in this study might be expected to cause less variability than that seen with ophthalmoscopy. Although it could be argued that an alternative study design involving serial ophthalmoscopic examinations by multiple examiners would be more realistic, ROP examinations by indirect ophthalmoscopy can be stressful for the neonate, resulting in adverse events such as pain, bradycardia, and oxygen desaturation. Therefore, we feel that a study involving serial ophthalmoscopic examinations by retinal fellows may be impractical because of concerns about infant safety. (2) Studies have demonstrated that there is disagreement in diagnosis of severe ROP among experts, even when reviewing the exact same images.[19] In addition, the CRYO-ROP trial found disagreement between two unmasked, certified examiners as to whether threshold disease was present in 12% of eyes.5 Similarly, studies have suggested that there can be considerable variation in the diagnosis of severe ROP between different centers, even after correcting for case-mix and sampling variability.24›25 Given all of these issues, we examined the diagnostic variability among fellowship trained examiners by measuring the face validity of the approach used in this study by measuring the diagnostic accuracy of 2 additional retinal specialists. The mean sensitivity and specificity for detection of both type-2 or worse ROP and treatment-requiring ROP by these two retinal specialists was higher than that of the 7 retinal fellows. This supports the validity of our study methods by showing that experienced retinal specialists performed better than fellows for diagnosis of type-2 or worse ROP. Although these findings reinforce the importance of training and experience for accurate ROP diagnosis, future studies involving larger numbers of experts may be warranted.

To identify reasons for incorrect responses by fellows, the most commonly misdiagnosed images were reviewed retrospectively to determine the most likely source of error based on author consensus (RVPC, SLW, TCL, MFC) (Table 3). Among 248 total image sets, 57 (23%) were diagnosed correctly by <50% of fellows. Of these 57 image sets, 41 (72%) had errors in identification of stage, 14 (25%) had errors in identification of plus disease, 8 (14%) had errors in identification of zone, and 5 (9%) had suboptimal image quality. Recognition of these frequent sources of error may help guide education and training programs. However, it is important to note that this study was not specifically designed to elucidate the discrepancies between the experienced retinal specialist and the fellows. Further studies designed to precisely define reasons for error may be warranted.

In addition to the issues regarding indirect ophthalmoscopy versus interpretation of retinal photographs, other study limitations include: (1) Subjects may have had differences in exposure to ROP examinations and diagnosis during their residency and fellowship training, prior to the inception of this study. For example, this study included 1 second-year fellow and 6 first-year fellows. This variation in experience may be similar to that of the larger community of general ophthalmologists who perform ROP examination. Although there was no clear difference between the diagnostic performance of the second-year fellow and the group of first-year fellows in this study, additional research involving the influence of ROP exposure or image-based ROP diagnosis may be warranted. (2) Two co-authors (RVPC, TCL) are also responsible for the ROP training of all participating fellows. This may influence how ROP is diagnosed, as fellows may adopt similar diagnostic biases as the authors. This may make the study conclusions less generalizable to other ophthalmologists. However, this bias might be expected to result in diagnostic responses that are more similar, rather than more divergent.

We believe that this is the most extensive study of ROP diagnosis among ophthalmologists who have not yet completed fellowship training. Our results suggest that there is variability in the diagnosis of clinically-significant ROP (type 2 or worse) among individual fellows, although they generally achieved high accuracy for diagnosis of treatment-requiring ROP. This has implications for clinical ROP management and raises potential concerns about the quality of ROP screening performed by inexperienced examiners. These findings suggest a need for more formalized ROP training protocols.

Acknowledgments

This study was supported by the St. Giles Foundation, New York, NY (RVPC), by the Starr Foundation, New York, NY (SLW), by the Doris Duke Charitable Foundation, New York, NY (SLW), by a Career Development Award (MFC) and Institutional Grant (RVPC) from Research to Prevent Blindness, New York, NY, and by grant EY13972 from the National Institutes of Health, Bethesda, MD (MFC). The authors have no commercial, proprietary, or financial interest in any of the products or companies described in this article. MFC is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems (Pleasanton, CA).

REFERENCES

1. Flynn JT, Bancalari E, Bachynski BN, et al. Retinopathy of prematurity. Diagnosis, severity, and natural history. Ophthalmology 1987;94:620–629. [PubMed: 3627710]
2. Martin JA, Hamilton BE, Sutton PD, et al. Births: final data for 2005. Natl Vital Stat Rep 2007;56:1–103. [PubMed: 18277471]
3. National Eye Institute. Retinopathy of Prematurity. [Accessed May 16, 2008]. Available at: http://www.nei.nih.gov/health/rop/.
4. Phelps DL. Retinopathy of prematurity: an estimate of vision loss in the United State—1979. Pediatrics 1981;67:924–925. [PubMed: 6894488]
5. Multicenter trial of cryotherapy for retinopathy of prematurity. Preliminary results. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Arch Ophthalmol 1988;106:471–479. [PubMed: 2895630]
6. Early Treatment For Retinopathy of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. Arch Ophthalmol 2003;121:1684–1694. [PubMed: 14662586]
7. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. Arch Ophthalmol 1984;102:1130–1134. [PubMed: 6547831]
8. An International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. Arch Ophthalmol 2005;123:991–999. [PubMed: 16009843]
9. Gilbert C, Foster A. Childhood blindness in the context of VISION 2020--the right to sight. Bull World Health Organ 2001;79(3):227–232. Epub 2003 Jul 7. [PubMed: 11285667]
10. Munoz B, West SK. Blindness and visual impairment in the Americas and the Caribbean. Br J Ophthalmol 2002;86:498–504. [PubMed: 11973241]
11. Steinkuller PG, Du L, Gilbert C, et al. Childhood blindness. J AAPOS 1999;3:26–32. [PubMed: 10071898]
12. American Academy of Pediatrics; Section on Ophthalmology. Screening examination of premature infants for retinopathy of prematurity. Pediatrics 2001;108:809–811. [PubMed: 11533356]
13. Section on Ophthalmology; American Academy of Pediatrics; American Academy of Ophthalmology and American Association for Pediatric Ophthalmology and Strabismus. Screening examination of premature infants for retinopathy of prematurity. Pediatrics 2006;117:572–576. [PubMed: 16452383]
14. Flynn JT, Chan-Ling T. Retinopathy of prematurity: two distinct mechanisms that underlie zone 1 and zone 2 disease. Am J Ophthalmol 2006;142:46–59. [PubMed: 16815250]
15. American Academy of Ophthalmology. Ophthalmologists Warn of Shortage in Specialists Who Treat Premature Babies with Blinding Eye Condition. [Accessed October 4, 2008]. Available at: http://www.aao.org/newsroom/release/20060713.cfm.
16. Kemper AR, Freedman SF, Wallace DK. Retinopathy of prematurity care: patterns of care and workforce analysis. J AAPOS 2008;12:344–348. [PubMed: 18440256]

17. Day S, Menke AM, Abbott RL. Retinopathy of prematurity malpractice claims: the Ophthalmic Mutual Insurance Company experience. Arch Ophthalmol Jun;2009 127(6):794–8. [PubMed: 19506200]

18. Mills MD. Retinopathy of prematurity malpractice claims. Arch Ophthalmol Jun;2009 127(6):803–4. [PubMed: 19506203]

19. Chiang MF, Jiang L, Gelman R, et al. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol 2007;125:875–880. [PubMed: 17620564]

20. Chiang MF, Wang L, Busuioc M, et al. Telemedical retinopathy of prematurity diagnosis: accuracy, reliability, and image quality. Arch Ophthalmol 2007;125:1531–1538. [PubMed: 17998515]

21. Chiang MF, Starren J, Du YE, et al. Remote image based retinopathy of prematurity diagnosis: a receiver operating characteristic analysis of accuracy. Br J Ophthalmol 2006;90:1292–1296. [PubMed: 16613919]

22. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. Educational and Psychological Measurement 1960;20:37–46.

23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174. [PubMed: 843571]

24. Darlow BA, Hutchinson JL, Simpson JM, et al. Variation in rates of severe retinopathy of prematurity among neonatal intensive care units in the Australian and New Zealand Neonatal Network. Br J Ophthalmol 2005;89:1592–1596. [PubMed: 16299138]

25. Darlow BA, Elder MJ, Horwood LJ, et al. Does observer bias contribute to variations in the rate of retinopathy of prematurity between centres? Clin Experiment Ophthalmol 2008;36:43–46. [PubMed: 18290953]

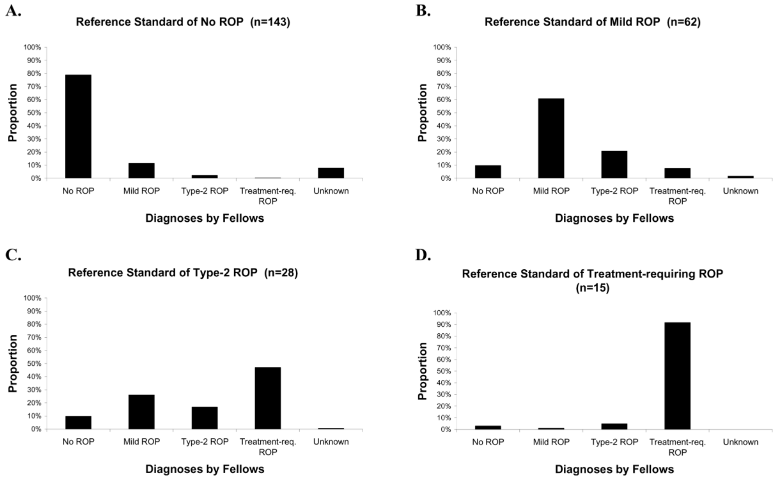
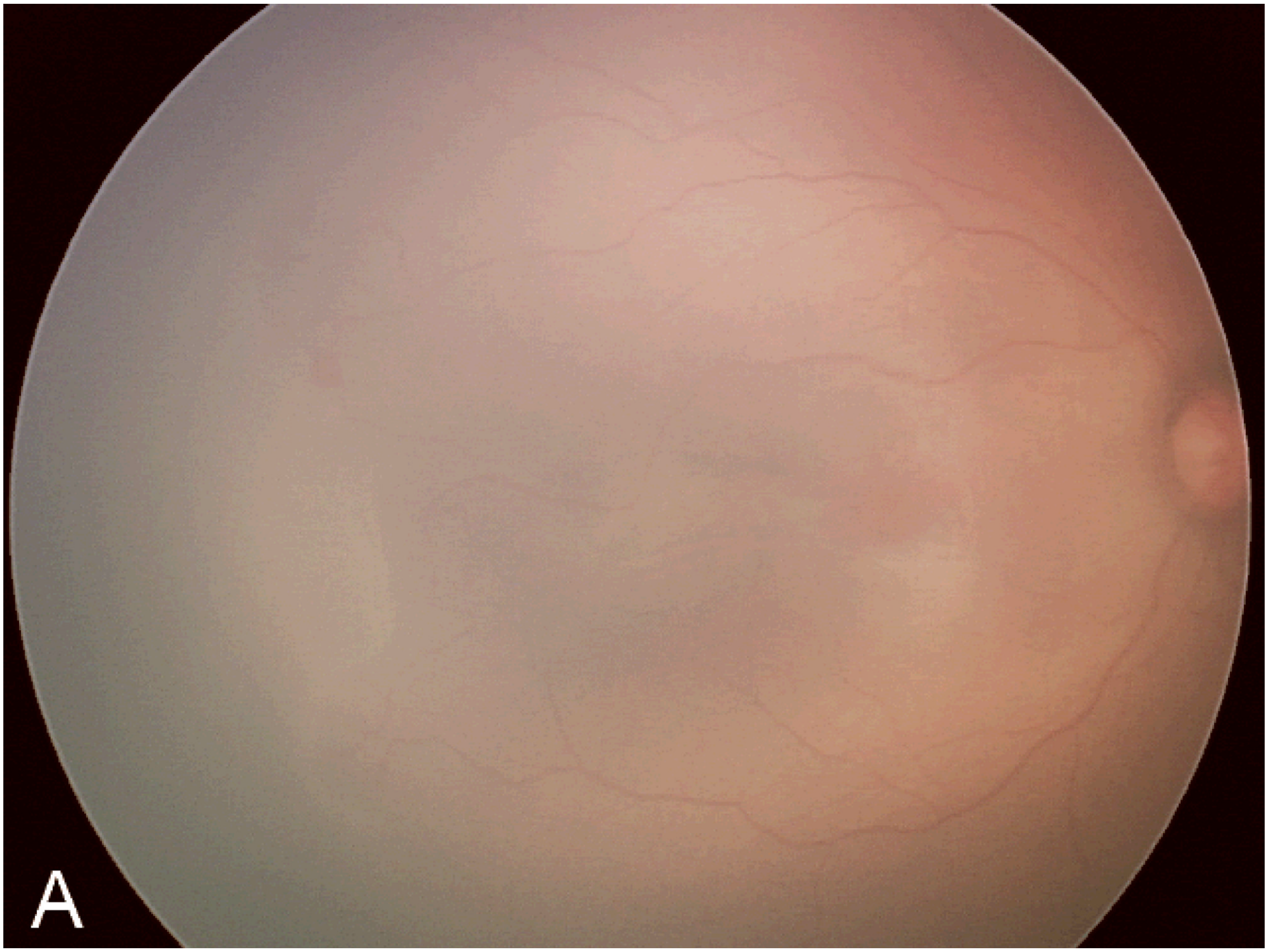
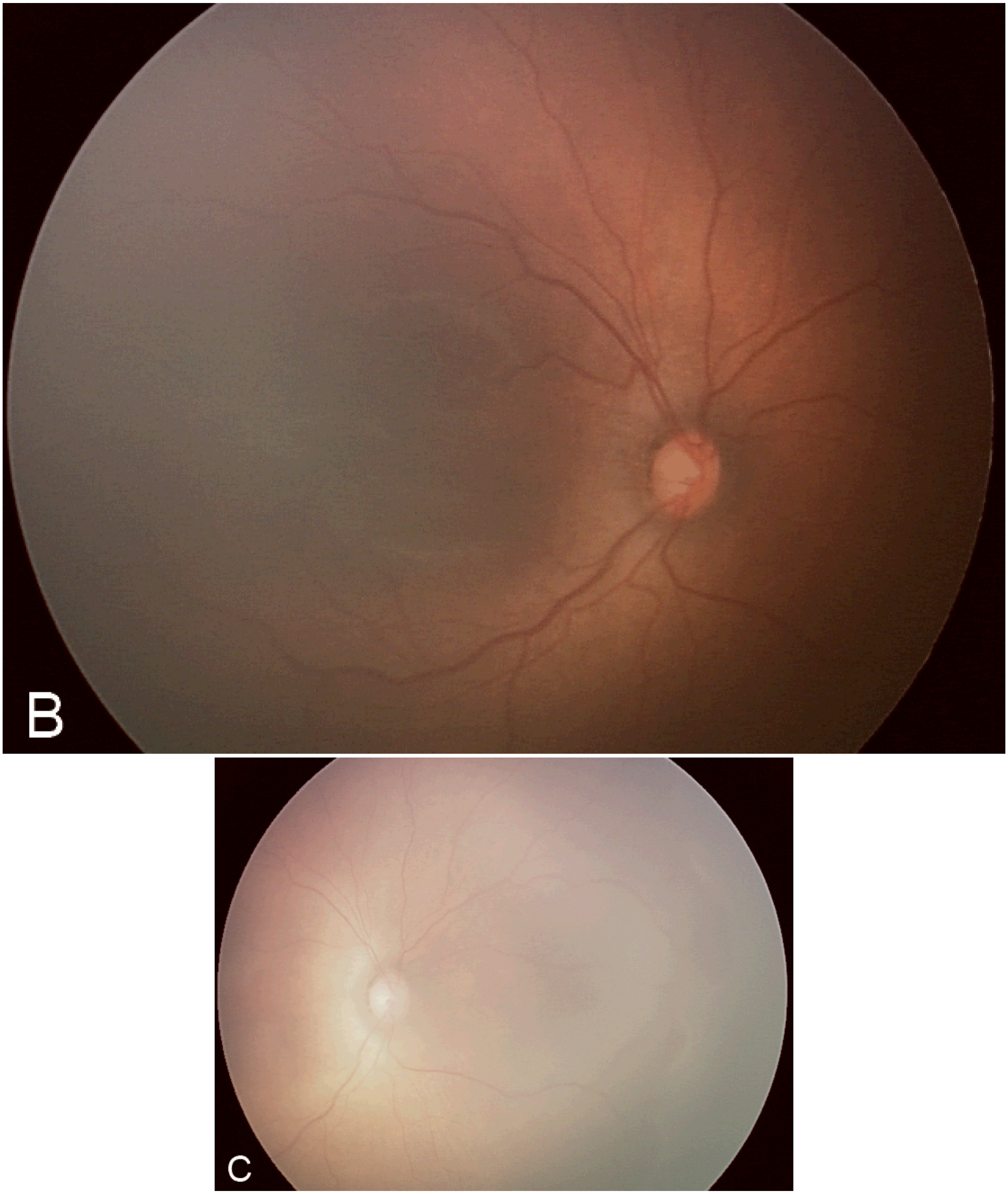
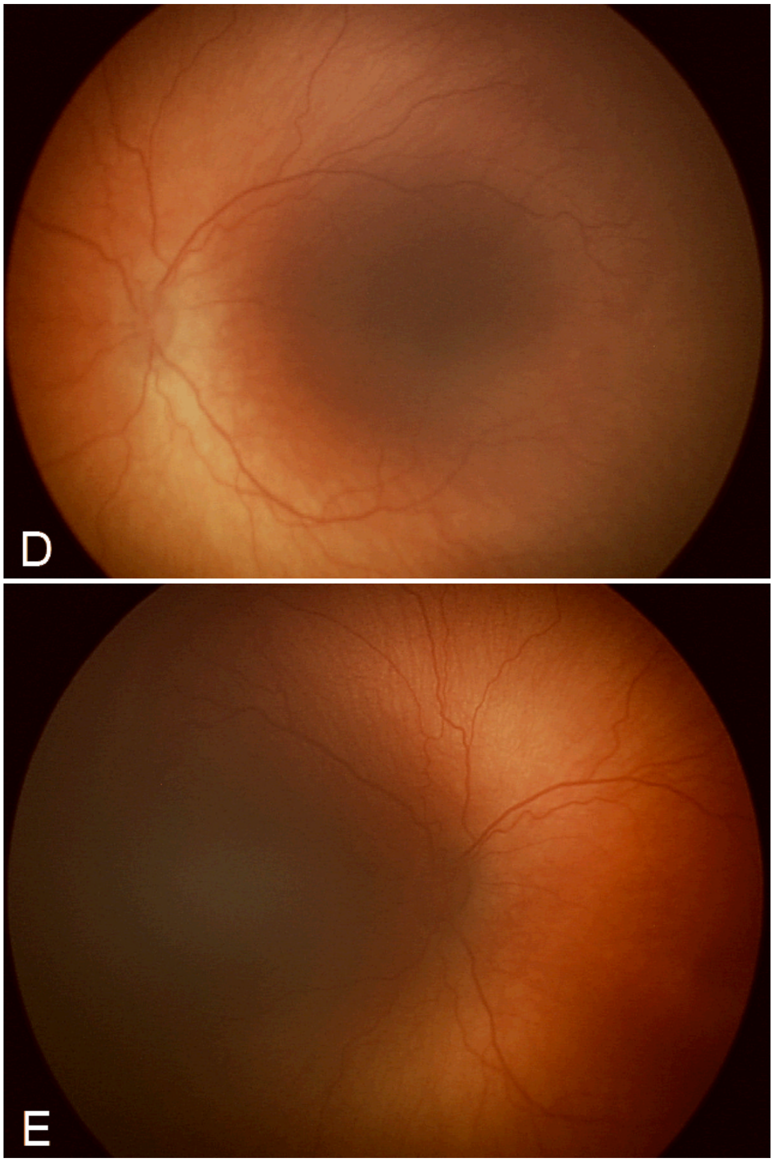


Figure 1. Mean distribution of ROP diagnoses by 7 retinal fellows

(A) Among the 143 eyes with a reference standard diagnosis of no ROP, the 7 fellows diagnosed no ROP in 113 (79%) eyes. **(B)** Among the 62 eyes with a reference standard diagnosis of mild ROP, the 7 fellows diagnosed mild ROP in 38 (61%) eyes. **(C)** Among the 28 eyes with a reference standard diagnosis of type-2 prethreshold ROP, the 7 fellows diagnosed type-2 prethreshold ROP in 5 (17%) eyes. **(D)** Among the 15 eyes with a reference standard diagnosis of treatment-requiring ROP, the 7 fellows diagnosed treatment-requiring ROP in 14 (91%) eyes.


A


B
C


D
E

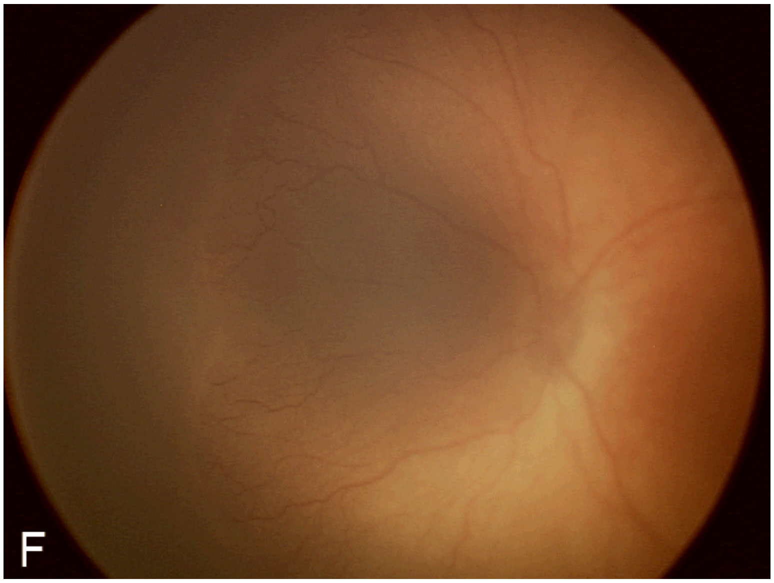


Figure 2. Examples of study images that were frequently misdiagnosed by retinal fellows

(A), (B), and (C) display temporal, posterior, and nasal images from an infant diagnosed as mild ROP by reference standard exam diagnosis and was diagnosed as type-2 ROP by 5/7 (71%) fellows, and as treatment-requiring ROP by 2/7 (29%) fellows. **(D), (E), and (F)** display temporal, posterior, and nasal images from an infant diagnosed as type-2 ROP by reference standard exam diagnosis and was diagnosed as no ROP by 2/7 (29%) fellows, as type-2 ROP by 1/7 (14%) fellows, and as treatment-requiring ROP by 4/7 (57%) fellows.

Table 1

Sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) for ROP diagnosis by 7 retinal fellows.[*]

Responses are compared to a reference standard of diagnosis by an expert pediatric retinal specialist. Analysis is performed by eye for diagnosis of mild or worse ROP, type-2 or worse ROP, and treatment-requiring ROP. Results are displayed as proportion (standard error), and unknown diagnoses by graders are considered incorrect responses.[†]

	Mild or worse ROP			Type-2 or worse ROP			Treatment-requiring ROP		
Physician	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**
Fellow 1	0.952 (0.021)	0.727 (0.037)	0.840 (0.026)	0.884 (0.049)	0.727 (0.031)	0.805 (0.034)	1.000 (…)	0.914 (0.018)	0.957 (0.013)
Fellow 2	0.914 (0.027)	0.608 (0.041)	0.785 (0.029)	0.907 (0.044)	0.707 (0.032)	0.819 (0.031)	1.000 (…)	0.678 (0.031)	0.839 (0.031)
Fellow 3	0.867 (0.033)	0.902 (0.025)	0.884 (0.024)	0.512 (0.076)	0.976 (0.011)	0.744 (0.050)	0.933 (0.064)	0.987 (0.007)	0.960 (0.038)
Fellow 4	0.933 (0.024)	0.720 (0.038)	0.827 (0.027)	0.814 (0.059)	0.937 (0.017)	0.875 (0.036)	1.000 (…)	0.893 (0.020)	0.946 (0.014)
Fellow 5	0.838 (0.035)	0.930 (0.021)	0.889 (0.024)	0.651 (0.073)	0.946 (0.016)	0.799 (0.045)	0.867 (0.088)	0.910 (0.019)	0.888 (0.052)
Fellow 6	0.838 (0.056)	0.832 (0.031)	0.845 (0.027)	0.535 (0.076)	0.883 (0.022)	0.709 (0.049)	0.667 (0.122)	0.906 (0.192)	0.786 (0.074)
Fellow 7	0.971 (0.016)	0.797 (0.034)	0.884 (0.023)	0.953 (0.032)	0.712 (0.032)	0.833 (0.029)	0.933 (0.064)	0.811 (0.026)	0.872 (0.041)

[*] Standard error values are reported for each sensitivity, specificity and AUC value.

[†] Unknown diagnoses were provided for 0 (0%) eyes by fellow 1, 41 (17%) eyes by fellow 2, 0 (0%) eyes by fellow 3, 7 (3%) eyes by fellow 4, 7 (3%) eyes by fellow 5, 19 (8%) eyes by fellow 6, and 0 (0%) eyes by fellow 7.

Table 2

Intra-physician reliability for ROP diagnosis by one pediatric retinal specialist (reference standard) and seven retinal fellows

Results are displayed as kappa statistic (standard error) for ability to detect mild or worse ROP, type-2 or worse ROP, and treatment-requiring ROP.

Physician	Mild or worse ROP	Type-2 or worse ROP	Treatment-requiring ROP
Reference Standard	0.904 (0.066)	1.000 (--)	0.730 (0.145)
Fellow 1	0.900 (0.071)	0.615 (0.120)	1.000 (--)
Fellow 2	1.000 (--)	0.935 (0.064)	0.935 (0.064)
Fellow 3	0.855 (0.080)	1.000 (--)	1.000 (--)
Fellow 4	0.854 (0.099)	0.882 (0.081)	1.000 (--)
Fellow 5	0.787 (0.099)	0.730 (0.126)	0.713 (0.130)
Fellow 6	0.876 (0.084)	0.656 (0.140)	0.231 (0.141)
Fellow 7	0.801 (0.110)	0.771 (0.107)	0.942 (0.057)

Table 3
Reasons for discrepancy between the expert retina specialist and retina fellows

Reasons for discrepancy	Number of images[*]	Percentage (%)
Identification of stage	41	71.9
Identification of plus disease	23	40.4
Identification of zone	8	14.0
Poor image quality	5	8.8
Total number of images	57[†]	

[*] Number of images that were diagnosed correctly by <50% of fellows. The sum is larger than the total number of images because some images were diagnosed incorrectly by fellows for different reasons.

[†] Total number of images that were diagnosed correctly by <50% of fellows.