Inter- and Intra-Observer Reliability Among Retinopathy of Prematurity (ROP) Screeners

Kristine Corpus, MD,[1] Jubaida Aquino, MD,[1,2] Macario Reandelar Jr., MD, MSPH,[1,3] Retinopathy of Prematurity Working Group

[1]St. Luke's Medical Center, E. Rodriguez Sr. Avenue, Quezon City
[2]East Avenue Medical Center, East Avenue, Quezon City
[3]Far Eastern University-Nicanor Reyes Medical Foundation Medical Center, Quezon City, Philippines

Correspondence: Kristine Corpus, M.D.
International Eye Institute, St. Luke's Medical Center
E. Rodriguez Sr. Avenue, Quezon City, Philippines
Tel. no.: +63-2-7230101 local 5546 or 4143
Email: krstncorpus@yahoo.com

Disclosure: No financial assistance was received for this study. The authors have no proprietary or financial interest in any product used or cited in the study.

ABSTRACT

Objective: (1) To determine the inter and intra-observer reliability in diagnosing ROP in terms of the stage, zone, and presence of plus disease among local ROP screeners involved in the ROP Working Group; and (2) to determine the inter-observer reliability between 2 groups of subspecialties – retina specialists and pediatric ophthalmologists.

Methods: This is a prospective observational study that analyzed the inter- and intra-observer reliability in describing ROP in 3 key observations: stage, zone, and presence of plus disease. This study utilized a test with 32 sets of fundus images from 27 cases, five of which were repeated. Images from previously photographed infants with and without ROP were collated into a downloadable powerpoint test and tested against retina specialists and pediatric ophthalmologists of the ROP Working Group. Outcome measures included presence of variability in ROP diagnosis in terms of the stage, zone, and presence of plus disease among screeners, and reliability coefficient (intra-class coefficient or ICC) in 2 levels: (1) individual and 2-group inter-observer reliability, and (2) intra-observer reliability.

Results: There were 11 respondents: 5 retina specialists and 6 pediatric ophthalmologists. Seven (46%) reported prior experience with RetCam image review. There was high inter-observer reliability (ICC 1.0) in the staging of ROP, but poor reliability in the identification of zone (ICC 0.3) and plus disease (ICC 0.5). The group of retina specialists and pediatric ophthalmologists scored high reliability for diagnosis of stage (ICC 1.0 vs 0.9) and plus disease (ICC 0.9 vs 0.9), while both showed poor reliability in the identification of zone (ICC 0.5 vs 0.4). Majority had high intra-observer reliability with regard to the stage (55%) and zone (73%) of ROP and most (73%) had acceptable intra-observer reliability in identifying plus disease. None of the respondents had poor intra-observer reliability.

Conclusion: The diagnosis of the stage of ROP was consistently reliable for both inter- and intra-observer parameters. However, identification of zone of ROP and plus disease were sources of significant discrepancies.

Keywords: ROP, Retinopathy of prematurity, Screening, Variability, ROP grading

Philipp J Ophthalmol 2013;38:80-85

Retinopathy of prematurity (ROP) is a vaso-proliferative retinal disease of preterm infants which can lead to retinal detachment and severe visual impairment.[1,2] It is a significant and a potentially avoidable cause of childhood blindness.[3] The Philippines is one of the countries with high risk of blindness due to ROP, with an incidence of 9-60 blind infants per 1,000 births resulting from inadequate neonatal care and ROP screening.[4] Prompt screening and treatment have been proven to optimize management and reduce complications of ROP among high-risk babies.[5,6] The financial burden to the society of one child blinded by ROP has been estimated to outweigh the screening and treatment cost.[7,8] Hence, the need for ROP screening that is not only timely but reliable. Although the International Classification of ROP (ICROP) has made the evaluation more objective by providing standard photographs,[9] ROP screening remains to be a subjective exam.[10-15]

In the CRYO-ROP study, 12% of eyes initially diagnosed as threshold disease were diagnosed as less-than-threshold during confirmatory examination by a second screener, within 3 days from each other. [11] There are inter-observer studies that described discrepancy in the diagnosis of plus disease and zone I ROP.[12-15] The goal of an effective ROP screening program is to identify the infants who could benefit from ROP treatment.[1] However, it necessitates screening that is accurate as well as consistent, especially in the identification of treatment-requiring ROP.[11] Inconsistencies in diagnosing can lead to over or under treatment of preterm infants.[11,13]

In this light, this study (1) determined the inter- and intra-observer reliability in diagnosing ROP in terms of the stage, zone, and presence of plus disease among local ROP screeners involved in the ROP Working Group; and (2) determined the inter-observer reliability between 2 groups of subspecialties – retina specialists and pediatric ophthalmologists – within the ROP Working Group.

The ROP Working Group is the initiative of the Philippine Academy of Ophthalmology aimed to decrease the prevalence of visual impairment secondary to ROP. It is a group comprised of retina specialists and pediatric ophthalmologists who are active local ROP screeners and strongly dedicated to ROP advocacy. As each country has been encouraged to create ROP screening criteria appropriate for the babies of their locale,[1] assessment of the reliability of our local ROP experts is also necessary as the competency and quality of screeners and experts vary in different populations. Moreover, no study at present has systematically evaluated reliability in the diagnosis of stage and zone of ROP, as well as the reliability of 2 groups of ROP experts – retina specialists and pediatric ophthalmologists.

METHODOLOGY

Test Fundus Photographs

This is a prospective observational study that utilized a test composed of 32 sets of images from 27 cases, five cases of which were repeated (131 fundus photographs). The fundus images were from 27 eyes of 24 infants, with and without ROP, who were previously photographed using a wide field digital imaging system (RetCam II-III; Clarity Medical Systems, Pleasanton, California, USA) at St. Luke's Medical Center Quezon City and East Avenue Medical Center.

The fundus images collected were part of the routine ROP care, as well as objective ophthalmologic documentation, and none of the infants could be identified from the retinal images. None of the photographs were taken for the sole purpose of this study. These photographs were obtained by their attending ophthalmologists who adhered to the standard protocols and the prerequisite of which entailed parental consent both for the screening exam and the photodocumentation and storage. There were no patient identifiers or annotations as to clinical data, gestational age, birth weight, laterality, or location of disc and lesions.

The primary author, who was not included among the test readers, collected as many high-quality images as possible. Poor quality photographs and those with no official diagnoses from their attending ophthalmologists were excluded. An online picture editor (www.pixlr.com/editor) was used to enhance the images. Enhancement was minimal and was limited to adjustments to exposure, brightness, and contrast. No photographs were cropped. The images were then collated into a downloadable Microsoft Office Powerpoint test. The images were randomly ordered. The five sets of images that were repeated were inverted 180º to avoid recognition of the recurring cases.

Of the 32 sets of images that comprised the downloadable test, 6 (19%) sets of images (Figure 1a) were allotted per stage (stage 1, 2, 3, 4 ROP, and immature retina). APROP and stage 5 had 1 case each (3%). Majority (19 sets of images or 6%) of the images were zone II while there were 4 sets of images for zone III, and 1 image for zone I (Figure 1b). Two sets of images (6%) were plus disease while 4 sets were pre-plus (Figure 1c). The rest (18 sets of images) had no plus. Of the total 32 sets tested, 5 (16%) were repeated for intra-observer testing (Figure 1d).

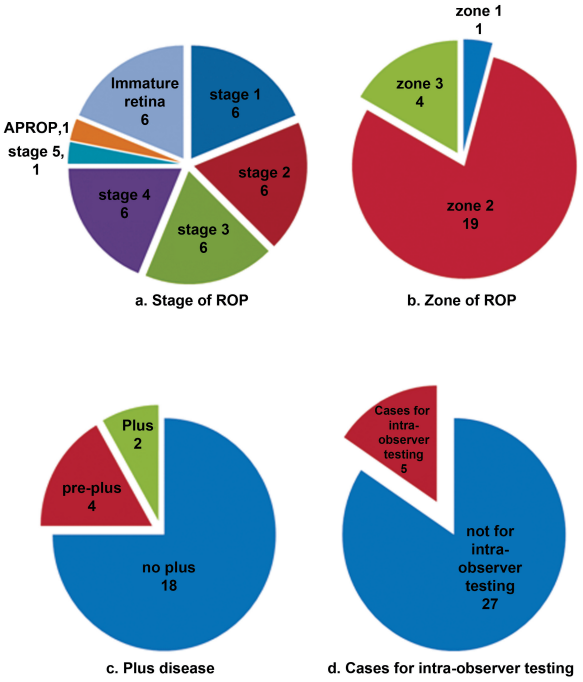

Figure 1. Distribution of images.

Participants

The target sample size included 16 members of the ROP Working Group. It comprised actively practicing retina specialists and pediatric ophthalmologists. They all had years of experiences in ROP screening and treatment, but their experience in RetCam image evaluation were varied. Informed consent was obtained from each participant. They independently downloaded the powerpoint test from a secure website. They were blinded to any information on the infants, except that they knew that cases of ROP and no ROP were included. They independently reviewed the powerpoint test and e-mailed their diagnosis to a secure e-mail address created exclusively for this study.

Image Interpretation

They were asked to evaluate each image, according to 3 key ROP observations based on the ICROP study,[9] each with an ordinal scale: (1) diagnosis; (2) zone; and (3) presence of plus disease. "Diagnosis" included ROP stage 1, 2, 3, 4, 5, aggressive posterior retinopathy of prematurity (APROP), and immature retina. "Zone" referred to as zone I, II, and III while plus disease was subdivided into no plus, pre-plus, and plus disease. Participants were also asked to comment on their confidence rating of the RetCam image review as confident, somewhat confident, not confident. Options in each categorization were exclusive. The diagnosis of the attending ophthalmologist was used as the reference standard. The answers of the participants were compared against (1) each other as a group, (2) within themselves, and (3) as 2 groups (retina specialists and pediatric ophthalmologists).

Outcome Measures & Statistical Analyses

Outcome measures included: (1) presence or absence of variability of ROP diagnosis among participants, (2) over-all inter-observer reliability, (3) intra-observer reliability, and (3) subspecialty inter-observer reliability (between retina specialists and pediatric ophthalmologists). Reliability was evaluated at 3 levels: stage, zone, plus disease.

Reliability pertains to the degree to which a measurement technique can be depended upon to secure consistent results upon repeated application.[17] Descriptive statistics was utilized and reliability was measured using intra-class correlation coefficient (ICC) with a range of values from 0 to 1.0 (highly unreliable to perfect reliability).[17] Acceptable reliability is assigned an ICC value of at least 0.6 (Table 1).[18] Calculations were made using SSPS software for Windows (SSPS Inc, Chicago, Illinois, USA).

Table 1. ICC classification and reliability.

ICC	Reliability
0.5 and below	Poor[17]
0.6-0.8	Acceptable[18]
0.9-1.0	High[17]

RESULTS

Characteristics of Participants

Of the 16 experts invited, 11 participated in the study with 5 retina specialists and 6 pediatric ophthalmologists. Most of the participants (64%) had, at the least, some experience with RetCam image review, while 18% (2) were experienced. Four (36%) participants reported no prior experience.

Reliability

There was high (ICC=1.0) over-all inter-observer reliability with regard to staging of ROP. However, poor reliability was observed in identification of the zone and presence of plus disease, with an ICC of 0.3 and 0.5, respectively (Table 2). There was high inter-observer reliability between retina and pediatric ophthalmology respondents (Table 3) in terms of the staging (1.0 vs 0.9) and diagnosis of plus disease (0.9 vs 0.9), but poor reliability with identification of the zone of ROP (0.5 vs 0.4).The answers to the repeated 5 sets of images showed that majority had high intra-observer reliability (Table 4) with regard to the stage (55%) and zone (73%) of ROP, while most (73%) had acceptable intra-observer reliability with diagnosing plus disease. None of the respondents had poor intra-observer reliability.

Table 2. Over-all inter-observer reliability.

	ICC	Reliability
Stage	1.0	High
Zone	0.3	Poor
Plus Disease	0.5	Poor

Table 3. Inter-observer reliability by subspecialty.

	Retina specialists		Pediatric Ophthalmologists	
	ICC	Reliability	ICC	Reliability
Stage	1.0	High	0.9	High
Zone	0.5	Poor	0.4	Poor
Plus Disease	0.9	High	0.9	High

Table 4. Intra-observer reliability.

Observer No.	ICC		
	Stage	Zone	Plus Disease
1	0.9	1	0.8
2	1	1	0.7
3	0.9	1	0.8
4	0.9	0.8	1
5	0.7	1	0.7
6	0.9	1	0.9
7	0.7	0.7	0.7
8	1	1	1
9	0.8	1	1
10	0.7	1	1
11	0.8	0.8	0.8

DISCUSSION

There was high over-all inter-observer reliability with the staging of ROP but poor reliability with the identification of the zone and presence of plus disease. This is consistent with previous inter-expert agreement studies wherein classification of the zone and presence of plus disease showed significant disagreement. Chiang and colleagues described a 33% disagreement (7 of 21 images) among 10 expert screeners in the diagnosis of zone I disease.[14] Meanwhile, Wallace reported inter-observer disagreement on the presence of plus disease in 18 of 181 (10%) images and disagreement on distinguishing plus from pre-plus disease for 18 of 67 (27%) among 3 examiners.[12] Similarly, Slidsborg concluded poor inter-observer agreement on plus disease as 4 screeners agreed on only 40 of the 948 (7.38%) quadrants reviewed.[13]

When compared as groups, the reliability between the groups of retina specialists and pediatric ophthalmologists consistently displayed high reliability for stage and poor reliability for zone of ROP. Both also scored high in diagnosing plus disease within each group. However, when compared across the groups, there was difference in reliability due to differences in interpretation of plus disease.

On the other hand, intra-observer reliability was high for both staging and identification of the zone and acceptable for determining the presence of plus disease. This was consistent with the conclusion of Scott and associates that intra-physician agreement in a study of 2 examiners in ROP screening was high.[15]

Over-all diagnosis of the stage of ROP was consistently reliable for both inter- and intra-observer parameters. However, identification of zone of ROP and plus disease were sources of significant discrepancies. The poor reliability in identifying the zone may be due to the limitations of using fundus photographs. Limited images were available; some sets of photos for cases involving zones 2 and 3 did

not include images of the disc or peripheral nasal area which served as landmarks when doing indirect ophthalmoscopy. These limitations may be overcome by capturing more quality images that show these landmarks.

According to Weiner and associates, factors affecting reliability included the observer, the instrument, or may be situational.[20] There was variability in the diagnosis of ROP based on the different inter-observer reliability indices among the participants. Chiang and Jiang explained that the varying interpretations among the observers even while viewing the same images resulted from subjective differences in judgment among the ROP screeners.[11]

Instrument-related factors affecting reliability pertain to the limitations of using wide field fundus photography. Although binocular indirect ophthalmoscopy remains to be the gold standard for ROP diagnosis,[9] we used fundus images obtained by RetCam as serial examination on the same infant for those with significant safety concerns. RetCam is a commonly used instrument for pediatric retinal imaging with a wider field of view of more than 100º. However, it does not have the 3-dimensional advantage of indirect ophthalmoscopy for better assessment of retinal contours.[21] This difference in perspective could have affected our findings especially since 4 participants had no prior experience with interpreting RetCam images.

Situational factors affecting reliability refer to the individual computer monitors of the participants which they used after downloading the powerpoint test containing the RetCam fundus images. Although the use of photographs ensured that all participants evaluated the exact same images and more or less ensured that the precision of the evaluation was not determined by the cooperation of the infant with the examination, it also has limitations. The luminance or resolution of the participant's computer monitor display was not standardized nor the conditions during which they reviewed the images (i.e. clinic hours vs. home).

Indeed, ROP screening remains to be subjective because it is based on photographic standards and descriptive qualifiers rather than quantifiable measurements.[11-13] The authors recommend including the disc and the peripheral nasal quadrant when doing wide-field fundus photography on infants using RetCam to facilitate identification of the zone of ROP. We also propose that standard images from the ICROP Revisited Study, especially images on pre-plus and plus disease, be placed at bedside during routine ROP screening to facilitate diagnosis. Adequate training in ROP diagnosis among screeners is crucial to ensure the highest level of patient care so that findings from major studies such as ICROP, CRYO-ROP, and ETROP trials can be applied properly and consistently.

ACKNOWLEDGMENT

The authors would like to extend our deepest gratitude to the fundus photographers and attending ophthalmologists who helped us collect the fundus photographs; namely, Jose Melvin Jimenez IV, MD, Ricardo Ventura, MD, Fay Cruz MD, and Dulce Peralta, MD.

REFERENCES

1. American Academy of Pediatrics, American Academy of Ophthalmology, American Association of Pediatric Ophthalmology and Strabismus, and American Association of Certified Orthoptists. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics* 2013;131:189-195.
2. Chaudhari S, Patwardhan V, Vaidya U, et al. Retinopathy of prematurity in a tertiary care center – incidence, risk factors and outcome. *Indian Pediatr* 2009;46:219-224.
3. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk, and implications for control. *Early Human Development* 2008;84:77-82.
4. World Health Organization, March of Dimes, The Partnership for Maternal, Newborn and Child Health, and Save the Children. Born Too Soon: The Global Action Report on Preterm Birth. 2012;1-124.
5. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for ROP: three-month outcome. *Arch Ophthalmol* 1990;108:195-204.
6. Early Treatment for Retinopathy of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity. *Arch Ophthalmol* 2003;121:1684-1696.
7. Azad R. Retinopathy of prematurity: a giant in the developing world. *Indian Pediatr* 2009;46:211-212.
8. Elder JE. Is it time to review the screening guidelines for retinopathy of prematurity? *J Paediatr Child Health* 2008; 44:159-160.
9. International Committee for the Classification of Retinopathy of Prematurity (ICROP). The international classification of retinopathy of prematurity revisited. *Arch Ophthalmol* 2005; 123:991-999.
10. Reynolds JD, Dobson V, Quinn GE, et al; CRYO-ROP and LIGHT-ROP Cooperative Groups. Evidence-based screening criteria for ROP: natural history data from the

CRYO-ROP and LIGHT-ROP studies. *Arch Ophthalmol* 2002;120:1470-6.
11. Chiang MF, Jiang L, Gelman R, et al. Inter-expert agreement of plus disease diagnosis in ROP. *Arch Ophthalmol* 2007; 125:875-880.
12. Wallace DK, Quinn GE, Freedman SF, et al. Agreement among pediatric ophthalmologists in diagnosis of plus and pre-plus disease in ROP. *J AAPOS* 2008;12:352-6.
13. Slidsborg C, Forman JL, Fielder AR, et al. Experts do not agree when to treat ROP based on plus disease. *Br J Ophthalmol* 2012;96:549-553.
14. Chiang MF, Thyparampil PJ, Rabinowitz D. Inter-expert agreement in the identification of macular location in infants at risk for ROP. *Arch Ophthalmol* 2010;128:1153-1159.
15. Scott KE, Kim DY, Wang L, et al. Telemedical diagnosis of ROP: intraphysician agreement between ophthalmoscopic and image-based interpretation. *Ophthalmology* 2008;115: 1222-1228.
16. Rasa S. Retinopathy of prematurity: is it time to change screening limits in Lithuania? Vilnius University Children's Hospital, World ROP Congress, 2006.
17. De Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *Clin Epid* 2006;59:1033-1039.
18. Chinn S. The assessment of methods of measurement. *Statistics in Medicine* 1990;9:351-362.
19. Gilbert C, Fielder A, Gordillo L, et al; International NO-ROP Group. Characteristics of infants with severe retinopathy of prematurity in countries with low, moderate, and high levels of development: implications for screening programs. *Pediatrics* 2005;155:e518-e52.
20. Weiner J. Measurement: reliability and validity measures. John Hopkins Bloomberg School of Public Health. http://ocw.jhsph.edu/courses/hsre/PDFs/HSRE_lect7_weiner.pdf (accessed online May 10, 2013).
21. Chiang MF, Wang L, Busuioc M, et al. Telemedical ROP diagnosis: accuracy, reliability and image quality. *Arch Ophthalmol* 2007;125:1531-1538.