Diagnostic Performance of a Telemedicine System for Ophthalmology: Advantages in Accuracy and Speed Compared to Standard Care

Michael F. Chiang, MD[1A,2A], Lu Wang, MD, MS[2A], David Kim, MD[2A], Karen Scott, MD[3A], Grace Richter, MD[3A], Steven Kane, MD, PhD[2A], John Flynn, MD[2A], Justin Starren, MD, PhD[1B]

Departments of Biomedical Informatics[1], Ophthalmology[2], and Pediatrics[3]
Columbia University College of Physicians and Surgeons[A], New York, NY
Marshfield Clinic[B], Marshfield, WI

Abstract

Telemedicine has potential to improve quality and delivery of medical care, particularly in image-oriented specialties where decisions are based on appearance of morphological features during examination. In the ophthalmology domain, nearly all published telemedicine studies have measured accuracy against a gold standard of ophthalmoscopic examination. The purposes of this study are to examine difficulties in defining an absolute gold standard and to compare diagnostic speed in a representative disease, retinopathy of prematurity. We compare results from ophthalmoscopic and telemedicine examinations by the same physicians. In 180 (86.5%) of 208 eyes, the two examinations produced the same diagnosis. In some discrepancies, there was rationale suggesting that telemedicine may have provided a more accurate diagnosis than ophthalmoscopic examination. The quantity and nature of these disagreements has important implications for evaluation of telemedicine systems in image-based specialties, and for the definition of gold standards in future studies.

Introduction

Traditional medical diagnosis in virtually all specialties occurs after examination by a physician. In image-oriented specialties such as ophthalmology, radiology, cardiology, and dermatology, diagnostic decisions are based largely on review of photographic studies captured by technicians.[1] Therefore, remote diagnosis using store-and-forward telemedicine may be a promising strategy for improving the delivery and accessibility of care in image-oriented fields.[2-3]

Meanwhile, modern health care trends are placing increasing emphasis on improved quality and adherence to evidence-based guidelines.[4] However, numerous studies have shown that there is significant practice variation among physicians.[5-6] Although diagnosis in image-oriented specialties is based on the appearance of morphological features visualized from examination, multiple studies in ophthalmology have shown that there is often disagreement among experts presented with the exact same clinical scenarios.[7-9] Telemedicine might create opportunities to improve the quality of health care in image-based specialties because the capture and interpretation of medical data may be standardized and monitored.

Nearly all published studies involving the accuracy of image-based telemedicine systems have examined their performance compared to a gold standard of examination by an expert physician. However, it is not clear that the accuracy of in-person examinations is inherently superior to that of image review by a remote expert. Understanding the factors contributing to accurate diagnosis, as well as having a clear definition of the correct diagnosis, is essential for evaluating performance of telemedicine systems.

The purpose of this paper is to measure the diagnostic performance of an image-based telemedicine system, and to examine difficulties in defining an absolute gold standard. Retinopathy of prematurity (ROP), an ophthalmic disease affecting low birth-weight infants during the first several months of life, is used as the evaluation domain. Results from ophthalmoscopic exams by one of two experts on a study cohort of infants are compared to results from telemedicine exams by the same ophthalmologist on the same infants. Diagnostic speed by one examiner using the two modalities is also compared. By analyzing findings obtained by the same physician using ophthalmoscopic and telemedicine exams, we control for variations in technique and interpretation among individuals. In this way, we isolate the impact of differences between these modalities that are relevant for diagnosis. Clinical implications of this work have previously been reported in separate manuscripts.[10-11] In this paper, we emphasize the methodological implications for biomedical informatics.

Evaluation Domain

ROP is diagnosed from dilated retinal examination by an experienced ophthalmologist, and there are guidelines for identifying high-risk premature infants who need serial screening examinations.[12] When ROP occurs, approximately 90% of cases improve

spontaneously and require only close follow-up examinations every 1-2 weeks. However, the remaining 10% are at high risk for complications leading to blindness and require surgical treatment.[13,15]

ROP an ideal disease for applications and research in telemedicine because: (1) Diagnosis is based solely on the appearance of disease in the retina. (2) There is a universally-accepted, evidence-based, diagnostic classification standard for ROP.[14] (3) Although it is treatable if detected early, ROP continues to be a leading cause of childhood blindness throughout the world because of inadequacies in screening.[16] (4) Current ROP exam methods are time-intensive and physiologically stressful to infants. (5) Clinical expertise is often limited to larger academic centers, and is therefore unavailable at the point of care.

Methods

Ophthalmoscopic Examination and Image Capture

This study was approved by the Columbia University IRB. Infants in the Columbia neonatal intensive care unit (NICU) were included if they met existing criteria for ROP exam from November 2005 to October 2006, and if their parents provided informed consent for imaging and study participation.

Each infant underwent two examinations, which were sequentially performed under topical anesthesia at the NICU bedside: (1) Dilated ophthalmoscopy with scleral depression, based on standard protocols.[12] This was performed by one of two authors (SAK, MFC), who are both pediatric ophthalmologists, and documented according to the international classification standard.[14] (2) Retinal imaging using a wide-angle camera (RetCam; Clarity Medical Systems, Pleasanton, CA). This was performed by a single trained NICU nurse based on manufacturer guidelines. A protocol was established to capture three standard central and peripheral images of each eye, along with up to two additional images if felt by the nurse to contribute diagnostic information. No infants were excluded because of poor image quality.

Image-based Telemedicine Examination

A store-and-forward telemedicine system was developed by the authors (MFC, LW), consisting of a front-end web interface created in Java (Tomcat 6.0, Apache, Forest Hill, MD; Sun Microsystems, Santa Clara, CA) and a back-end database (SQL 2005; Microsoft, Redmond, WA). This SSL-encrypted system included an upload module for the nurse to select and transmit the best images, and a review module displaying images and demographic data for study physicians (Figure 1). Two authors (SAK, MFC) used this system to perform telemedicine exams on the same eyes that they had previously examined using ophthalmoscopy. This system represented data about ophthalmoscopic exam findings, imaging exam properties, and telemedicine exam findings in separate tables. This permitted analysis of diagnostic accuracy, inter-grader reliability through presentation of the same images to multiple graders, intra-grader reliability through repetition of random images to the same graders, and image quality opinions from graders.[17] To simulate a real-world scenario, the system displayed images from both eyes side-by-side, along with the birth weight, gestational age, and age at time of exam.

Physicians interpreted images using a scale based on well-known criteria from NIH-sponsored trials.[13,15] Eyes were classified as: (1) No ROP, meaning that infants are re-examined in 2 weeks for surveillance; (2) Mild ROP, meaning that infants are re-examined in 1-2 weeks; (3) Moderate ROP, meaning "type-2 prethreshold" disease requiring close monitoring in ≤1 week; and (4) Severe ROP, meaning that surgical treatment is required.[12,13,15] To minimize likelihood of physicians remembering data about specific patients, no identifiers were displayed, images were shown in random order, and telemedicine exams were performed 4-12 months after ophthalmoscopy.

The telemedicine system recorded timestamps reflecting examiner speed. Telemedicine diagnosis time was considered to start when a new patient page was opened, and to end when the examiner submitted all responses for that patient. For one physician (MFC), this was compared to ophthalmoscopic exam speed from an independent set of 150 consecutive infants. Ophthalmoscopic diagnosis time was recorded by an outside observer, and considered to start upon arrival to the infant bedside and to end after completion of a standard paper-based note.

Figure 1. Example of image-based telemedicine diagnosis interface. Multiple images from both eyes are displayed, and basic demographic information is shown.

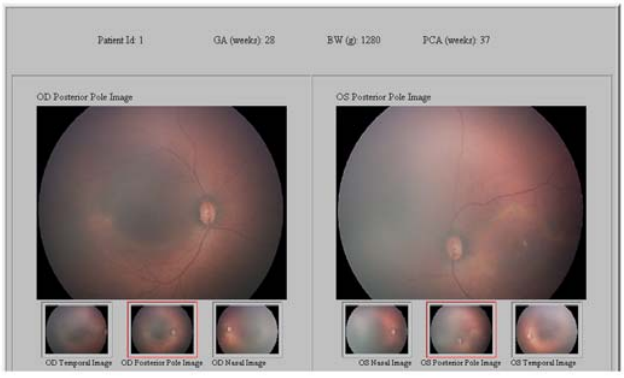


Table 1. *Results from ROP diagnosis on 208 study eyes by two physicians, both of whom performed ophthalmoscopic and masked image-based examinations on the same patients. ROP is classified ordinally as: none, mild, moderate (type-2 prethreshold), or severe (treatment-requiring) disease.*

Ophthalmoscopy Exam	Image-based Exam: None	Mild	Moderate	Severe
None	103 (49.5%)	9 (4.3%)	0 (0%)	0 (0%)
Mild	6 (2.9%)	62 (29.8%)	5 (2.4%)	2 (1.0%)
Moderate	0 (0%)	4 (1.9%)	11 (5.3%)	0 (0%)
Severe	0 (0%)	0 (0%)	2 (1.0%)	4 (1.9%)

Data Analysis

Data were analyzed using spreadsheet software (Excel 2003; Microsoft, Redmond, WA). Agreement between ophthalmoscopic and image-based exams was calculated for each physician, using the ordinal classification scale described above. Cases in which these two examinations by the same physician resulted in different diagnoses were identified and reviewed by an independent examiner (DYK). After consensus of three authors (DYK, KES, MFC), the reason for disagreements was identified as one of the following: no ROP identified by ophthalmoscopic exam, no ROP identified by image-based exam, disagreement in classification of ROP severity ("stage"), disagreement in classification of ROP location ("zone"), or disagreement in classification of blood vessel appearance ("dilation" and "tortuosity").

Results

Characteristics of Study Population

A total of 68 infants were enrolled in the study. Both eyes of each infant underwent ophthalmoscopic and telemedicine exams by one of two physicians. Each infant received up to two sets of examinations during different weeks, for a total of 208 study eyes. Among these eyes, 158 (76.0%) received both examinations from Physician #1, and 50 (24.0%) received both examinations from Physician #2. According to ophthalmoscopic exams, 112 (53.8%) eyes had no ROP, 75 (36.1%) had mild ROP, 15 (7.2%) had moderate ROP, and 6 (2.9%) had severe ROP. According to image-based exams, 109 (52.4%) eyes had no ROP, 75 (36.1%) had mild ROP, 18 (8.6%) had moderate, and 6 (2.9%) had severe ROP.

Agreement between Examinations

Table 1 shows pooled results from ophthalmoscopic and telemedicine exams by the same physicians. In 180 (86.5%) eyes, the same diagnostic classification resulted from both exams. Image-based examination resulted in a more severe diagnosis in 16 (7.7%) eyes, while ophthalmoscopy gave a more severe diagnosis in 12 (5.8%) eyes. Table 2 shows underlying reasons for the 28 (13.5%) disagreements between exams. All discrepancies involved a single level (e.g. "mild" vs. "moderate"), except two cases in which "mild ROP" was diagnosed by ophthalmoscopy but image-based exam reported "severe ROP."

To compare intra-physician agreement between telemedicine and ophthalmoscopy to inter-grader agreement, Physician #1 performed secondary review of every telemedicine image reviewed by Physician #2. This resulted in 39/50 (78.0%) inter-physician agreement among telemedicine eye examinations.

Speed of Examinations

The mean (range) ± standard deviation (SD) time of ophthalmoscopic diagnosis by Physician #1 was 4.17 (1-11) ± 1.34 minutes, whereas the mean (range) ± SD time of telemedicine diagnosis by the same physician was 1.75 (1-7) ± 0.80 minutes. This difference was highly statistically significant ($p<0.0001$). Data comparing diagnostic speed by additional physicians using these two modalities are reported in a more detailed clinical manuscript.[11]

Table 2. *Reasons for disagreement between ophthalmoscopic and image-based telemedicine diagnosis.*

Reason for disagreement	Number (%) eyes
No ROP identified by ophthalmoscopic exam	9 (32.1%)
No ROP identified by image-based exam	6 (21.4%)
Disagreement in classification of severity of peripheral ROP	1 (3.6%)
Disagreement in classification of location of ROP	8 (28.6%)
Disagreement in classification of blood vessel appearance	4 (14.3%)

Discussion

Telemedicine has been promoted as a strategy for improving accessibility to health care.[2-3] The key findings from this study are that: (1) diagnostic agreement in ROP classification between ophthalmoscopic and image-based exam of the same eyes by the same physicians is imperfect, and that (2) telemedicine diagnosis was significantly faster than standard ophthalmoscopic diagnosis. These findings suggest that telemedicine may also have important benefits with regard to improving the accuracy and speed of health care delivery. This study design controls for variation among multiple examiners, and thereby isolates differences between these two modalities that are relevant for diagnosis. This has implications for the evaluation of image-based telemedicine systems, and for the definition of "gold standards" in these studies. It is instructive to review three specific categories of diagnostic disagreements between these two modalities (Table 2).

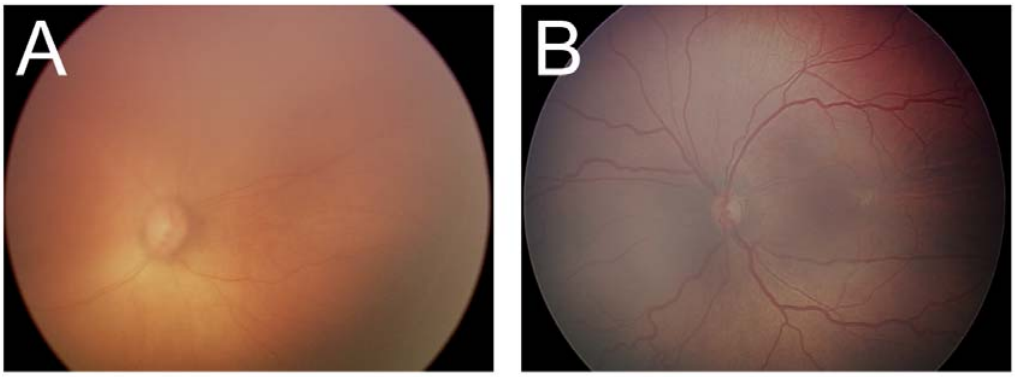
First, 9 (32.1%) of the 28 disagreements in this study occurred because disease was identified by image-based exam but not by ophthalmoscopy (Figure 2A). Given that disease was photographically documented in all cases, these scenarios likely reflect "false-negative" errors by ophthalmoscopy, rather than "false-positives" by telemedicine. In the opposite situation, 6 (21.4%) eyes were found to have disease that was identified by ophthalmoscopy but not by telemedicine. We cannot determine whether these latter cases reflected "false-negative" errors by telemedicine or "false-positives" by ophthalmoscopy.

Second, 4 (14.3%) eyes with disagreements had differing classifications regarding blood vessel appearance (Figure 2B). Vascular "dilation" and "tortuosity" are characteristic of severe ROP, and a standard published photograph selected by expert consensus defines the minimum amount of vascular abnormality that warrants surgical treatment.[13] Although this standard photographic definition is familiar to every ophthalmologist, we have previously shown that there is significant inter-expert variation in diagnosis, presumably because of differing qualitative interpretations of the precise meaning of "dilated" and "tortuous."[7] This may be an advantage of telemedicine because exam findings could be directly compared to photographic standards, thereby decreasing the impact of subjective differences in examiner judgment.[7-9]

Third, there was disagreement in classification of ROP location in 8 (28.6%) eyes. An international disease classification system defines three anatomic "zones" of the retina, and ROP disease that is located in the most central zone has been shown to represent the most severe disease with the worst prognosis.[13-14] Although the borders between these zones are anatomically defined, the landmarks separating these zones are often very difficult to distinguish during ophthalmoscopic exam. It is conceivable that image-based exam may produce more accurate and reproducible identification of the zone in which ROP disease is located, because these landmarks could be precisely measured on retinal photographs.

The accuracy of telemedicine systems has been studied in many image-based specialties. In ROP, as in most other domains, all published outcome studies to our knowledge have compared the results of image-based diagnosis to a gold standard of ophthalmoscopic exam.[17] Using that approach, the sensitivity/specificity for diagnosis of type-2 prethreshold or worse ROP would be 0.895/0.971 for Physician #1, and 0.810/0.985 for Physician #2. However, this study identified 9 (32.1%) eyes with disagreement between ophthalmoscopic and telemedicine exams where there was photographic

Figure 2. Examples of disagreements. In (A), ROP was documented by image-based exam (arrows point at vertical white line denoting peripheral disease) but not by ophthalmoscopy. In (B), there was disagreement regarding whether dilation and tortuosity of vessels was sufficient to meet criteria for severe ROP.



documentation that image-based diagnosis was correct (Table 2). There were also 12 (42.9%) highly clinically-significant disagreements regarding vascular appearance or disease location, in which there is rationale that telemedicine may be inherently more accurate (Table 2).[13,15] We feel that these issues may be generalizable across many diseases that are diagnosed and classified from appearance of images, and that our findings raise significant concerns about the design of evaluation studies involving these diseases. In the future, a more rigid design for these telemedicine studies could involve manual review of images for photographic evidence of disease, in order to establish a more rigorous gold standard.

There are several other factors and limitations regarding this study: (1) No standardization of reading conditions was performed, such as resolution and color correction for monitor displays. (2) This study was designed to determine whether the same diagnosis was reached using ophthalmoscopic and image-based examinations, and not to analyze the accuracy of either approach. (3) The agreement between ophthalmoscopic and image-based examinations by the *same* physician in this study is considerably higher than that reported by previous research about agreement among *multiple* graders reviewing the same images.[17] In the current study, the intra-physician agreement between these two modalities was also higher than the inter-physician agreement for telemedicine exams. These findings suggest that inter-observer variation may exceed the variation due to technology. However, follow-up studies are required to understand how the inter-physician variation in telemedicine exams compares to that of standard clinical examination. (4) Physicians were aware that they were being monitored while recording speed to telemedicine and clinical exams. This may have created a Hawthorne effect bias, or additional bias to the extent that physicians may have had an interest in telemedicine.

Telemedicine offers potential benefits for health care delivery, and may be especially well-suited for image-based specialties. Research involving accuracy of these systems has traditionally been performed in comparison to a gold standard of in-person examination. We show that telemedicine may have additional benefits regarding accuracy and speed of diagnosis, and that alternative approaches may be needed to define a true gold standard for rigorous technology evaluation in the future.

References

1. ETDRS Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs. ETDRS report no. 10. Ophthalmology 1991; 98: 786-806.

2. Grigsby J, Sanders JH. Telemedicine: where it is and where it's going. Ann Int Med 1998; 129: 123-7.

3. Bashshur RL, Reardon TF, Shannon GW. Telemedicine: a new health care delivery system. Annu Rev Public Health 2000; 21: 613-37.

4. Institute of Medicine. To Err is Human: Building a Safer Health Care System. Washington: National Academies Press, 2000.

5. Peabody JW, Luck J, Glassman P. Comparison of vignettes, standardized patients, and chart abstraction: a prospective study of 3 methods for measuring quality. JAMA 2000; 283: 1715-22.

6. Veloski J, Tai S, Evans AS, Nash DB. Clinical vignette-based surveys: a tool for assessing physician practice variation. Am J Med Qual 2005; 20: 151-7.

7. Chiang MF, Jiang L, Gelman R, et al. Inter-expert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol 2007; 125:875-80.

8. Moss SE, Klein R, Kessler SD, Richie KA. Comparison between ophthalmoscopy and fundus photography in determining severity of diabetic retinopathy. Ophthalmology 1985; 92: 62-7.

9. Kinyoun JL, Martin DC, Fujimoto WY, Leonetti DL. Ophthalmoscopy versus fundus photographs for detecting and grading diabetic retinopathy. Invest Ophthalmol Vis Sci 1992; 33: 1888-93.

10. Scott KE, Kim D, Wang L, et al. Telemedical ROP diagnosis: agreement between ophthalmoscopic and image-based examinations. Ophthalmology 2008; 115:1222-8.e3.

11. Richter GM, Sun G, Lee TC, et al. Speed of telemedicine versus ophthalmoscopy for ROP diagnosis. Am J Ophthalmol 2009; 148:136-42.e2.

12. Section on Ophthalmology, et al. Screening examination of premature infants for retinopathy of prematurity. Pediatrics 2006; 117: 572-6.

13. Cryotherapy for ROP Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity: preliminary results. Arch Ophthalmol 1988; 106: 471-9.

14. Committee for the classification of retinopathy of prematurity. The international classification of ROP revisited. Arch Ophthalmol 2005; 123: 991-9.

15. Early Treatment for ROP Cooperative Group. Revised indications for the treatment of ROP. Arch Ophthalmol 2003; 121: 1684-94.

16. Steinkuller PG, Du L, Gilbert C, et al. Childhood blindness. J AAPOS 1999; 3: 26–32.

17. Chiang MF, Wang L, Busuioc M, et al. Telemedical retinopathy of prematurity diagnosis: accuracy, reliability, and image quality. Arch Ophthalmol 2007; 125:1531-8.

Acknowledgements

Supported by grant EY13972 from the National Institutes of Health, Bethesda, MD (MFC), and by a Career Development Award from Research to Prevent Blindness, New York, NY (MFC).