University of York Department of Health Sciences

M.Sc. in Evidence Based Practice

Measurement in Health and Disease

Assessment, June 2006

Time allowed: 60 minutes

ANSWER TWO QUESTIONS

Each part of a question carries equal marks.

Question 1.

The following is the abstract of a recent paper:

Title: Detection of diabetic retinopathy: a comparison between red-free digital images and colour transparencies.

Background: The aim of this study was to compare how diabetic retinopathy was detected from red-free digital images and colour transparencies.

Methods: Two ophthalmologists graded two-field, nonstereoscopic, 60^{o}; red-free digital images and colour transparencies utilizing an ETDRS-based grading scale, from 107 mainly type 2 diabetic patients. The discordantly scored eyes were graded by the graders together to obtain a consensus level of retinopathy for each method. The eyes with discordant consensus grading results were further graded using all available photographic material to reach a final consensus level of diabetic retinopathy. Intermethod variations were presented as percentages and using kappa (*k*) and weighted kappa (*wk*) statistics. The errors of the two consensus gradings with respect to the final consensus grading were compared using McNemar's test.

Results: For the colour transparencies there was an agreement between the individual and the consensus grading results in 93% ($k = 0.90$, $wk = 0.97$) and 86% (k = 0.79, wk 0.88) for grader 1 and grader 2. Corresponding figures for red-free digital images were 88% ($k = 0.83$, $wk = 0.96$) and 84% ($k = 0.78$, wk 0.91). Agreement between methods was obtained in 76/107 eyes (71%; $k = 0.58$ and $wk = 0.79$). In the 31 discordantly graded eyes the level of retinopathy was underestimated in 20/31 (64%) vs 7/31 eyes (23%) and overestimated in 1/31 (3%) vs 3/31 eyes (10%) from colour transparencies and red-free digital images, respectively. The error tendencies were significantly lower when using red-free digital images ($p< 0.008$).

Conclusions: Red-free digital images are comparable with two-field colour transparencies in the identification of mild to moderate nonproliferative diabetic retinopathy.

(Von Wendt G, Summanen P, Hallnas K, Algvere P, Heikkila K, Seregard S. (2005) Detection of diabetic retinopathy: a comparison between red-free digital images and colour transparencies. *Graefes Archive for Clinical and Experimental Ophthalmology* **243**, 427-432.)

(a) What is a kappa statistic? How could we interpret $k = 0.90$ and $k = 0.58$?

(b) What is a weighted kappa statistic and how does it differ from kappa? Why are the weighted kappa statistics all larger than the corresponding kappa statistics and what does this tell us?

(c) Percentage agreement is given for each of the measures used. Why might this be a misleading statistics and why is it always greater than the corresponding kappa?

Question 2.

The following is the abstract of a recent paper:

Title: Test-retest reliability of isokinetic dynamometry for the assessment of spasticity of the knee flexors and knee extensors in children with cerebral palsy.

Abstract: Objective: To assess test-retest reliability of the peak resistance torque and slope of work methods of spasticity measurement of the knee flexors and extensors in children with cerebral palsy (CP).

Design: Test-retest reliability study.

Setting: Pediatric orthopedic hospital.

Participants: Fifteen children with CP.

Intervention: Knee extensor and flexor spasticity was assessed with an isokinetic dynamometer using passive movements at 15 degrees, 90 degrees, and 180 degrees/s taken 1 hour apart.

Main Outcome Measures: Peak resistive torque and work were calculated. The relative and absolute test-retest reliability was calculated by using intraclass correlation coefficients (ICCs) and Bland-Altman plots, respectively.

Results: Relative reliability was good (ICC>.75) for slope-of-work and peak resistance torque measurements at a velocity of 180 degrees/s. whereas reliability of peak torque measurements was decreased (ICC<.51) at slower velocities for both muscle groups. The 95% limits of agreement of Bland-Altman plots contained most data points for both methods, but the width of the limits of agreement were wide.

Conclusions: The measurement of spasticity of the knee extensors and flexors in children with CP using peak-resistance torque at 180 degrees/s and the slope of work method has acceptable relative test-retest reliability. However, the absolute reliability of spasticity data should be considered cautiously.

(Pierce SR, Lauer RT, Shewokis PA, Rubertone JA, Orlin MN. Test-retest reliability of isokinetic dynamometry for the assessment of spasticity of the knee flexors and knee extensors in children with cerebral palsy. *Archives of Physical Medicine and Rehabilitation* 2006; **87**: 697-702.)

(a) What is an intraclass correlation coefficient? Why is this described as measuring 'relative' reliability?

(b) What are limits of agreement and how could they be used to investigate test-retest reliability? Why is this described as measuring absolute reliability?

(c) The authors report that "the 95% limits of agreement of Bland-Altman plots contained most data points for both methods". Why doesn't this tell us anything useful?

Question 3.

The following is the abstract of a recent paper:

Title: Child mania rating scale: Development, reliability, and validity

Objective: To develop a reliable and valid parent-report screening instrument for mania, based on DSM-IV symptoms.

Method: A 21-item Child Mania Rating Scale-Parent version (CMRS-P) was completed by parents of 150 children (42.3% female) ages 10.3 +/- 2.9 years (healthy controls = 50; bipolar disorder = 50; attention-deficit/hyperactivity disorder [ADHD] = 50). The Washington University Schedule for Affective Disorders and Schizophrenia was used to determine DSM-IV diagnosis. The Young Mania Rating Scale, Schedule for Affective Disorders and Schizophrenia Mania Rating Scale, Child Behavior Checklist, and Child Depression Inventory were completed to estimate the construct validity of the measure.

Results: Exploratory and confirmatory factor analysis of the CMRS-P indicated that the scale was unidimensional. The internal consistency and retest reliability were both 0.96. Convergence of the CMRS-P with the Washington University Schedule for Affective Disorders and Schizophrenia mania module, the Schedule for Affective Disorders and Schizophrenia Mania Rating Scale, and the Young Mania Rating Scale was excellent (0.78-0.83). The scale did not correlate as strongly with the Conners parent-rated ADHD scale, the Child Behavior Checklist -Attention Problems and Aggressive Behavior subscales, or the child self-report Child Depression Inventory (0.29-0.51). Criterion validity was demonstrated in analysis of receiver operating characteristics curves, which showed excellent sensitivity and specificity in differentiating children with mania from either healthy controls or children with ADHD (areas under the curve of 0.91 to 0.96).

Conclusion: The CMRS-P is a promising parent-report scale that can be used in screening for pediatric mania.

(Pavuluri MN, Henry DB, Devineni B, Carbray JA, Birmaher B. Child mania rating scale: Development, reliability, and validity. *Journal of the American Academy of Child and Adolescent Psychiatry* 2006; **45**: 550-560.)

(a) What does the authors mean by 'Exploratory . . . factor analysis of the CMRS-P indicated that the scale was unidimensional.'?

(b) The internal consistency (alpha) and test-retest reliability were both 0.96. What do these terms mean and how would we interpret 0.96?

(c) What is 'construct validity' and how could the analysis described demonstrate it?