Evaluation of Interobserver Agreement In Gonioscopy

Dr. Jenny Thomas Jacob DO, Dr. Shima M. MS, Dr. Thomas George T. M.S.

Abstract

Aim:- To define the learning curve in gonioscopy.

Materials and Methods :- A single blind study using one resident with one year experience in Ophthalmology and another resident with two years experience in ophthalmology and three weeks in Glaucoma speciality clinic was done. Each student independently did gonioscopy on the fellow eye of fifty patients admitted for cataract surgery. Tests of agreement(reliability) was performed and kappa was derived.

Results: - Observed agreement for Shaffer’s grading: 0.62 (k=0.46), occludability : 0.88 (k=0.45), posteriormost structure : 0.62 (k=0.46).

Keywords:- Gonioscopy, interobserver agreement, kappa, glaucoma, Shaffer's grading, occludability, posteriormost structure.

Introduction

Gonioscopy is a clinical technique used to examine structures in the anterior chamber angle, which is essential in the management of different types of glaucoma. But like many other physical findings, radiographic interpretations and other diagnostic tests, gonioscopy too often rely on some degree of subjective interpretation by observers. Thus, like any other diagnostic test, gonioscopy too has its own learning curve effect.

Aims and Objectives

To define the learning curve involved in gonioscopy and to assess the level of agreement between two postgraduate students at different levels of training in Ophthalmology, in the interpretation of gonioscopy findings.

Materials and Methods

The study was conducted in the Regional Institute of Ophthalmology, Trivandrum over a period of 3 months in 2007. A single blind study was conducted using two observers – observer 1, a postgraduate student with one year experience in Ophthalmology and observer 2, a postgraduate student with two years experience in Ophthalmology and three weeks in Glaucoma speciality clinic. Each student independently did gonioscopy on the fellow eye of 50 patients admitted for cataract surgery and recorded findings with respect to Schaffer’s grading, occludability and posteriormost structure seen (without the knowledge of the other’s interpretation).
The test of agreement (reliability) was performed and kappa was derived which is the statistical method of analysis adopted in our study.

Method of Analysis

Accuracy versus precision

When assessing the ability of a test (radiograph, physical finding etc.) to be helpful to clinicians, it is important that its interpretation is not a product of guesswork. This concept is often referred to as ‘precision’ (though some incorrectly use the term ‘accuracy’). For example, if we actually hit the bull’s-eye of a target (representing agreement with the gold standard), we are accurate. If all our shots land together, we have good precision (good reliability). If all our shots land together and we hit the bull’s-eye, we are accurate as well as precise. It is however possible to hit the bull’s-eye purely by chance.

Precision, as it pertains to agreement between observers (interobserver agreement), is often reported as a kappa statistic, which is intended to give a quantitative measurement of the magnitude of agreement between two or more observers. For example, comparing the presence of wheezes on lung examination to the presence of an infiltrate on a chest radiograph assesses the validity of the exam findings to diagnose pneumonia. Assessing whether the examiners agree on the presence or absence of wheezes (regardless of validity) assesses precision (reliability).

Kappa Statistics

<table>
<thead>
<tr>
<th>Kappa Value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 0</td>
<td>Less than chance agreement</td>
</tr>
<tr>
<td>0.01 – 0.20</td>
<td>Slight agreement</td>
</tr>
<tr>
<td>0.21 – 0.40</td>
<td>Fair agreement</td>
</tr>
<tr>
<td>0.41 – 0.60</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.61 – 0.80</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>0.81 – 0.99</td>
<td>Almost perfect agreement</td>
</tr>
<tr>
<td>+ 1</td>
<td>Perfect agreement</td>
</tr>
</tbody>
</table>

Interobserver variation can be measured in any situation in which two or more observers are evaluating the same thing. For example, let us imagine a study in which two medical students are evaluating the usefulness of a series of hundred morning lectures. They agree that the lectures are useful 15 percent of the time, while it is not useful 70 percent of time (in other words the remaining 15 percent of the lectures the 2 students disagreed with each other i.e., 1 felt it was good and the other called it bad).

The calculation is based on the difference between how much agreement is actually present (PO = observed agreement) compared to how much agreement would be expected to be present by chance alone (PE = expected agreement). In the example cited above, the observed agreement is the percent of all lectures for which the two residents’ evaluations agree.

Also, kappa is a measure of the difference between the two i.e., agreement beyond chance (k = PO – PE/1 – PE). The values are standardized to lie on a scale from -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, i.e., potential systematic disagreement between the observers.

Sometimes, we are more interested in the agreement across major categories in which there is a meaningful difference. For example, let’s suppose we had five categories of ‘helpfulness of noon lectures’: ‘very helpful’, ‘somewhat helpful’, ‘neutral’, ‘somewhat a waste’, ‘complete waste’. Here, we may not care if one resident categorizes as ‘somewhat helpful’ and another categorizes as ‘very helpful’, but we might care if one resident categorizes as ‘very helpful’ and another categorizes as ‘complete waste’.

Using a clinical example, we may not care if one radiologist categorizes a mammogram finding as normal and another categorizes it as benign, but we do care if one categorizes it as normal and another categorizes it as cancer.

Here, is where the weighted kappa assumes its significance, which is an appropriate ‘chance adjusted measure of agreement’ between two observers, when there are more than two ordered categories of classification. This statistic ranges from -1 (agreement less than chance) to +1 (perfect agreement). In our previous example, a disagreement of normal versus benign would still be credited with partial agreement, but a disagreement of normal versus cancer would be counted as no agreement.
Limitation of Kappa

It may not be reliable for rare findings. Thus, very low values of kappa in such cases, may not necessarily reflect low rates of overall agreement.

Our Results

Observed agreement (PO)

- Schaffer's grading $\rightarrow 0.62$ (k = 0.46)
- Occludability $\rightarrow 0.88$ (k = 0.45)
- Posterior most structure $\rightarrow 0.62$ (k = 0.46)

Discussion

We found moderate agreement with Schaffer's grading (k = 0.46), occludability (k = 0.45) and posterior most structure (k = 0.46). Thus, we have demonstrated that a junior resident can achieve moderate levels of agreement with a senior resident, in the gonioscopic evaluation even without specific training. Further consensus training can increase the level of agreement to substantial to almost perfect.

The two observers showed fair agreement (k = 0.393) for Shaffer's grading, less than chance agreement (k = 0.071) for occludability and fair agreement (k = 0.393) for posteriormost structure seen, for the first twenty patients seen among the fifty selected for the study. These values improved to moderate agreement (k = 0.516) for Shaffer's grading, substantial agreement (k = 0.615) for occludability and moderate agreement (k = 0.516) for posteriormost structure, seen for the last twenty patients. This demonstrates a learning curve for gonioscopy and suggests that the stage of training might have influenced the degree of improvement. It can also be concluded that occludability showed a steeper learning curve when compared to the other two parameters.

Development of standardized criteria and reporting forms, pilot testing and training of raters through the review of disagreements are some of the methods of maximizing agreement in a wide variety of clinical ratings.

Conclusions

- As with any other diagnostic test, gonioscopy too has its own leaning curve, with a steeper curve for occludability.
- As the junior resident has demonstrated moderate agreement with the senior resident even without specific training in gonioscopy, it can be concluded that with consensus training, the junior resident can be given sole responsibility for assessment of gonioscopy and thus, the patient needs can be addressed in a better way.
- Multiple clinicians involved in clinical trials, should seriously consider pilot training and assessment of the level of agreement in making clinical and diagnostic test ratings, to enhance the power and accuracy of their studies.

References

2. A J Viera, MD; J M Garrett, MD, Ph.D. - Understanding Interobserver agreement ; The Kappa statistic (Fam Med 2005;37 (5) 360 – 3.