THE YALE JOURNAL OF BIOLOGY AND MEDICINE 55 (1982), 81-88

Assessment of Observer Variability in the Classification of Human Cataracts

DOMENIC V. CICCHETTI, Ph.D.,[a] YOG SHARMA, M.D.,[b] AND EDWARD COTLIER, M.D.[b]

[a] *VA Medical Center, West Haven, Connecticut;*
[b] *Departments of Visual Science and Ophthalmology,*
Yale University, School of Medicine, New Haven, Connecticut

Received February 16, 1982

An *in vitro* cataract classification system was developed in our laboratories and used to demonstrate a relationship between sustained aspirin intake and the apparent deceleration or retardation of human cataract formation. The purpose of this investigation was to assess the reliability of this cataract classification schema. Sets of extracted human cataractous lenses, which had been photographed *in vitro*, were randomly assigned to five observers. The task was to classify the lenses on the basis of nuclear and cortical involvement, as reflected in color and area changes along five groupings. Assessments were made on the basis of both intraobserver and interobserver agreement levels, corrected for chance (weighted kappa values). All five examiners evidenced levels of *intra*observer agreement which ranged between "Good" and "Excellent" (.73-.92). The *inter*observer chance-corrected agreement levels ranged between "Fair" and "Excellent" (.46-.83). Each of the five observers was ranked on the basis of his agreement levels with the remaining four observers. The results followed a predictable pattern such that the more experienced the observer in classifying cataracts, the more consistent his rankings vis-à-vis the remaining four evaluators. These results are discussed in the general context of observer variability studies in the field of medicine.

INTRODUCTION

A number of previous investigators have made reference to the rather extensive levels of observer variability, in the classification of clinical signs and symptoms, across a wide range of medical phenomena [1-4].

Within the past several years, the phenomenon of observer variability has appeared in the writings of a number of research workers in the field of ophthalmic pathology [5-9]. Although observer variation was mentioned in varying degrees in each of these studies, only one of them (Kahn [6]) provided any formal documentation as to the extent of examiner differences in the assessment of various aspects of ophthalmic pathology. Kahn's research is an important contribution to the field of ophthalmic pathology because it demonstrates that although observer variability appears as prevalent here as in many other specific areas of medicine, it is possible to *reduce* the extent of interobserver variation, thereby increasing the precision or reliability of the measurement of crucial aspects of ophthalmic pathology.

Despite the obvious heuristic value of Kahn's research for the field of ophthalmic pathology, the statistics that were employed (chi squared and average interexaminer variance ratios) rendered impossible the answers to a question we regard as

Address reprint requests to: Domenic V. Cicchetti, Ph.D., VA Medical Center, West Haven, CT 06516

somewhat fundamental to the assessment of observer variability in visual research; namely, the extent to which different observers, examining the same ocular variables, agree both within and between each other on the basis of levels of both statistical and clinical significance. The latter is an important distinction since levels of statistical significance which are of doubtful clinical or practical value can be easily obtained with large enough samples of patients. Our research attempted, by the use of appropriate statistics and guidelines for judging levels of clinical or practical significance, to obviate these problems. We had as our primary objective the assessment of observer variation in the classification of stages of human cataract development, on the basis of changes in nuclear color. The classification schema we employed has been used in our laboratories for many years to investigate the relationship between aspirin intake and cataract formation in elderly human adults, e.g., Cotlier [10]. Somewhat similar schema have been described by Marcantonio and colleagues [8] based upon the previous work of Pirie [11]. This former group of investigators studied the relationship between nuclear color changes and levels of sodium content in human cataractous lenses. However, no attempt was made to either quantitate this relationship or to assess the specific role of observer variation in the classification of stages of cataract development.

On the basis of previous work by Kahn [6] and Koran [3–4] we were interested in the following assessments:

1. *Intra*observer agreement levels (or the extent to which observers would agree with their own previous classification of stages of cataract development); and *inter*observer agreement levels (agreement with other observers) in the classification of stages of human cataract development and
2. The role of experience in the classification of these same human cataract stages

MATERIALS AND METHODS

Observers (or Examiners)

Five observers participated in this study. They varied in levels of experience as follows: (*1*) the most senior observer, a 42-year-old board-certified ophthalmologist, had more than five years of experience in both the specific classification of human cataracts and in general classification methods in the field of vision; (*2*) the next most experienced examiner was a 40-year-old senior technician with two years of the specific experience just described; (*3*) third in experience level was a 31-year-old postdoctoral fellow, who was board-certified, with general visual classification experience but none specific to differentiating among stages of human cataract development; (*4*) next in experience level was a 28-year-old postdoctoral fellow with one year of experience in the field of ophthalmology but neither specific nor general experience in ophthalmic classification procedures; and (*5*) the least experienced examiner was a 22-year-old junior technician with no general experience at all and only occasional or sporadic involvement in classifying human cataracts.

Extracted Lenses

Twenty-two extracted human cataractous lenses, which had been photographed *in vitro*, were obtained from elderly males, ranging in age from 55 to 91.

Cataract Scale

Five copies of each of the 22 photographs were made for a total of 110 slides. These photographs were randomly assigned to the five observers who were asked to classify a given lens using the aforementioned method of Cotlier [10], a system which is based upon changes in nuclear color and distortion of grid plot lines. Each observer made his assessments independently, using a standard set of five photographs of increasing cataract involvement, as photographed on a grid plot which could be described as:

Category A: Distortion of less than 4 lines; photograph light yellow, with percentage area of opacity between 5 and 25 percent.

Category B: Distortion of 4–6 lines, photograph yellow, with percentage opacity between 26 and 50 percent.

Category C: Distortion of > 6 lines, photograph dark yellow, with percentage opacity between 51 and 75 percent.

Category D: Distortion of > 6 lines, photograph yellow brown, with percentage opacity between 76 and 100 percent.

Category E: Distortion of all grid lines, photograph brown, with area of opacity between 95 and 100 percent.

Methods of Analysis

The determination of levels of intraobserver and interobserver agreement in classifying the five stages of cataract development was based upon the following rationale: Since the development of human cataracts appears to progress in stages based, in part, on changes in nuclear color and density, then the "distance" between observer ratings of the same cataract photograph should indeed be graded progressively as well. Thus, a B–C discrepancy (one cataract stage apart) is viewed as less serious than a C–E discrepancy (two stages apart). These two levels of discrepancies are in turn seen as less serious than A–E pairings (four cataract stages apart, the maximum "distance" possible). We adopted a scoring strategy recommended by us and previous investigators [12–13] which can be described as the following: complete observer agreement pairings (A–A, B–B, . . . E–E) are scored as 1; complete disagreement pairings (A–E or E–A) are scored as zero (0), since they are as far apart as a five-category rank-ordered scale will allow; agreement within a single cataract stage is the least serious discrepancy possible (e.g., A–B, C–D pairings) and such pairings receive an examiner agreement score of .75; stages two categories apart (e.g., C–E, D–B) receive a score of .50; and, finally, cataract stage discrepancies three categories apart (e.g., A–D, E–B intra- or interexaminer pairings) receive a score of .25.

Intraobserver Variability

The question of interest here was to what extent would each of the five observers agree with himself on the assignment of a given photograph to a particular stage of cataract development? It will be recalled that each observer received, in random order, each of the same 22 slides repeated five times. Thus, for each reading (R) of each of the 22 photographs, 10 paired comparisons could be made (according to the formula R(R − 1)/2): (*1*) R1 vs. R2; (*2*) R1 vs. R3; (*3*) R1 vs. R4; (*4*) R1 vs. R5; (*5*) R2 vs. R3; (*6*) R2 vs. R4; (*7*) R2 vs. R5; (*8*) R3 vs. R4; (*9*) R3 vs. R5; and (*10*) R4 vs. R5.

We sought answers to the following questions: (*1*) What is the level (proportion or percentage) of intraobserver agreement (PO)? (*2*) What is the proportion of intraobserver agreement to be expected on the basis of chance alone (PC)? (*3*) What is the *difference* between observed and expected intraobserver agreement (or PO −

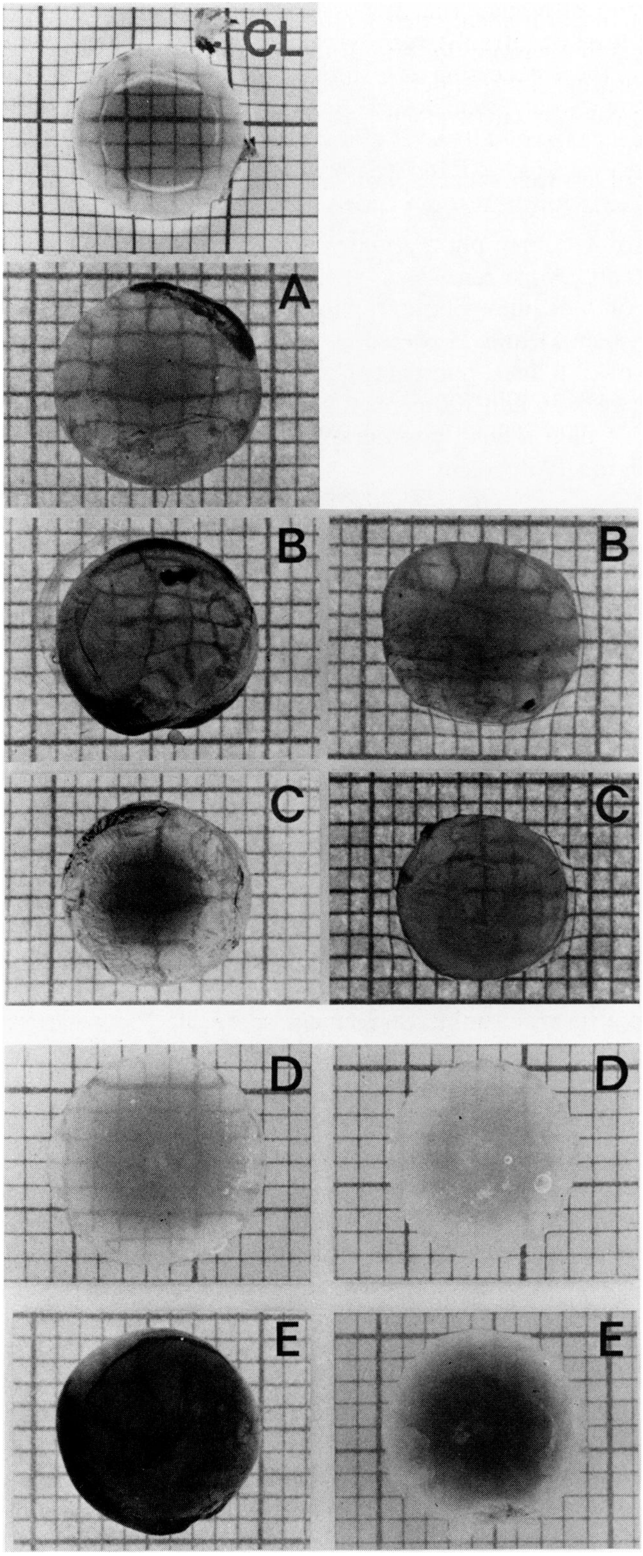

FIG. 1. Photographic classification of human cataracts. *Legend:* **CL** = clear lens, no cataract present; **A, B, C, D,** and **E** refer to examples of progressive stages of cataract development.

PC) *relative* to the maximum difference that is possible (or 1 − PC)? (*4*) Is the chance-corrected or kappa level of intraobserver agreement (Cohen and colleagues [14–16]) defined in question three occurring at a statistically significant level (e.g., $p \leq .05$)? and (*5*) Is the "chance-corrected" level of intraobserver agreement of any practical or clinical significance *beyond* its level of statistical significance? Utilizing guidelines recently published by Cicchetti and Sparrow [17] and Fleiss [18], which are a simplified version of previously recommended guidelines (Landis and Koch [19]) we defined clinical, practical, or substantive levels of chance-corrected observer agreement as follows: <.40 = *Poor*; .40–.59 = *Fair*; .60–.74 = *Good*; and ≥.75 = *Excellent.*

RESULTS

The data in Table 1 indicate the following:

1. Levels of intraobserver agreement (uncorrected for chance) varied between 85.2 and 98.9 percent.
2. Chance-corrected levels of intraobserver agreement (weighted kappa or κ_w values) were all highly statistically significant at $p < .0005$.
3. From a practical, clinical, or substantive point of view, individual κ_w values ranged between .60 ("*Good*") and .96 ("*Excellent*"). When κ_w values were averaged across the 10 possible orders of presentation for each observer, they varied between .73 ("*Good*") and .92 ("*Excellent*").

Interobserver Variability

The data for interobserver agreement levels were also analyzed by employing the weighted kappa statistic. Since the intraobserver agreement levels among the 10 possible pairings of five duplicate readings were all between "*Good*" and "*Excellent,*" we shall focus upon cataract stage scores *averaged* over these five readings. We utilized a modal cataract score based upon the most frequently cited

TABLE 1
Levels of Intraobserver Agreement for Each of Five Observers Across Each of Five Independent Readings of the Same Cataract Slide

Observer 1		Observer 2		Observer 3		Observer 4		Observer 5	
ORP	Weighted Kappa	ORP	Weighted Kappa	ORP	Weighted Kappa	ORP	Weighted Kappa	ORP	Weighted Kappa
4,5	.92	1,2	.89	3,4	.90	3,4	.96	3,4	.94
2,3	.87	3,4	.83	4,5	.81	4,5	.96	1,5	.82
3,4	.84	1,3	.76	1,2	.80	2,5	.96	4,5	.81
3,5	.84	3,5	.74	1,3	.79	1,2	.96	2,3	.81
1,5	.80	2,3	.73	3,5	.77	3,5	.92	1,2	.77
1,3	.80	4,5	.72	1,4	.76	2,4	.92	2,4	.76
2,4	.79	1,4	.66	2,3	.74	1,5	.91	3,5	.75
2,5	.79	1,5	.65	2,4	.71	2,3	.88	1,3	.70
1,2	.74	2,4	.64	1,5	.71	1,4	.87	2,5	.70
1,4	.72	2,5	.63	2,5	.60	1,3	.84	1,4	.65
Means	.81		.73		.76		.92		.77

ORP = Ordering of Reading Pairs by levels of chance-corrected agreement Here < .40 = *Poor*; .40–.59 = *Fair*; .60–.74 = *Good*; .75–1.00 = *Excellent* Percentages of observed agreement (PO) ranged between 85.2 and 98.9.

cataract stage within each set of five replicate readings for a given slide. Thus, if the five readings made upon slide 1, by observer 2, were BAABA, respectively, then the modal stage score for that observer would be A. Although it was certainly possible for some observers to evaluate any given set of replicate slides such that a mode could not be calculated (e.g., ABABC or ABCDE) this never, in fact, occurred. The 10 possible pairings across the five observers were, once again, ordered on the basis of the size of chance-corrected interrater agreement (or weighted kappa) levels. These appear in Table 2. All levels of $\varkappa_w$ were highly statistically significant ($p < .0005$). Levels of clinical or practical significance of $\varkappa_w$ values ranged between .46 ("*Fair*") and .83 "(*Excellent*").

Rank Orderings of Observers

A third question we sought to answer is the extent to which the five observers were consistent in their scorings of cataract stages vis-à-vis each other. Another way of stating this is that we were interested in rank ordering the five observers from most to least reliable, given all the possible interobserver pairings. The ultimate aim here was to correlate rank ordering or consistency level with the extent of experience each observer had in both general morphologic classification systems as well as the more specific experience required for classifying stages of cataract development. As with previous analyses, we analyzed the data with respect to individual and average readings of cataract stages. Since the results were quite similar whether based upon each replicate reading or modal score, we shall focus again upon the latter.

For each of the five observers his levels of PO and PC, on modal reading, were obtained with each of the remaining observers, across the 22 cataract photographs. From these data, $\varkappa_w$ values were calculated from average PO and PC values. For example, observer 5 had $\varkappa_w$ values of .49, .49, .52, and .46 with observers 2, 1, 4, and 3, respectively (Table 2). The average $\varkappa_w$ value for observer 2 (derived from average PO and PC values) was then found to be .49. After this process was repeated for all five observers, their respective average $\varkappa_w$ values were then rank ordered from highest to lowest. The results in Table 3 show that the $\varkappa_w$ values for observers 1, 4,

TABLE 2
Levels of Agreement Among Five Observers Evaluating Stages of Cataract Development, on the Basis of Modal Stage[a]

Observer Pairings	Kappa[b] (Weighted)
2,4	.83
1,2	.75
1,4	.68
1,3	.67
3,4	.60
4,5	.52
2,3	.52
1,5	.49
2,5	.49
3,5	.46

[a] The most frequently occurring cataract stage among the five readings on a given slide (e.g., AAABB means stage A is the modal response).

[b] PO values ranged between 85.2 percent and 95.5 percent.

TABLE 3
Rank Orderings of Chance-Corrected Interrater Agreement Levels ($\varkappa_w$ values) based upon Modal Cataract Stages, Across 22 Cataractous Slides

Observer	Average Chance-Corrected Agreement Level ($\varkappa_w$) With Remaining Observers
1	.66 (*Good*)
4	.66 (*Good*)
2	.65 (*Good*)
3	.57 (*Fair*)
5	.49 (*Fair*)

and 2 were interchangeable (i.e., .66, .66, and .65, respectively). Each of these values indicates a level of "*Good*" average interrater agreement, corrected for chance. The remaining two observers (3 and 5) evidenced average $\varkappa_w$ values of .57 and .49, respectively, which reflect levels of "*Fair*" agreement, in a clinical or practical sense.

DISCUSSION

These data appear rather straightforward. Both trained and untrained observers are quite consistent in classifying replicate cataract photographs into the same stage of development upon repeated readings made in a randomized order. These range between 85.2 percent and 98.9 percent agreement, with chance-corrected levels considered "*Excellent*" (.76–.92) or very "*Good*" (.73).

However, with respect to levels of interobserver agreement, those with more training tend to be more consistent observers than those with less training. Our two most inexperienced observers, 3 and 5 (Table 2) were a postdoctoral fellow junior technician with no general experience in classifying morphologic variables and only occasional or very limited experience in classifying human cataract stages (observer 3); and a postdoctoral fellow with neither type of experience (observer 5). The remaining three observers (as previously noted) can be classified or rank ordered on the basis of experience as follows: the most senior (observer 1, author EC) had more than five years of experience both with respect to general ophthalmologic classification methods and the specific classification of human cataracts; the next most experienced examiner (observer 4) was a senior technician with two years of specific classificatory experience but no general experience; the next in experience level was a board-certified postdoctoral fellow with general knowledge about ophthalmologic classification procedures but none specific to differentiating among stages of human cataract development (observer 2).

In summary, then, this observer variability study, investigating the classification of human stages of cataract development among elderly males, indicates the following: (*1*) Both experienced and inexperienced observers are capable of very high degrees of intraexaminer reliability (or chance-corrected agreement levels), that is, ranging between very "*Good*" and "*Excellent*"; and (*2*) Experience level appears highly correlated with levels of interobserver agreement (that is to say, the more experienced the more consistent in the classification of cataract photographs vis-à-vis other examiners viewing the same materials).

In the context of our research program, this *in vitro* visual classification system has been quite valuable for demonstrating the relationship between aspirin intake and the apparent retardation of the development of human cataracts [10]. However, one basic limitation of this research is that it is based upon retrospective data with its known potential for introducing uncontrolled sources of bias [20-21]. In order to help resolve this problem, we have recently designed a prospective, double-blind study of the effects of aspirin on the frequency and extent of cataract development.

Viewed in a more general methodologic context of studies assessing the reliability of clinical signs and symptoms [3-4], our results suggest that a visual classification system, based upon purely qualitative criteria, can be quite reliable, relative to many other areas previously investigated, including even some of the more quantitative visual variables [6].

ACKNOWLEDGEMENT

We thank Charles James (M.R.C.P.), Bart Toftness, and Tracy Nissen for their collaboration in this research project.

REFERENCES

1. Cicchetti DV, Conn HO: A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. Yale J Biol Med 49:373-383, 1976
2. Etter LE, Dunn JP, Kammer AG, et al: Gastroduodenal X-ray diagnosis: A comparison of radiographic technics and interpretations. Radiology 74:766-770, 1960
3. Koran LM: The reliability of clinical methods, data and judgments. N Engl J Med 293:642-644, 1975
4. Koran LM: The reliability of clinical methods, data and judgments. N Engl J Med 293:695-701, 1975
5. Chylack LT: Classification of human cataracts. Arch Ophthalmol 96:888-892, 1978
6. Kahn MA: Diagnostic standardization. Clin Pharmacol Ther 25:703-711, 1979
7. Kahn HA, Liebowitz HM, Ganley JP, et al: The Framingham eye study: I. Outline and major prevalence findings. Am J Epidemiol 106:17-32, 1977
8. Marcantonio JM, Duncan G, Davies PD, et al: Classification of human senile cataracts by nuclear colour and sodium content. Exp Eye Res 31:227-237, 1980
9. Smith VC, Pokorny J, Starr SJ: Variability of color mixture data—I. Inter-observer variability in the unit coordinates. Vision Res 16:1087-1094, 1976
10. Cotlier E: Senile cataracts: Evidence for acceleration by diabetes and deceleration by salicylate. Can J Ophthalmol 16:113-118, 1981
11. Pirie A: Colour and solubility of the proteins of human cataracts. Invest Ophthalmol 7:634-642, 1968
12. Cicchetti DV: Assessing inter-rater reliability for rating scales: Resolving some basic issues. Br J Psychiat 129:452-456, 1976
13. Hall JN: Inter-rater reliability of ward rating scales. Br J Psychiat 125:248-255, 1974
14. Cohen J: A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37-46, 1960
15. Cohen J: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213-220, 1968
16. Fleiss JL, Cohen J, Everitt BS: Large sample standard errors of kappa and weighted kappa. Psychol Bull 72:323-327, 1969
17. Cicchetti DV, Sparrow SS: Developing criteria for establishing inter-rater reliability of specific items: Applications to assessment of adaptive behavior. Am J Ment Def 86:127-137, 1981
18. Fleiss JL: Statistical Methods for Rates and Proportions. New York, Wiley, 1981
19. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 33:159-174, 1977
20. Dorn H: Some problems arising in prospective and retrospective studies of the etiology of disease. N Engl J Med 261:571-579, 1959
21. Feinstein AR: The epidemiologic trohoc, the ablative risk ratio, and 'retrospective' research. In Clinical Biostatistics. St. Louis, Missouri, CV Mosby, 1977 (Chapter 14)