

Reliability of the Interpretation of Coronary Angiography by the Simple Visual Method

Jorge Augusto Nunes Guimarães, Edgar Guimarães Victor, Maria do Rosário de Britto Leite, José Maria Pereira Gomes, Edgar Victor Filho, Jesus Reyes Liveras

Recife, PE - Brazil

Objective – Evaluation of inter and intraobserver reproducibility of by the visual method interpretation of cineangiogram in a clinically based context.

Methods – Five interventional cardiologists analyzed 11 segments of 8 coronary cineangiograms at a two month apart sessions. The percent luminal reduction by the lesions were analyzed by two different classifications: in one (A) the lesions were graded in 0% = absent, 1-50% = mild, 51 - 69 = moderate, and $\geq 70\%$ = severe; the other classification (B) was a dichotomic one : $< 70\%$ = nonsignificant and $\geq 70\%$ = significant lesions. The agreement were measured by the kappa (k) index.

Results - Interobserver agreement was moderate for classification A (1st measurement, $k = 0.36 - 0.63$, $k_m = 0.49$; 2nd measurement, $k = 0.39-0.68$, $k_m = 0.52$) and good for classification B (1st measurement, $k = 0.55-0.73$, $k_m = 0.63$; 2nd measurement, $k = 0.37-0.82$, $k_m = 0.61$). Intraobserver levels of agreement were $k = 0.57-0.95$ for classification A and $0.62-1.0$ for classification B.

Conclusion – The higher level of reproducibility obtained by adopting the dichotomous criteria usually considered for ischemic limits demonstrates that in the present clinical context, the reliability of the simple visual method is adequate for the identification of patients with clinically significant lesions and candidates for myocardial revascularization procedures.

Key words: coronary angiography, observer variation, kappa statistic

Any measurement procedure, has some degree of variability in its results. In scientific investigation, one of the criteria that evaluate measurement variability is reliability (reproducibility, precision), defined as the capacity to produce agreeing results for measuring procedures repeated over time or when the same phenomenon is measured by several individuals at the same time^{1,2}. Reliability can be estimated from the agreement between the analysis of a phenomenon by different examiners (interobserver agreement) or by the consistency of the results obtained by repeated analysis of the phenomenon by the same examiner (intraobserver agreement)¹.

In the field of cardiology, cinecoronariography remains as the main method for diagnosis of atherosclerotic coronary disease and for definition of their therapeutic strategies³, however, there is relevant questions about its reproducibility and accuracy. The observation and estimation of the magnitude of an obstructive lesion usually depend to a great extent on the opinion of one sole clinician who analyses the result of his own procedures. Such involvement can lead to underestimation of some factors that influence the operator and may limit the credibility of the results obtained⁴.

Studies on the reliability of simple visual interpretation of cinecoronariography arose in the 1970s, but are still relatively scant. Methodological differences between data and indexes used to evaluate reliability render effective comparisons between them difficult⁵⁻⁹. The large margin of variability, especially for results expressed in terms of percent values, has become evident. Although some authors¹⁰⁻¹³ have suggested alternative approaches aimed at increasing the precision of visual estimates, application of quantitative digital analysis has become a standard in scientific literature¹³⁻¹⁸. Yet, the use of this method in clinical practice is not automatic. As a diagnostic method, the most important information derived from cinecoronariography (CCG) is above the presence of obstructive coronary atherosclerotic disease capable of evoking myocardial ischemia. Gould et al¹⁹⁻²¹ identified ischemic limits correlating the

Hospital das Clínicas UFPE and Real Hospital Português, Recife.
Mailing address: Jorge Augusto Nunes Guimarães - Rua Alfredo Fernandes, 136/
401 - 52060-320 - Recife, PE - Brazil. e-mail: jang@realcor.com.br

degree of coronary lumen obstruction and coronary flow reserve. These results influenced the tendency to downgrade the precision of the quantification of lesions of less than 50% obstruction^{4,22,23}, by describing them in classes according to their magnitude as discrete, moderate and severe. Furthermore, several studies have revealed important limitations of the routine application of quantitative digital analysis to diagnostic procedures^{4,23-27}.

At this time, although equipment for digital quantitative cineangiography analysis is present in the major of cathlabs, the great majority of examinations are still interpreted by the same specialist who performed them using the traditional form of visual examination. In this case an incongruity between evidence of the scientific reliability of the method and its use in clinical practice becomes noticeable. Furthermore, the reliability of the way in which the visual method is routinely applied for diagnostic purposes has not been adequately evaluated.

The purpose of the present study was to evaluate the reproducibility of the interpretation of CCG by the simple visual method performed by interventional cardiologists, using a model capable of estimating inter- and intraobserver agreement in a clinically based context.

Methods

Five of 17 interventional cardiologists, actives in nine public and private hospitals in Recife were chosen as observers by one of this work's authors, who also selected from his personal files 23 CCGs of patients who had not had myocardial revascularization. The only prerequisites of the chosen CCGs were that they showed multiarterial coronary atherosclerotic disease of any degree and to be of an adequate technical quality to show arterial opacity, and with projections sufficient to enable a clear identification of the various arterial segments.

All examinations were performed by the Judkins²⁸ technique with 6F catheters in a cineangiographic Philips, Poly Diagnostic U.P.I. apparatus with a 6.5 in image intensifier, at the Real Hospital Português de Beneficência in Recife, Pernambuco. Images were recorded on 35mm Kodak® CFT cinefilms at 30 frames/sec. Eight exams were chosen to obtain an ample spectrum of lesion sizes in all segments to be studied.

Coronary arteries were divided into 11 segments for film analysis (chart I).

The five observers were codified and kept anonymous during the study. Each received a form containing a table with lines for the recording of the 8 films and columns for each of the 11 segments referred to in chart 1. No clinical information of the patients was provided; identification records of the cinefilms were obscured.

Analyses were performed using a TAGARNO® 35CX film projector, individually and independently, without a time limitation. Each observer recorded a single 0 to 100 percent obstruction value for each segment. In cases where more than one lesion was observed in a segment the highest

Chart I - Segments of coronary arteries selected for analysis.

- 1) Left main coronary artery
- 2) Proximal half of left descending artery (ostium included)
- 3) Distal half of left descending artery
- 4) Main diagonal branch regarding extension and caliber
- 5) Proximal segment of the circumflex artery (ostium included)
- 6) Atrio-ventricular segment of the circumflex artery
- 7) Main marginal branch of the circumflex artery regarding extension and caliber
- 8) Proximal half of the right coronary artery (ostium included)
- 9) Distal half of the right coronary artery (ostium included)
- 10) Descending posterior branch of the right coronary
- 11) Ventricular posterior branch of the right coronary

value was recorded. To avoid visual fatigue and to minimize the natural tendency of observers to dedicate greater attention to the first cinefilms examined possibly leading to a biased interpretation of the last ones, analyses were made in two times with the interpretation of four films in each. The sequence of observations of the films was performed informally.

Using the same above described protocol, each observer re-evaluated the films after a minimum period of two months. None knew the values annotated by the other observers in either step or his own values prior to initiating the 2nd step.

It was emphasized to observers that they should avoid merely classifying or placing lesions between border values. In cases where they were eventually unable to establish a percentage value for a given lesion, they were allowed to place, at appropriate sites, a question mark (?) followed by an interpretation of the degree of obstruction as being mild (mi), moderate (M) or severe (S).

For comparative analysis, two charts, one for each manner of classification of the lesions according to the grade of obstruction, were set up (chart II).

The analysis of the frequency distribution of the lesions indicated by each observer was performed by Friedman's²⁹ test with four classes of variables for the A classification A and by Cochran's³⁰ test with binary variables for the classification B.

Pairing of the 5 observers permitted the formation of 10 pairs of combinations, whose contingency tables were used for the statistical analysis of each classification of lesions in both phases of the study. The general rate of agreement, defined as the proportional agreement between observers relative to the whole sample, was calculated for each classification in both stages. The kappa (k) statistic, defined as the proportion of the observed agreement over the agreement expected by chance, is expressed by formula^{30,31}

$$k = \frac{(p_o - p_e)}{(1 - p_e)}$$

where p_o = proportion of observed agreements and p_e = proportion of expected agreements.

This calculation takes all discrepancies into account in the same way. When more than two categories are aligned,

Chart II - Classifications of lesions according to the degree of obstruction in the coronary segments selected						
Segment	Classification A				Classification B	
	Absence	Mild	Moderate	Severe	Non significant	Significant
	(Z)	(mi)	(M)	(S)	(N)	(S)
LM*	0%	1%-50%	-	≥50%	<50%	≥50%
Others	0%	1%-50%	51%-69%	≥70%	<70%	≥70%

*LM- left main coronary artery.

degrees of discrepancy between contiguous or more distant categories may have different clinical relevance. To correct for the evaluation of these discrepancies, the weighted kappa index³¹ was used to measure agreement between the four categories defined in classification A.

Criteria for the interpretation of values of kappa are described in chart III.

The level of statistical significance of the differences between kappa indexes of pairs of observers and between the values of each observer in both classifications was analyzed by the paired *t* test, with $\alpha=0.05$.

Charts were developed and statistical calculations performed with the aid of computer programs Microsoft® Excel – version 8.0, Epi Info – 6.02 and SPSS for Windows – 6.0.

Results

Each observer individually interpreted 11 coronary segments in the eight cinefilms on two independent occasions; a total of 440 observations were made in each step of the study. Only in 8 (0.9%) segments of the 880 evaluations were no percent obstruction values assigned.

Frequency distributions of the values determined according to the degree of obstruction defined by classification A in the 1st step of the study are described in Table 1. Observers 1, 2, 3 and 5 determined that the majority of the segments had no obstructive lesions (grade Z), with incidences between 62.5% and 73%, but for observer 4 the majority of the segments had mild lesions (grade Mi, 60.2%). All five observers determined that moderate (grade M) lesions were least frequent, with an incidence varying between 0% (observer 2) and 4.6% (observers 4 and 5). Severe lesions (grade S) were estimated at 8% (observer 1) and 17%

(observer 5). These differences were statistically significant (Friedman's test, $P<0.0001$). In the 2nd step of the study (table II), the average frequency of segments considered to be grade Z decreased due to the smaller proportion attributed to it by observers 1, 2, 3 and 5 (46.6; 68.2; 64.8; 67.1 %, respectively). Again, observer 4 interpreted the majority of the segments to be lesions of grade Mi (55.7%), and all observers considered that lesions of grade M occurred at a low frequency, with an incidence varying between 1.1% (observer 2) and 5.7% (observers 4 and 5). Lesions of grade S were considered to fall between 10.2% (observer 4) and 22.7% (observer 5). The results of the analysis of these differences were also statistically significant (Friedman's test, $P=0.00015$).

Distribution of the frequencies of the lesions attributed by the five observers according to classification criterion B in the first step are described in table III. Observers 1, 2, 3 and 4 pointed towards an incidence between 8% and 12.5% of significant (grade S) lesions, but observer 5 described them in 17% of the cases. The analysis of these diffe-

Chart III - Interpretation of kappa values*	
kappa	Level of agreement
0	Equal to chance
0 < k ≤ 0.20	Poor
0.21 < k ≤ 0.40	Fair
0.41 < k ≤ 0.60	Moderate
0.61 < k ≤ 0.80	Good
0.81 < k ≤ 0.99	Excellent
1	Perfect

* According to ref. 2.

Table I – Frequency of lesion values attributed by the five observers according to classification A for the lesion grade (Z- absent=0%; Mi-Mild=1%-5%; M- moderate=51%-69%; S- severe ≥ 70%) in the first step of the study						
Grade	Observers					$\bar{X} \pm SD$
	1	2	3	4	5	
Z	64 (72.7) *	65 (73.9)	55 (62.5)	22 (25)	62 (70.5)	53.6±18.1
Mi	16 (18.2)	14 (15.9)	19 (21.6)	53 (60.2)	7 (7.9)	21.8±18.0
M	1 (1.1)	0	3 (3.4)	4 (4.6)	4 (4.6)	2.4±1.8
S	7 (8)	9 (10.2)	11 (12.5)	9 (10.2)	15 (17)	10.2±3.0
Total	88 (100)	88 (100)	88 (100)	88 (100)	88 (100)	

Friedman's test, $P<0.0001$; * n (%); $\bar{X} \pm SD$ - mean ± standard deviation.

Table II - Frequency of lesion values attributed by the five observers according to classification A for the lesion grade (Z- absent=0%; Mi-Mild=1%-5%; M-moderate=51%-69%; S- severe ≥ 70%) in the second step of the study

Grade	Observers					$\bar{X} \pm SD$
	1	2	3	4	5	
Z	41 (46.6) *	60 (68.2)	57 (64.8)	25 (28.4)	59 (67.1)	48.4±15.2
Mi	34 (38.6)	17 (19.3)	18 (20.4)	49 (55.7)	4 (4.5)	24.4±17.4
M	3 (3.4)	1 (1.1)	3 (3.4)	5 (5.7)	5 (5.7)	3.4±1.7
S	10 (11.4)	10 (11.4)	10 (11.4)	9 (10.2)	20 (22.7)	11.8±4.6
Total	88 (100)	88 (100)	88 (100)	88 (100)	88 (100)	

Friedman's test, P<0.00015; * n (%); $\bar{X} \pm SD$ - mean ± standard deviation.

rences was statistically significant (Cochran's test, P=0.03). In the 2nd stage of the study (Table IV), this pattern repeated itself with a more marked difference between the results of observers 1, 2, 3, and 4 (grade S in 10.2 to 11.4%) and those of observer 5 (grade S = 22.7%). These differences were highly significant (Cochran's test, P=0.0008).

Interobserver agreement – Table V describes weighted kappa indexes among 10 combinations of pairs of observers, calculated to measure the agreement regarding grade of lesion according to classification A in the study's two steps, and the respective general agreement rates (GAR). In the 1st step, the GAR of the 10 possible combinations between the 5 observers varied between 38% (observer 4 vs. observer 5) and 81% (observer 1 vs. observer 2). Weighted kappa indexes varied between 0.36 (observer 4 vs. observer 5) and 0.63 (observer 1 vs. observer 2). In the 2nd step, the GAR varied between 42% (observer 4 vs. observer 5) and 78% (observer 2 vs. observer 3). Weighted kappa indexes varied between 0.39 (observer 3 vs. observer 5) and 0.68 (observer 1 vs. observer 4). Differences between kappa indexes of the 10 combinations in both steps were not statistically significant (paired t test, P=0.62).

Kappa indexes between the 10 combinations of pairs of observers calculated to measure agreement regarding lesion grade by classification B in the two steps of the study and their respective GAR are described on table VI. In the first step, GAR of the 10 combinations varied between 91% (observer 1 vs. observer 5; observer 2 vs. observers 3 and 5, observer 3 vs. observers 4 and 5; observer 4 vs. observer 5) and 95% (observer 1 vs. observers 2 and 4). kappa indexes varied between 0.37 (observer 3 vs. observer 5) and 0.82 (observer 1 vs observer 4). In the same manner as in relation to classification A in the 2nd step, the GAR varied from 42% (observer 3 vs. observer 5) and 97% (observer 1 vs. observer 4) and kappa indexes varied from k = 0,37 (observer 3 vs. observer 5) and k = 0,82 (observer 1 vs. observer 4) differences between kappa indexes of the 10 combinations were not statistically significant in both steps (paired t test, p=0.65).

With the objective of estimating a general index of agreement between all observers, by each classification, the average of the 10 combined analyses of the kappa indexes was considered for each step. In relation to classification A, the averages of the weighted kappa indexes were $k_{m=}$ 0.49 and $k_{m=}$ 0.52 in the first and second step, respective-

Table III - Frequency of lesion values attributed by the five observers according to classification B for the lesion grade (N- non-significant <70%; S- significant ≥ 70%) in the first step of the study

Grade	Observers					$\bar{X} \pm SD$
	1	2	3	4	5	
N	81 (92) *	79 (89.8)	77 (87.5)	79 (89.8)	73 (83)	77.8±3.0
S	7 (8)	9 (10.2)	11 (12.5)	9 (10.2)	15 (17)	10.2±3.0
Total	88 (100)	88 (100)	88 (100)	88 (100)	88 (100)	

Cochran test, P = 0.003; * n (%); $\bar{X} \pm SD$ - mean ± standard deviation.

Table IV - Frequency of lesion values attributed by the five observers according to classification B for the lesion grade (N- non-significant <70%; S- significant ≥70%) in the second step of the study

Grade	Observers					$\bar{X} \pm SD$
	1	2	3	4	5	
N	78 (88.6) *	78 (88.6)	78 (88.6)	79 (89.8)	68 (77.3)	76.2±4.6
S	10 (11.4)	10 (11.4)	10 (11.4)	9 (10.2)	20 (22.7)	11.8±4.6
Total	88 (100)	88 (100)	88 (100)	88 (100)	88 (100)	

Cochran test, P- 0.0008; * n (%); $\bar{X} \pm SD$ - mean ± standard deviation.

Table V – General agreement rates (GAR) and weighted kappa indexes according to classification A of the degree of lesion, by pairs of observers (OBS) in the two stages of the study

Pairs of OBS	1 st Step		2 nd Step	
	GAR	kappa	GAR	kappa
1-2	81	0.63	72	0.59
1-3	70	0.54	66	0.59
1-4	43	0.37	70	0.68
1-5	70	0.48	57	0.45
2-3	74	0.55	78	0.66
2-4	44	0.39	52	0.45
2-5	80	0.61	70	0.51
3-4	45	0.40	47	0.42
3-5	73	0.60	66	0.39
4-5	38	0.36	42	0.41
Mean	62	0.49	62	0.52

Paired *t* test, weighted kappa 1st step vs.2nd step, P=0.62.

Table VI – General agreement rates (GAR) and weighted kappa indexes according to classification B of the degree of lesion, by pairs of observers (OBS) in the two stages of the study

Pairs of OBS	1 st Step		2 nd Step	
	GAR	kappa	GAR	kappa
1-2	95	0,73	92	0,62
1-3	93	0,63	95	0,77
1-4	95	0,73	97	0,82
1-5	91	0,59	84	0,45
2-3	91	0,55	92	0,62
2-4	93	0,63	93	0,66
2-5	91	0,62	88	0,58
3-4	91	0,55	94	0,71
3-5	91	0,64	82	0,37
4-5	91	0,62	85	0,48
Mean	92	0,63	90	0,61

Paired *t* test, weighted kappa 1st step vs.2nd step, P=0.65.

ly. By applying the criteria defined in chart III, the level of agreement between observers on interpreting the absence or three possible categories of lesion degrees (mild, moderate or severe) was characterized as moderate for both steps. In relation to step B, averaged kappa indexes were $k_m = 0.63$ and $k_m = 0.61$ for the first and second steps, respectively. Therefore, the agreement between observers when evaluating the presence or not of a clinically significant lesion, reached a good level in both steps.

The extent of the variation of the kappa indexes of the 10 pairs of observers in both classifications in both steps of the study is represented in figure 1. It will be observed that in classification A, the variation was similar for both steps, differing from the pattern observed for classification B, which was noticeably wider in the 2nd step.

The performance of each observer in the results concerning the presence or absence of a significant lesion (Classification B) was subsequently analyzed using kappa indexes of pairs of observers in which the observer was included. Each observer participated in four pairs of combinations and the averaged kappa (k_m) of each relative to the

others in both steps of the study, is described in figure 2. In the first step, these averages varied between 0.59 (observer 3) and 0.67 (observer 1). In the second step, the averages varied between 0.47 (observer 5) and 0.67 (observers 1 and 4). The analysis of these data confirms the homogeneous pattern of the interpretation of the observers in the first step and indicates that one of the observers (observer 5) was responsible for the broader amplitude of the kappa indexes verified in the 2nd step of the study.

Intraobserver agreement – kappa indexes calculated to determine the reproducibility of the evaluations by each observer in the two steps of the study are demonstrated in table VII. In relation to classification A, results varied between $k=0.57$ (observed 1) and $k=0.95$ (observer 4). Concerning classification B, results varied between 0.62 (observer 3) and 1.0 (observer 4). Analysis of the differences between the values obtained by each observer; in both classifications; demonstrated that the level of intraobserver agreement was significantly higher when they judged the presence or not of clinically significant lesions than when they evaluated them according to a greater number of categories of the degree of obstruction (paired *t* test, P=0.03).

Regarding classification A, observers 2, 3 and 5 achieved good levels of agreement and observer 4 reached an excellent kappa level; observer 1 obtained a moderate level. Relative to classification B, observers 3 and 5 maintained their good level, but the others improved the levels of reproducibility of their own evaluations (observer 1: good; observer 2, excellent; observer 4: perfect).

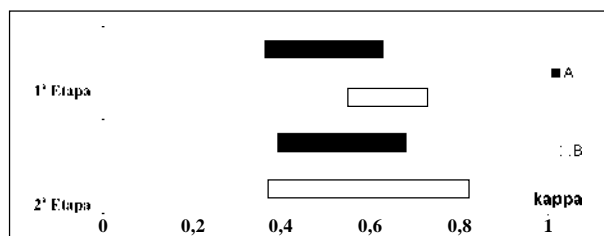


Fig. 1 - Extent of the variation of kappa indexes between 10 combinations of pairs of observers regarding grade of classifications A and B of the lesions in both steps of the study

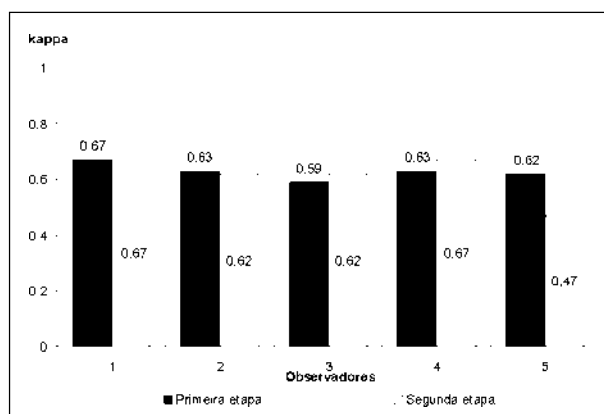


Fig. 2 - Averages of kappa indexes from each observer in relation to the others, regarding classification B of the grades of obstruction, in the two steps of the study.

Table VII – kappa indexes between evaluations by the same observer, in the two steps of the study, relative to classifications A and B of the lesion grade

Classification	Observer				
	1	2	3	4	5
A	0.57	0.76	0.61	0.95	0.62
B	0.67	0.89	0.62	1	0.68

Paired *t* test, classification A vs. classification B; P=0.03.

With the objective of analyzing whether the pattern of intraobserver agreement influenced the interobserver agreement in relation to the identification of clinically significant lesions (classification B), the following results were grouped for each observer: kappa indexes of intraobserver agreement and its averaged kappa indexes relative to the other interobserver agreements in both steps of the study (table VIII). It was noted that the level of interobserver agreement was relatively homogeneous in both steps of the study with the exception of that obtained by observer 5 in the 2nd step. The indexes that measured intraobserver agreement varied in a broader way from good and perfect. We did not find relationship between intraobserver level of consistency and level of interobserver agreement. Observer 4, whom reached maximal reproducibility in his interpretations in both steps, obtained an average similar to those of the others regarding interobserver agreement. On the other hand, observer 5, in spite of having an average lower than the others in the 2nd step, had a good level of reproducibility of his own interpretations, similar to the level reached by observers 1 and 3.

Discussion

Few studies have been designed especially to analyze the reliability of the visual interpretation of cinecoronariography^{5-8, 32,33}. Our results are similar to those observed in other studies^{9,11,22,34}, however their objective and design were much different so hampering the determination of a suitable pattern for estimating the precision of the method employed. Under the impact of the appearance of coronary angioplasty, several studies proposing to enhance reliability

Table VIII – Kappa indexes from each observer in relation to his interpretations in the two steps of the study (INTRA) and averages of kappa indexes of each observer relative to those of the others in the first and second stages of the study (INTER) regarding the presence or not of a significant lesion (classification B)

Observer	Intra	Inter	
		1 st Step	2 nd Step
1	0.67	0.67	0.67
2	0.89	0.63	0.62
3	0.62	0.59	0.62
4	1	0.63	0.67
5	0.68	0.62	0.47

of the interpretation of CCG have been published^{10-13, 15-18}. Technological developments led to improvement in quantitative digital analysis methodology and evolution of their indexes of reproducibility^{14,27,35,36}. Differently from the visual method^{22,37}, digital quantification leads to the attainment of normal distribution curves for its measurements. At the present time, even considering the limitations pointed out in the medical literature^{4,23-26}, the demand for digital quantification in research based on angiographic interpretation of coronary atherosclerosis has become consensual.

The use of CCG as a complementary diagnosis tool has however a different meaning. In clinical practice, its major area of information concerns the presence or not of obstructive disease able to cause myocardial ischemia, even though a tendency to disregard lesions of less than 50% exists⁴. According to Fleming et al²² observers tend to categorize lesions by the visual method, even when their objective is to quantify percent obstruction. In his study, this fact is shown to result in greater variability in lesions interpreted as mild (<50%) and in a tendency towards underestimation, in comparison with results of quantitative digital analysis. A similar conclusion has been formulated by Gurley et al²³.

What should be compared and how should it be collected? – The value of percent obstruction, by being a continuous variable, permits the calculation of variability indexes based on the standard deviation of their means. Some authors suggest that the standard error of the estimated percent value of a given lesion could be used as a numerical parameter of the variation of the method, describing variability indexes of 28-36%^{7,11,32}.

However, when we analyzed for example, the results of Derouen et al⁷ in which standard deviations of the segments analyzed varied between 0 and 51.3%, the generalization based on the obtained mean (18%) appears to be of a debatable practical utility.

The present study was planned to evaluate the method of visual CCG in the way it is currently routinely performed, favoring the comparison based on class variables and adopting as major references the best known values for these ischemic limits – obstructive lesion ≥50% in the trunk of the left coronary and ≥70% in the other arteries. In reality, simple observation in tables I and II of the high values of the standard deviations of the frequencies attributed to grades Z and D compared to with the lower frequencies given to degrees M and S indicates that the observers varied more when quantifying clinically insignificant lesions (≤50% obstruction), confirming results commented on above.

Studies on the variability in CCG evaluated in terms of time of activity or experience of the observers, either did not demonstrate significant differences related to these criteria^{11,22} or found a positive correlation with the maintenance of regular activity in the area⁵. The five observers chosen for the present study make up 29.4% of the total number of professionals qualified by the Brazilian Society of Hemodynamics and Interventional Cardiology regularly active in the various clinics of the city of Recife.

The chosen protocol adopted a model similar to daily practice. Cinefilms were selected without major restrictions and the observers handled them freely. With the criteria defined by classification A with four classes of variables, we tried to establish the level of agreement regarding the most detailed evaluation of the degree of obstruction. The main objective of the present study was to analyze the agreement rate when the dichotomous criterion to classify lesion (classification B) is used.

Which is the most adequate index of reliability? – The use of different indexes to measure agreement is an important limitation for the comparison between studies. The simplest form to evaluate the agreement between categorical variables, the general agreement rate has been adopted by some authors^{5,6,9} in spite of important restrictions concerning the significance of their results. The general degree of agreement reached does not identify the proportion of cases in which chance was responsible for agreement and is influenced by the proportion of positive findings and cannot be compared with degrees of agreement resulting from other studies^{31,38}. In the present work, these rates were shown with merely descriptive objectives. Actually, values such as those obtained, for example, in classification B (table VI), with a general degree of agreement between 91 and 95% in the first step and between 82 and 97% in the second, do not permit clear interpretations of the quality of the level of agreement between observers.

The kappa index is a coefficient that excludes chance when the agreement between paired observations is calculated, enabling the qualification of the degree of agreement and the comparison with indexes obtained from other studies^{2,31}. The main criticism to its application in the present study is the impossibility of estimating the agreement of the results of all of the observers, because it deals with an index that measures agreement between pairs of observers. With the aim of obtaining a general idea of the agreement of all of the observers, we adopted the principle of the mean of the indexes, a resource applied by others in studies cited in the medical literature^{7,34}.

Weighted kappa index is recommended to minimize the discrepancies between levels of disagreement in situations in which more than two variables are considered^{30,31}. The manner of weighting the indexes in classification A, however, does not correct for the distortions regarding the clinical relevance of these disagreements. Disagreement on the attribution of moderate or severe degrees to a given lesion, has greater therapeutic and prognostic implications than disagreement about moderate or discrete degrees. Only an arbitrary intervention in the application of weights would correct such discrepancies; this, however, would affect the possibility of comparing the results of this study with those of others.

Interobserver agreement – Detre et al⁵ considering a criterion of a significant lesion to be an obstruction of $\geq 50\%$, applied indexes derived from the standard deviation of the positive findings and concluded that the levels of agreement among observers resided in the middle between

perfect agreement and that due to chance. Correlating with the kappa function, an index of 0.50 would be considered as moderate. The same level ($k = 0.55$) was shown by Derouen et al⁷ whose criterion was similar to that adopted by us in classification B (significant lesion $\geq 70\%$).

Although intended to compare the visual method with the use of the *caliper*, the study related by Holder et al³⁴ yielded results amenable to comparison with ours. Using means of weighted kappa indexes, the agreement between the five observers classifying lesions in three categories was of a good level ($k = 0.62$). kappa averages of each observer relative to those of the others were similar for all observers.

In CASS a study there was a moderate level of agreement about the number of significant lesions per exam ($k = 0.57$)⁸. Although the criterion for significant lesions was similar to ours, this index actually reflected the agreement between interpretations of one of fourteen participating clinical centers and those of four quality control centers.

In our study, the repetition of the analyses of the five observers allowed levels of interobserver agreement to be established between 10 pairs of observers in each step, as well as between these same pairs in both steps. Results demonstrated that to identify in greater detail the degree of coronary obstruction (classification A), the simple visual method of CCG interpretation achieved a merely moderate level of reproducibility between results of different observers. However, it resulted in a good level of reproducibility for the estimation of the existence or not of obstructions capable of causing ischemia (classification B).

Intraobserver agreement – In the study by Detre et al⁵ intraobserver agreement was estimated from the general rate of agreement and varied between 72 and 91%. Holder et al³⁴ applying weighted kappa indexes, described intraobserver levels of consistency from moderate to good ($k = 0.57$ to 0.79).

In our study, observers were more consistent in their own evaluations when the dichotomous criterion (B) was applied. The results demonstrated that all observers reproduced their interpretations with a minimum level considered good regarding the presence or not of a significant lesion. In reality, as described on table VII, one observer obtained an excellent level [observer 2: ($k = 0.89$)] and another reached maximal agreement [(observer 4: ($k = 1.0$))]. This qualitative difference of the intraobserver reproducibility did not agree with the more homogeneous pattern presented by the observers regarding the agreement between them (table VIII). At the other extreme, the lower index of observer 5 relative to the others in the 2nd step did not prevent him from achieving a good level of reproducibility of his own evaluation. These observations demonstrate that no relationship exists between the level of intra-observer consistency and the levels of interobserver agreement.

Final considerations – In the medical literature, quantitative digital angiography has become a standard procedure for angiographic interpretation of coronary arteries. In clinical practice, however, its routine application has important limitations and does not eliminate the operator's subjectivity in various stages of the examination and of selec-

tion of images and segments to be interpreted. The tendency to classify lesions according to the degree of obstruction, and to define conduct based on dichotomous criteria, renders the level of precision of the estimation of lesions below limits considered capable of evoking myocardial ischemia irrelevant.

The present study demonstrated that interpretation of cinecoronarioangiography (CCG) by the simple visual method based on kappa statistics-only reached a good level of reproducibility among interventional cardiologist when the adopted criterion was that routinely used to consider the indication of a treatment of myocardial revascularization (lesions clinically significant or not). On expressing opinions about grades of obstruction classically considered in clinical practice (absence of obstructions, discrete, moderate or severe lesions), only a moderate level of reliability between observers was noted. As expected, because it is more plausible for each individual to agree with himself

rather than with others, the level of intraobserver agreement was higher by each of the criteria adopted, although we did not find a relationship between the degree of precision among observers and the level of consistency of their own opinions.

In the clinical field, therefore, where the main objective is to diagnose and define the extent of coronary atherosclerotic disease in order to set up a therapeutic program, the present study demonstrated that the simple visual method, still most used in clinical practice, fulfills the requisites regarding reliability of its results.

Acknowledgements

To Profs. Eulálio Cabral and José N. Figueiroa for aid in the analysis of the results and to Prof. Sandra N. Coelho, MD for patience and judicious analyses during the elaboration and execution of this study.

References

1. Dawson-Saunders B, Trapp RG. Summarizing Data. In: Dawson-Saunders B, Trapp RG, eds. *Basic & Clinical Biostatistics*. 2nd ed. Norwalk: Appleton & Lange, 1994: 41-63.
2. Pereira MG. Aferição dos Eventos. In: Pereira MG, ed. *Epidemiologia - Teoria e Prática*. Rio de Janeiro: Guanabara Koogan, 1995: 358-76.
3. Bittl JA, Levin DC. Coronary Arteriography. In: Braunwald E, ed. *Heart Disease: A Textbook of Cardiovascular Medicine*. 5th ed. Philadelphia: WB Saunders Co., 1997: 240-72.
4. Stadius ML, Alderman EL. Coronary artery revascularization: Critical need for, and consequences of, objective angiographic assessment of lesion severity. *Circulation* 1990; 82: 2231-4.
5. Detre KM, Wright E, Murphy ML, Takaro T. Observer agreement in evaluating coronary angiograms. *Circulation* 1975; 52: 979-86.
6. Zir LM, Miller SW, Dismore RE, Gilbert JP, Harthorne JW. Interobserver variability in coronary angiography. *Circulation* 1976; 53: 627-32.
7. Derouen TA, Murray JA, Owen W. Variability in the analysis of coronary angiograms. *Circulation* 1977; 55: 324-8.
8. Fisher LD, Judkins MP, Lesperance J, et al. Reproducibility of coronary arteriographic reading in the Coronary Artery Surgery Study (CASS). *Cathet Cardiovasc Diagn* 1982; 8: 565-75.
9. Trask N, Califf RM, Conley MJ, et al. Accuracy and interobserver variability of coronary cineangiography: a comparison with postmortem evaluation. *J Am Coll Cardiol* 1984; 3: 1145-54.
10. Meier B, Gruentzig AR, Goebel N, Pyle R, Von Goslar W, Schlumpf F. Assessment of stenoses in coronary angioplasty: inter- and intraobserver variability. *Int J Cardiol* 1983; 3: 159-69.
11. Beauman GJ, Vogel RA. Accuracy of individual and panel visual interpretations of coronary arteriograms: Implications for clinical decisions. *J Am Coll Cardiol* 1990; 16: 108-13.
12. Danchin N, Juilliere Y, Foley D, Serruys PW. Visual versus quantitative assessment of the severity of coronary artery stenoses: can the angiographer's eye be reeducated? *Am Heart J* 1993; 126: 594-600.
13. Katrisis D, Lythall DA, Anderson MH, Cooper IC, Webb-Peploe MM. Assessment of coronary angioplasty by an automated digital angiographic method. *Am Heart J* 1988; 116: 1181-7.
14. Selzer RH, Hagerty C, Azen SP, et al. Precision and reproducibility of quantitative coronary angiography with application to controlled clinical trials: a sampling study. *J Clin Invest* 1989; 83: 520-6.
15. Goldberg RK, Kleiman NS, Minor ST, Abukhalil J, Raizner AE. Comparison of quantitative coronary angiography to visual estimates of lesion severity pre and post PTCA. *Am Heart J* 1990; 119: 178-84.
16. Kalbfleisch SJ, McGillem MJ, Pinto IMF, Kavanaugh KM, Deboe SF, Mancini GBJ. Comparison of automated quantitative coronary angiography with caliper measurements of percent diameter stenosis. *Am J Cardiol* 1990; 65: 1181-4.
17. Kimball BP, Bui S, Cohen EA, Cheung PK, Lima V. Systematic bias in the reporting of angioplasty outcomes: accuracy of visual estimates of absolute lumen diameters. *Can J Cardiol* 1994; 10: 815-20.
18. Desmet W, Willems J, Van Lierde J, Piessens J. Discrepancy between visual estimation and computer-assisted measurement of lesion severity before and after coronary angioplasty. *Cathet Cardiovasc Diagn* 1994; 31: 192-8.
19. Gould KL, Lipscomb K, Hamilton GW. Physiologic basis for assessing critical coronary stenosis. *Am J Cardiol* 1974; 33: 87-94.
20. Gould KL, Lipscomb K. Effects of coronary stenoses on coronary flow reserve and resistance. *Am J Cardiol* 1974; 34: 48-55.
21. Gould KL. Quantification of coronary artery stenosis in vivo. *Circ Res* 1985; 47: 341-53.
22. Fleming RM, Kirkeeide RL, Smalling RW, Gould KL. Patterns in visual interpretation of coronary arteriograms as detected by quantitative coronary arteriography. *J Am Coll Cardiol* 1991; 18: 945-51.
23. Gurley JC, Nissen SE, Booth DC, et al. Comparison of simultaneously performed digital and film-based angiography in assessment of coronary artery disease. *Circulation* 1988; 78: 1411-20.
24. Gurley JC, Nissen SE, Booth DC, Demaria NA. Influence of operator- and patient-dependent variables on the suitability of automated quantitative coronary arteriography for routine clinical use. *J Am Coll Cardiol* 1992; 19: 1237-43.
25. Herrington DM, Siebes M, Walford GD. Sources of error in quantitative coronary angiography. *Cathet Cardiovasc Diagn* 1993; 29: 314-21.
26. Herrington DM, Siebes M, Sokol DK, Siu CO, Walford GD. Variability in measures of coronary lumen dimensions using quantitative coronary angiography. *J Am Coll Cardiol* 1993; 22: 1068-74.
27. Jost S, Deckers J, Nikutta P, et al. Influence of the selection of angiographic projections on the results of coronary angiographic follow-up trials. International nifedipine trial on antiatherosclerotic therapy investigators. *Am Heart J* 1995; 130: 433-9.
28. Baim DS, Grossman W. Coronary angiography. In: Baim DS, Grossman W, eds. *Cardiac Catheterization, Angiography, and Intervention*. 4th ed. Philadelphia: Lea & Febiger, 1991: 185-214.
29. Altman DG. Relation between several variables. In: Altman DG, ed. *Practical Statistics for Medical Research*. London: Chapman & Hall, 1995: 325-64.
30. Armitage P, Berry G. Further Analysis of Categorical Data. In: Armitage P, Berry G, eds. *Statistical Methods in Medical Research*, 3rd ed. Oxford: Blackwell Scientific Publications, 1994: 402-447.
31. Altman DG. Some Common problems in medical research. In: Altman DG, ed. *Practical Statistics for Medical Research*. London: Chapman & Hall, 1995: 396-439.
32. Sanmarco ME, Brooks SH, Blankenhorn DH. Reproducibility of a consensus panel in the interpretation of coronary angiograms. *Am Heart J* 1978; 96: 430-7.

33. Kussmaul III WG, Popp RL, Norcini J. Accuracy and reproducibility of visual coronary stenosis using information from multiple observers. *Clin Cardiol* 1992; 15: 154-62.
34. Holder DA, Johnson AL, Stolberg HO, et al. Inability of caliper measurement to enhance observer agreement in the interpretation of coronary cineangiograms. *Can J Cardiol* 1985; 1: 24-9.
35. Serruys PW, Reiber JHC, Wijns W, et al. Assessment of percutaneous transluminal coronary angioplasty by quantitative coronary arteriography: diameter versus densitometric area measurements. *Am J Cardiol* 1984; 54: 482-8.
36. Reiber JH, Van Eldik-Helleman P, Visser-Akkerman N, Kooijman CJ, Serruys PW. Variabilities in measurement of coronary arterial dimensions resulting from variations in cineframe selection. *Cathet Cardiovasc Diagn* 1988; 14: 221-8.
37. Bertrand ME, Lablanche JM, Bauters C, Leroy F, Mac Fadden E. Discordant results of visual and quantitative estimates of stenosis severity before and after coronary angioplasty. *Cathet Cardiovasc Diagn* 1993; 28: 1-6.
38. Koran LM. The reliability of clinical methods, data and judgements (First part). *N Eng J Med* 1975; 293: 642-6.