STUDY ON INTEROBSERVER AGREEMENT IN A GOITRE PREVALENCE SURVEY

Lt Col PB PILLAI*, Lt Col GS SAIPRASAD*,
Mr DR BASANNAR+, Brig AG GOKARN#

ABSTRACT

Observers differ in their judgment while assessing physical signs in a patient. We had undertaken a goitre prevalence survey amongst school students in a Rural Health Training Centre, Pune district (Maharashtra) during October 1992. Four teams of trained observers were used for detection of goitre. This study was undertaken to estimate the extent and acceptability of interobserver agreement amongst the four teams. Observer variation/agreement was measured by two methods viz. kappa coefficient and proportion of agreement. The proportion of agreement appears to be a better measure of observer agreement as it could make a distinction between normality (absence of goitre) and abnormality (presence of goitre). In the present study, the proportion of agreement for abnormality ranged between 0.62 - 0.83. This measure was considered as indicating a good interobserver agreement in detecting goitre in the survey that was undertaken.

MJAFI 1995; 51 : 95-99

KEY WORDS : Confidence intervals; Epidemiologic methods; Health surveys; Observer variation.

Introduction

While carrying out a prevalence survey for goitre by a group of investigators, the validity and reproducibility of the observations are affected by inter- or intraobserver agreement in assessing the grades of goitre, characteristic of the neck of the patient, investigators' method of examining the neck and illumination during the examination [1]. The most important amongst the above being the inter- or intra-observer agreement. However, for the purposes of defining goitre as a public health problem, generally an error of not more than 25% in estimation of prevalence is accepted [2]. It was proved in a study of finger clubbing that no observer, however well qualified, was infallible and free of bias [3]. There is no gold standard in the case of physical examination of goitre against which observers' results can be compared. We planned a goitre prevalence survey amongst school children located in Rural Health Training Centre (RHTC) area,

Shirur, District Pune, Maharashtra, by a group of investigators. This study was undertaken to assess whether the observations of goitre grades made by these investigators were within acceptable limits. We also studied the validity of various statistical methods available to assess the inter- observer agreement or variation.

Methods

We had eight medical graduates, all practicing clinicians pursuing their postgraduate studies. They were divided into four teams of two each. All the teams were instructed regarding the purpose of the study, the technique of examination and the goitre classification [1,2] to be followed. All four teams were seated well away from each other so that each team could make an independent assessment.

A primary school with 99 students of a small village, Juna Shirur, Dist. Pune, was selected for the interobserver agreement

*Reader; +Scientist "C"; Dept of PSM, AFMC, Pune 411 040; #Deputy Commandant, Command Hospital, Western Command, Chandimandir.

study. (One student of the same age group, from the same village but not studying in the school was included to round off the figure to 100 to facilitate easy calculations). Each of the 100 children was assessed for goitre by all the four teams.

Assessment of goitre status and estimation of goitre prevalence

Goitres were classified by each team as grades 0, IA, IB, II or III. Average number of students in each grade was calculated after pooling the observations of all teams.

The grade of goitre given to each student separately by each of the four teams was tabulated. As there are no gold standards for the purposes of comparison, a consensus grade (C-grade) of goitre was devised for each study subject based on majority opinion by observing certain set of rules as given in Table 1.

Total goitre rate (TGR) was estimated by using the pooled data of all 4 teams as well as by using consensus grade. The 95% confidence intervals for prevalence of goitre was calculated as suggested in the WHO monograph [2].

Statistical methods for evaluation of interobserver agreement

All the four teams were compared separately with the consensus grade as well as with each other. Interobserver agreement of each team with the consensus grade was

evaluated by calculating the kappa coefficient along with Z scores. The proportion of agreement along with 95% confidence intervals of each team with the consensus grade as well as that within each team was calculated. The formulae and the general notations used in the formulae for calculation of the above statistical indices are as given by Grant [4].

Proportion of agreement between the individual teams of observers

As there were four observers the number of possible comparisons (trials) were six. Table 2 shows results of independent assessments by four observers. Thus by four observers there were six trials for each subject and with 100 subjects there were 600 trials as shown in the Table 5.

Results

It was noted that there was 100% consensus amongst the four teams in respect of 11% subjects, 75% consensus in 42% of subjects and 50% consensus in the remaining 47%. There was no instance where all the four teams differed in their observed grade of goitre. The goitre grades as assessed by each team, average of four teams and that of consensus grade is shown in Table 2. The Table also shows prevalence of goitre obtained from pooled data as well as that from consensus grade along with 95% confidence intervals. The kappa co-efficient along with Z score and

TABLE 1
The method of deriving consensus grade from goitre grades given by four teams

Name of the subject	Goitre grades given by team				Consensus* grade arrived	Remarks
	1	2	3	4		
"A"	IA	IA	IA	IA	IA	100% Consensus
"B"	IB	IA	IB	IB	IB	75% Consensus
"C"	IA	IA	0	0	0**	50% Consensus
"D"	IB	IA	IA	0	IA	50% Consensus
"E"	0	IA	IB	II	Reject	No Consensus

Note : Only five of the total 100 subjects are shown here for the sake of explanation

* The grade given by the majority of the four teams was taken as consensus grade.

** In case of 50% consensus wherein one grade was given by two teams and a second grade was given by both the remaining teams, the lower of the 2 grades was chosen as the C-grade.

TABLE 2

Grade of goitre given by different teams, the consensus grades and the prevalence of goitre

Grades of goitre	Number of students with a given grade of goitre				Average of pooled data*	Consensus grade
	Team-1	Team-2	Team-3	Team-4		
0	20	23	52	61	39.00	49
IA	38	65	33	11	36.75	33
IB	40	12	15	28	23.75	18
II	2	—	—	—	0.50	—
Total	100	100	100	100	100.00	100
TGR	80	77	48	39	61.00	51

Note : 1. TGR = Total goitre rate, (IA+IB+II); 2. * Average of Team 1 to Team 4; 3. 95% confidence interval for prevalence of goitre 51% (C-grade) are 63.75 and 38.25 (51×1.25 and 51×0.75).

TABLE 3

Comparative statement showing kappa coefficient, "Z" score and levels of significance of each of four teams with consensus grade

Comparison between	K	Inference on agreement	Z score	Level of significance
C - Grade and Team 1	0.37	Poor	3.39	$P < 0.001$
C - Grade and Team 2	0.39	Poor	3.34	$P < 0.001$
C - Grade and Team 3	0.82	Excellent	4.69	$P < 0.001$
C - Grade and Team 4	0.64	Moderate	3.41	$P < 0.001$

TABLE 4

Comparative statement showing proportion of agreement between C - grade and different teams

Agreement between	Proportion of agreement	
	For abnormality	For normality
C - Grade and Team 1	0.62 (0.51 - 0.72)	0.38 (0.25 - 0.51)
C - Grade and Team 2	0.62 (0.51 - 0.73)	0.41 (0.28 - 0.55)
C - Grade and Team 3	0.83 (0.73 - 0.93)	0.83 (0.73 - 0.93)
C - Grade And Team 4	0.67 (0.54 - 0.79)	0.72 (0.61 - 0.83)

Figures in parentheses () are the lower and upper limits of 95% confidence intervals

p value of agreement of each team with the consensus grade are given in Table 3. The proportion of agreement of each team with consensus grade both for normality and abnormality along with 95% confidence intervals are given in Table 4. The proportion of agreement for all the four observers together is given in Table 5.

Discussion

Interobserver agreement was calculated

using kappa coefficient and proportion of agreement for abnormality and normality along with 95% confidence intervals. The proportion of agreement calculated separately for normality and abnormality was found to be a better statistical method in arriving at interobserver agreement rather than the kappa coefficient.

Prevalence of goitre : TGR as obtained from the pooled data in this study was 61% and that obtained by consensus grade (C-grade)

TABLE 5

Proportion of agreement for more than two observers (interpretation of goitre by four observers teams)

Subject	Assessment Abnormal	Assessment normal	Result of trials*		
			Disagreement	Agreement normal	Agreement abnormal
1	4	0	0	0	6
2	1	3	3	3	0
3	3	1	3	0	3
4	4	0	0	0	6
5	1	3	3	3	0
100	2	2	4	1	1
Total	244	156	220	124	256

1. Proportion of agreement for abnormality

 $256/(256 + 220) = 0.54$ and 95% confidence intervals are 0.50 & 0.59.

2. Proportion of agreement for normality

 $124/(124 + 220) = 0.36$ and 95% confidence intervals are 0.31 & 0.41

3. * Total trials for each subject in case of four observers is six and for 100 subjects it is 600.

4. To avoid lengthy display, only a few subjects are shown.

was 51% (95% CI is 38.25 - 63.75). Hypothetically considering that 51% total goitre rate as the correct prevalence, the total goitre rate as obtained from the pooled data i.e. 61% is still within 95% confidence intervals (38.25 and 63.75).

Assessment of agreement : Kappa coefficient is generally accepted for measuring interobserver agreement [5]. However, its value was doubted as it measures only the association but not the level of agreement between two observers [6]. Kappa statistic also does not distinguish between normality and abnormality (i.e. absence or presence of disease). From Table 3 it was noted that kappa coefficient for Team 1 and Team 2 and C-grade indicated poor agreement whereas that of Team 3 and Team 4 with C-grade indicated excellent to moderate agreement. Z scores are statistically significant. In this kappa statistic does not distinguish between normality and abnormality.

To overcome this drawback the proportion of agreement for abnormality and normality was suggested as an alternate and a better method. Goitre prevalence surveys estimate the presence of abnormality (i.e. goitre) in a community. Thus the proportion of agree-

ment for abnormality between individual teams and consensus grade is more valuable in this study and was found to vary from 0.62 to 0.83. This agreement is considered good as the 95% CI are narrow and does not include 0.50 value. It is thus seen that the indices for measuring the interobserver agreement by kappa and proportion of agreement deal with different aspects of the agreement. The latter was found to be a better method during goitre prevalence studies. Table 5 shows that the proportion of agreement for abnormality within the four teams of observers was 0.54 with 95% CI of 0.50 and 0.59.

The study also highlights the advantages of calculating the goitre prevalence rates by pooling the data of several investigators rather than calculating the same from the observations of a single investigator. From Table 2 it can be inferred that had the prevalence rates been calculated based on the observations of Team 1 only, then we were likely to have exaggerated prevalence rates (80%), on the other hand had we calculated the rates based only on Team 4, we would have underestimated (39%). By pooling the data we could get a more balanced value and probably nearer to the true value. This was

also the view of other authors [7].

In conclusion the kappa statistic was not found to be a good indicator to measure the level of agreement between two observers whereas the proportion of agreement for normality and abnormality along with 95% CI was found to be a better measure.

REFERENCES

1. Thilly CH, Delange F, Stanbury JB. Epidemiological surveys in endemic goitre and cretinism. In : Stanbury JB, Hetzel BS, editors. Endemic goitre and endemic cretinism. New Delhi : Wiley Eastern Ltd., 1985; 157-83.
2. Perez C, Scrimshaw NS, Munoz JA. Technique of endemic goitre surveys. In : Endemic goitre. WHO

Monograph series number 44. Geneva : World Health Organisation 1960; 369-83.

3. Pyke DA. Finger clubbing - validity as a physical sign. Lancet 1954; 2 : 352-4.
4. Grant JM. The fetal heart rate is normal, isn't it? Observer agreement of categoricals. Lancet 1991; 337 : 215-8.
5. Fleiss H. Standard methods for rates and proportions, 2nd ed., New York : John Wiley, 1981; 212.
6. Bailey SM, Sarmandal P, Grant JM. A comparison of three methods of assessing inter-observer variation applied to measurement of symphysis - fundal height. Br J Obstet Gynecol 1989; 96 : 1261-5.
7. Schilling RSF, Huges JPW, Dingwall-Fordyce J. Disagreement between observers in an epidemiological study of respiratory diseases. BMJ 1955; 1 : 65-8.