

Systematic Observer Variation in Trachoma Studies

F. A. ASSAAD¹ & F. MAXWELL-LYONS¹

There has been increasing awareness in recent years among trachomatologists, as well as among workers in other fields of medical research, of the frequency and importance of observer variation in epidemiological studies and clinical trials. In trachoma, the lack of simple definitive laboratory diagnostic procedures suitable for wide application has placed the onus largely, and usually exclusively, on clinical observation.

The study reported is based on the recorded observations of two skilled ophthalmologists in an epidemiological survey covering more than 35 000 persons in Taiwan. Observer differences were found to lie not only in deciding whether a case is trachomatous or not but also in assigning cases diagnosed as trachoma to the appropriate evolutive stage of the WHO trachoma classification.

The conclusions reached are that: (a) inter- and intra-observer variations of some degree are inevitable if dependence is placed on clinical examination alone; (b) it is possible by preliminary testing of observers' interpretation of clinical signs to determine the nature of these differences, to assess their importance, and to reduce them; also to set base-lines for the detection of subsequent divergences over time; and (c) it is better to have two observers than one in any trachoma survey or clinical trial.

INTRODUCTION

The magnitude of the bias introduced by observer variation depends on how much of the variation is "systematic" (consistently in the same direction) and how much is "non-systematic". A systematic bias can be determined if two or more observers examine the same or comparable parts of a sample. Non-systematic bias, being independent from unit to unit, tends to average zero over a large sample and it is properly taken into account in the usual formulae for computing the standard errors of the estimates (Cochran, 1960). Although some cognizance has been taken by trachomatologists of non-systematic intra-observer variations (Bobb, 1966; Attiah et al. 1962), to the present authors' knowledge no studies have been made of systematic variation.

In trachoma surveys and clinical trials decision is needed on the following points:

1. Is a given case trachomatous or not?
2. If diagnosed as trachoma, into which of the four recognized evolutive stages of the disease does it fall?

3. What is the relative gravity of the case, i.e., the degree of disabling or potentially disabling lesions?

Differential diagnosis in trachoma presents difficulties under field conditions—firstly, because of the lack of satisfactory diagnostic instruments (the conventional biomicroscope is too unwieldy for routine use and attempts to produce a portable hand-held binocular model have so far been unsuccessful); secondly, because simple definitive laboratory diagnostic techniques suitable for large-scale application have not yet been developed. The situation is all the more difficult in some trachoma endemic areas where, as a result of improvements in biophysical conditions or of active intervention, the disease is changing to a milder form and cases with minimal characteristic lesions are occurring, sometimes in relatively large numbers. Clinical diagnosis under such conditions is particularly prone to observer variation.

The breakdown of the clinical course of trachoma into four stages, first proposed by MacCallan (1908) and later more clearly defined by the WHO Expert Committee on Trachoma (1952, 1962) serves a very useful purpose, especially in epidemiological studies in showing the pattern of the disease in a community.

¹ Medical Officer, Virus Diseases, Division of Communicable Diseases, World Health Organization, Geneva, Switzerland.

The disease process, however, is a continuum and borderline (inter-stage) cases present problems in clinical judgement. The following directive has provided a partial solution: when in doubt between Stages I and II, place in Stage II; when in doubt between Stages II and III, place in Stage III; when in doubt between Stages III and IV, place in Stage III. However, a tendency to load Stage II at the expense of Stage I may suggest a lower disease incidence than reality, and the holding of possible cures in Stage III may understate spontaneous cure rates or the results of treatment or other active intervention. Moreover, the range of clinical "doubt" is wider in some observers than in others so that, even following the above seemingly simple rules, two observers may not always reach the same conclusions.

Assessment of the relative gravity of trachoma in an individual case, and in a population group at a given time, presents yet another set of difficulties; this will be the subject of a separate study.

OBJECTIVES

The present study is aimed at:

- (1) defining the pattern and magnitude of systematic variations in the observations of two examiners in a specific survey, and
- (2) elucidating the factors which may influence these differences.

METHODS

The study makes use of the data from a trachoma survey carried out in Taiwan during 1960-61. In this survey two highly experienced and fully-briefed ophthalmologists (henceforth referred to as observer X and observer Y) examined 35 607 persons in 6092 households in 352 *lins* (neighbourhoods)—a 1/250 stratified sample of the island's population.

Before embarking on the survey the two examiners, using the same criteria of diagnosis and classification of trachoma and adopting the same examination procedure (using focal illumination and monocular loupe), worked intensively together in the field to determine and iron out points of difference and to attune their diagnosis.

During the survey the households in each sampling unit, the *lin*, were divided between the two examiners by systematic subsampling. Because of the administrative and organizational problems entailed in the conduct of a survey of this magnitude systematic subsampling of the households was decided upon

instead of the more precise (and informative) procedure of interpenetrating subsampling.

Included in the present study are cases which, in examinations under field conditions, presented only one cardinal sign of trachoma (i.e., characteristic follicles, or pannus or scars). An earlier study of the data had indicated that, at least in this particular study in Taiwan, most if not all of these cases were in fact trachomatous (Assaad & Maxwell-Lyons, 1966).

FINDINGS

Presurvey trials

Only one trial that is pertinent to the present study is reported here. The trial entailed the independent examination of a sample of 574 persons of the general population in Ta-an/Ta-chia area (Table 1A). Agreement between the two examiners on whether the case was trachomatous or not reached the high degree of approximately 80%. However (using a technique first described by Yerushalmy et al., 1950) a comparison based on the number of cases assigned by at least one examiner to the evolutive stages (Table 2) shows that in Stages I, II and III in only about one-fifth of the cases diagnosed by the examiners were the two observers in agreement. Moreover, the level of disagreement is almost the same for the three stages. On the other hand, agreement is reached on more than 70% of either total or healed trachoma.

The above comparison does not indicate the possible risk of misdiagnosis of a trachoma-free case as trachoma; in the absence of positive proof assessment of the risk can only be made indirectly. Calculating the proportion of the maximum number of non-trachomatous persons that could have been mislabelled as trachomatous to the maximum number of possibly trachomatous cases (Table 3) reveals that Stages II and IV are the least, and Stage I the most, liable to a misdiagnosis.

If the prevalence rates obtained by the two examiners (Table 1A) are considered, irrespective of agreement or disagreement on individual cases, it will be noted that total trachoma prevalence rates approach each other very closely: 61.1% for X as against 62.2% for Y. However, there is a slight tendency for X to report more active cases (23.7% as against Y's 21.3%).

In the same presurvey exercise the 182 persons for whom an agreement, by stage, was not reached were first subjected to a joint examination and diagnosis, both examiners consulting their earlier readings, and later, within a week, to another independent

TABLE 1

INTER-OBSERVER VARIATION IN PRESURVEY TRIAL IN TA-AN/TA-CHIA AREA

A. INITIAL EXAMINATION OF TOTAL SAMPLE POPULATION

		Examiner Y							
Stage		O	D	I	II	III	IV	Total	
Examiner X	O	157	4	14	1	14	30	220	38.3
	D	3	0	0	0	0	0	3	0.5
	I	21	0	15	1	6	0	43	7.5
	II	3	0	7	5	6	0	21	3.7
	III	20	2	9	1	25	15	72	12.5
	IV	7	0	1	0	17	190	215	37.5
Total		211	6	46	8	68	235	574	100.0
		36.8	1.0	8.0	1.4	11.8	40.9	99.9	
		37.8		21.3			62.2		

B. REPEAT EXAMINATION OF CASES ON WHICH AGREEMENT WAS NOT OBTAINED AT INITIAL EXAMINATION

		Examiner Y							
Stage		O	D	I	II	III	IV	Total	
Examiner X	O	48	1	14	0	8	15	86	55.8
	D	1	0	0	0	0	0	1	0.6
	I	6	0	9	0	9	0	24	15.6
	II	0	0	1	1	2	0	4	2.6
	III	4	0	6	1	9	6	26	16.9
	IV	1	0	0	0	5	7	13	8.4
Total		60	1	30	2	33	28	154	99.9
		39.0	0.6	19.5	1.3	21.4	18.2	100.0	
		39.6		42.2			60.4		

Areas of agreement: Total trachoma Active trachoma Evolutive stage

examination, the initial and joint diagnoses being withheld from the examiners. Altogether 154 persons attended both examinations. The two examiners agreed on the diagnosis by stage (Table 1B) in only 74 cases (48.1%). An equally poor intra-observer agreement was reached: 42.2% for X and 52.6% for Y (Table 4).

Prevalence survey

The systematic subsampling of the households resulted in an equitable division of the sample between the two ophthalmologists: 49.9% to X and 50.1% to Y. The population examined is, moreover, equally distributed by age between the two observers (Table 5).

TABLE 2

INTER-OBSERVER VARIATION, BASED ON PERCENTAGE DISAGREEMENT BETWEEN EXAMINERS, IN PRESURVEY TRIAL IN TA-AN/TA-CHIA AREA

Diagnosis (Tr stage)	Cases diagnosed by one examiner and disagreed upon by the other (a)	Cases agreed upon by both examiners (b)	Cases diagnosed by at least one examiner (a+b) (c)	Percentage disagreement ($\frac{a}{c} \times 100$)
I	59	15	74	79.7
II	19	5	24	79.2
III	90	25	115	78.3
IV	70	190	260	26.9
Subtotal I-III	108	75	183	59.0
Total I-IV	112	298	410	27.3
O (non-trachomatous)	117	157	274	42.7
D	9	0	9	100.0
Subtotal: O+D	112	164	276	40.6

TABLE 3
INTER-OBSERVER VARIATION, IN TERMS OF POSSIBLE MISDIAGNOSIS OF A TRACHOMA-FREE CASE, IN PRESURVEY TRIAL IN TA-AN/TA-CHIA AREA

Stage	Cases diagnosed by at least one examiner	Cases diagnosed Tr O or Tr D by at least one examiner	
		No.	%
I	74	35	47.3
II	24	4	16.7
III	115	34	29.6
IV	260	37	14.2
Subtotal: I-III	183	75	41.0
Total: I-IV	410	112	27.3

The results of the clinical examination indicate very little observer difference in the prevalence rates of total trachoma: 46.9% for examiner X and 49.1% for examiner Y. However, in X's subsample there is a lower rate of healed trachoma (24.6% as against Y's 30.4%) and a correspondingly higher rate of active trachoma (22.2% as against Y's 18.7%); the greatest relative difference is encountered in Stage I (Table 5).

Considering age-specific rates (Table 5) the following is noted.

Total trachoma. Both subsamples start with the same rate, in the age below 1 year, soon diverge, with the most marked relative difference (X's rate more than double that of Y) in the age-group 3-4 years, but converge again beyond the age of 9 years, X's rates being consistently lower than Y's (Fig. 1).

Active trachoma. Age-prevalence line graphs for active trachoma (Fig. 1) show rates of similar or near similar magnitude in the very young (below 1 year), in the very old (beyond 64 years) and in the ages of maximum prevalence (10-24 years). With the exception of the age-group 10-14 years, Y's rates are consistently lower than X's. A breakdown of active trachoma by presence or absence of cicatrices (Fig. 2) indicates that the difference in the total active trachoma rates in the younger age-groups, below 15 years, is a reflection of an excess of X's reported non-cicatricial trachoma overriding Y's higher rates of Tr III. Beyond 14 years of age, X gives higher rates of both non-cicatricial and cicatricial trachoma. Meantime, the excess in non-cicatricial trachoma in X's subsample is due to higher rates of Tr I than those encountered in Y's subsample; Tr II rates are nearly the same for the two examiners (Fig. 3 and 4).

Healed trachoma. X's rates are consistently lower (Fig. 1) than Y's.

To determine the effect of the "time" element (the survey lasted approximately 8 months) on the

TABLE 4
INTRA-OBSERVER VARIATION ON REPEAT EXAMINATION OF CASES ON WHICH AGREEMENT WAS NOT OBTAINED AT INITIAL EXAMINATION IN PRESURVEY TRIAL IN TA-AN/TA-CHIA AREA

		A. EXAMINER X								B. EXAMINER Y							
		Repeat Examination								Repeat Examination							
Initial Examination	Stage	O	D	I	II	III	IV	Total	}	}	}	}	}	}	}	}	
		O	33	0	7	0	6	3									49
	D	1	1	0	0	1	0	3	1.9	33.8							
	I	16	0	7	1	0	0	24	15.6								
	II	4	0	7	3	0	0	14	9.1	51.9							
	III	22	0	3	0	14	3	42	27.3								
	IV	10	0	0	0	5	7	22	14.3								
	Total	86	1	24	4	26	13	154	100.0								
		55.8	0.6	15.6	2.6	16.9	8.4	99.9									
		56.5		35.1				43.5									
		39.0		0.6		19.5		1.3		21.4		18.2		100.0			
		39.6		42.2				60.4									
	Total	60	1	30	2	33	28	154	99.9								

Areas of agreement:

 Total trachoma

Active trachoma

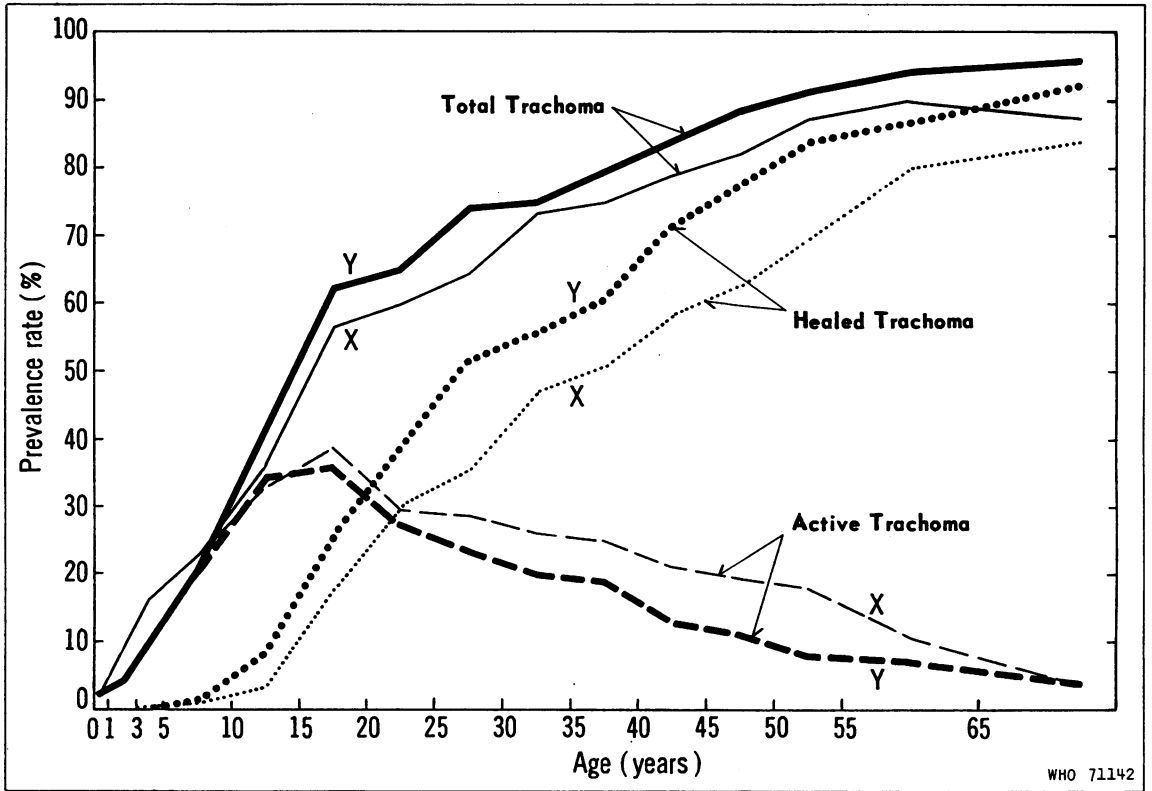
Evolutive stage

TABLE 5
CLINICAL FINDINGS BY AGE AND EXAMINER IN PREVALENCE SURVEY

Age (years)	Ex-aminer	Number ex-aminated	Trachoma cases											
			I		II		III		Subtotal I-III		IV		Total I-IV	
			No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
< 1	X	631	13	2.1					13	2.1			13	2.1
	Y	634	12	1.9				12	1.9			12	1.9	
1-2	X	1 457	106	7.3	11	0.8	2	0.1	119	8.2			119	8.2
	Y	1 473	46	3.1	11	0.7	1	0.1	58	3.9			58	3.9
3-4	X	1 365	171	12.5	38	2.8	7	0.5	216	15.8	3	0.2	219	16.0
	Y	1 368	85	6.2	39	2.8	2	0.1	126	9.2			126	9.2
5-9 ^a	X	3 288	487	14.8	159	4.8	80	2.4	726	22.1	8	0.2	734	22.3
	Y	3 320	317	9.5	195	5.9	159	4.8	671	20.2	34	1.0	705	21.2
10-14	X	2 241	325	14.5	133	5.9	275	12.3	733	32.7	72	3.2	805	35.9
	Y	2 228	150	6.7	168	7.5	431	19.3	749	33.6	180	8.1	929	41.7
15-19	X	1 370	64	4.7	42	3.1	420	30.6	526	38.4	244	17.8	770	56.2
	Y	1 430	34	2.4	55	3.8	423	29.6	512	35.8	370	25.9	882	61.7
20-24	X	1 148	15	1.3	10	0.9	311	27.1	336	29.3	344	30.0	680	59.2
	Y	1 164	10	0.8	15	1.3	286	24.6	311	26.7	444	38.1	755	64.9
25-29	X	1 190	10	0.8	3	0.2	330	27.7	343	28.8	417	35.0	760	63.9
	Y	1 212	7	0.6	5	0.4	268	22.1	280	23.1	617	50.9	897	74.0
30-34	X	1 134	6	0.5	5	0.4	283	25.0	294	25.9	530	46.7	824	72.7
	Y	1 154	2	0.2	2	0.2	221	19.2	225	19.5	635	55.0	860	74.5
35-39	X	938	1	0.1	4	0.4	227	24.2	232	24.7	470	50.1	702	74.8
	Y	960	1	0.1	1	0.1	177	18.4	179	18.6	577	60.1	756	78.8
40-44	X	847	1	0.1			174	20.5	175	20.7	492	58.1	667	78.7
	Y	810					102	12.6	102	12.6	575	71.0	677	83.6
45-49	X	685	1	0.1	2	0.3	128	18.7	131	19.1	427	62.3	558	81.4
	Y	723					79	10.9	79	10.9	558	77.2	637	88.1
50-54	X	634					112	17.7	112	17.7	439	69.2	551	86.9
	Y	652					49	7.5	49	7.5	543	83.3	592	90.8
55-64	X	775					79	10.2	79	10.2	616	79.5	695	89.7
	Y	735					51	6.9	51	6.9	638	86.8	689	93.7
≥ 65	X	506					17	3.4	17	3.4	424	83.8	441	87.2
	Y	434					15	3.4	15	3.4	399	91.9	414	95.4
Total	X	18 209	1 200	6.6	407	2.2	2 445	13.4	4 052	22.2	4 486	24.6	8 538	46.9
	Y	18 297	664	3.6	491	2.7	2 264	12.4	3 419	18.7	5 570	30.4	8 989	49.1
	Total	36 506	1 864	5.1	898	2.5	4 709	12.9	7 471	20.5	10 056	27.5	17 527	48.0

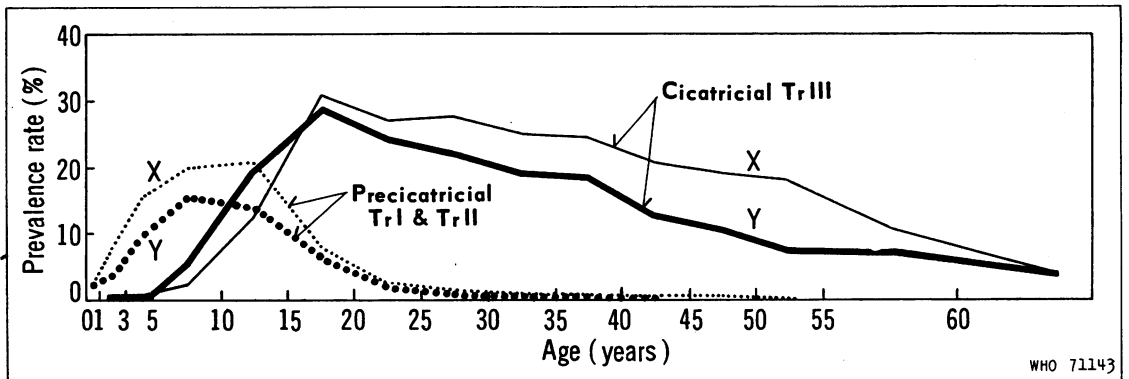
^a One individual (Tr O) in whose case the examiner could not be identified has been excluded from the study.

FIG. 1
TRACHOMA PREVALENCE, BY AGE, AS RECORDED BY EXAMINERS X AND Y



WHO 71142

FIG. 2
PRECICATRICAL AND CICATRICAL TRACHOMA PREVALENCE, BY AGE, AS RECORDED BY EXAMINERS X AND Y



WHO 71143

FIG. 3
Tr I PREVALENCE, BY AGE, AS RECORDED BY EXAMINERS X AND Y

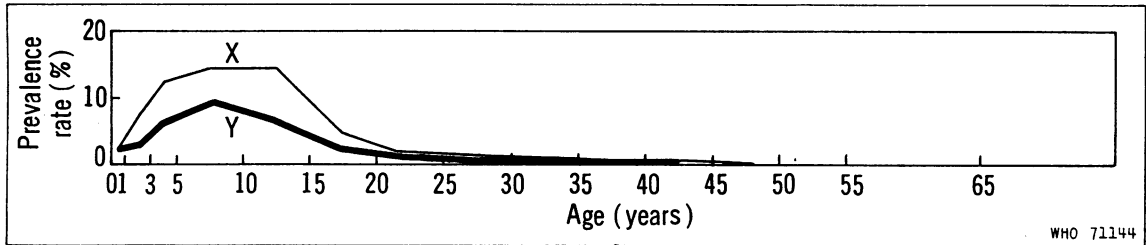
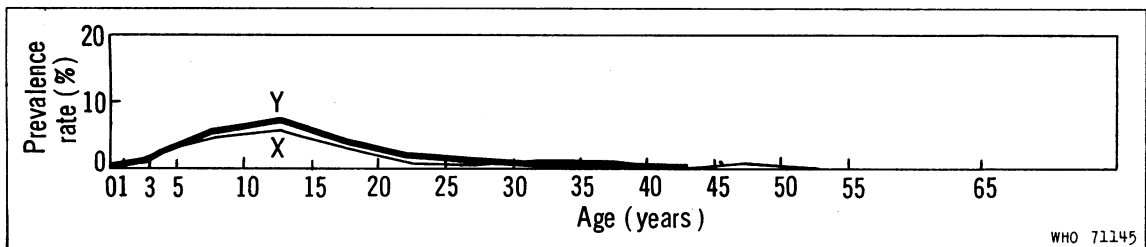


FIG. 4
Tr II PREVALENCE, BY AGE, AS RECORDED BY EXAMINERS X AND Y



examiner's diagnosis, and hence on a systematic observer variation, prevalence rates are calculated for each of the administrative divisions on the island arranged in the order in which they were taken in the survey. The sequence of field operations was determined by random ordering of the administrative divisions; within any one division the examinations were scheduled in the order most convenient to field work. The arrangement by administrative division does not represent equal intervals on a time scale, but a mere time sequence (Fig. 5A).¹ Except for the 5th to 8th and 21st and 22nd (last two) administrative divisions on the time scale, Y maintains a higher rate of total trachoma than X; the difference between the examiners showing an irregular pattern. On the other hand, a definite change is seen in the magnitude and direction of difference in both active and healed lesions. In the case of active trachoma, after an initial increase the difference between the examiners not only diminishes as the survey progresses but even changes direction;

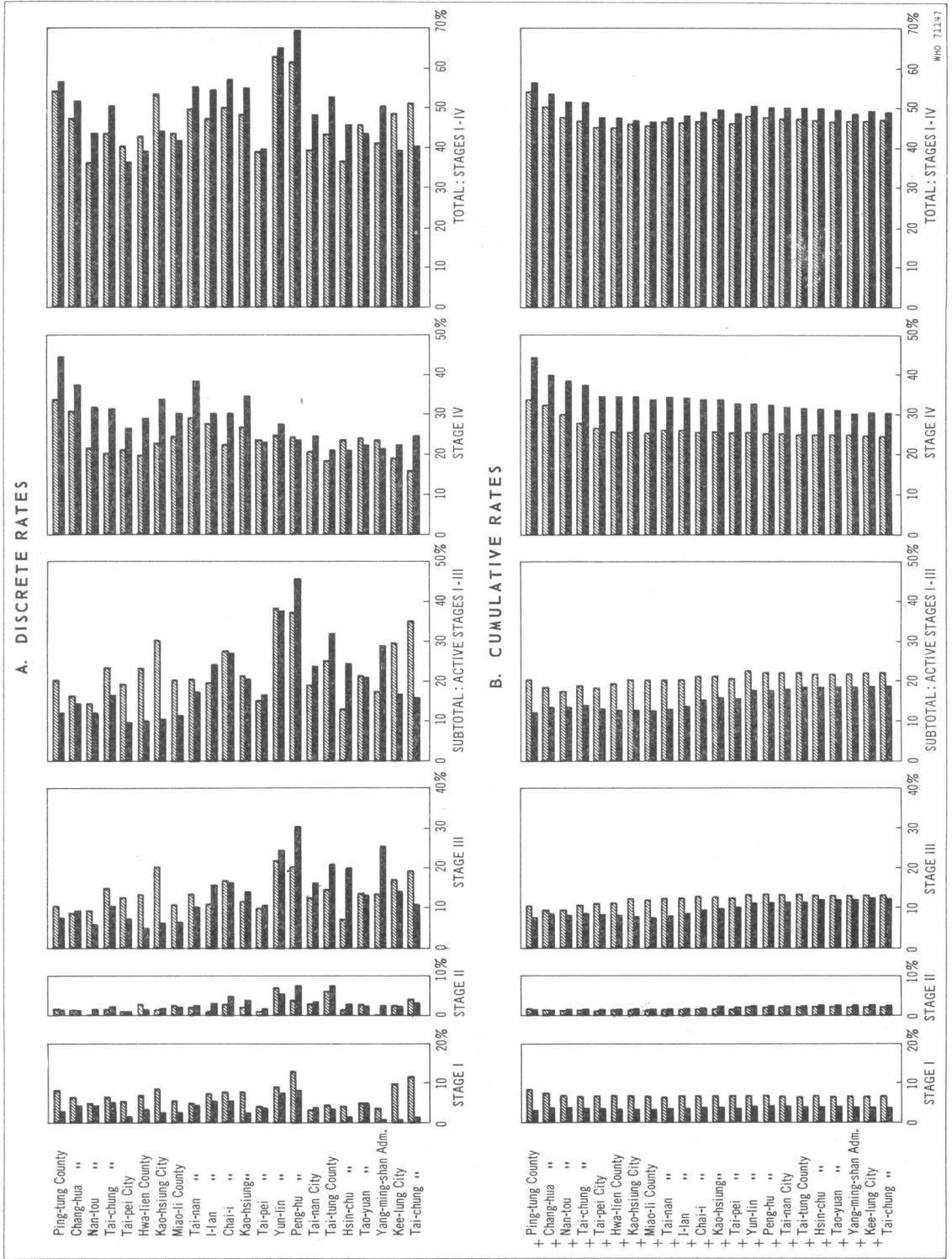
while up to Yun-lin County the trend is for X's prevalence rates to exceed Y's, X tends to diagnose less active trachoma than Y from Peng-hu County on to near the end of the survey. An abrupt reversal in this latter trend and a very marked difference between examiners sets in, however, in the last two divisions. The changes noted in observer variation in active trachoma are a reflection of the examiners' trends in diagnosing Stage III—Stage I contributing little (except in case of the last two divisions where X's rate is nearly 10 times that of Y), and Stage II hardly at all, to the total picture.

Healed trachoma shows a trend in the opposite direction to that noted in active lesions but the pattern is less regular. Again the trend is abruptly reversed in the last two divisions.

To disentangle a consistent systematic pattern from non-systematic observer and sampling variations, cumulative prevalence rates, as the survey progresses in time, are calculated and plotted (Fig. 5B). In using the cumulative rates the denominator becomes bigger as more divisions are added and the non-systematic variations are thereby

¹ The table on which Fig. 5 is based has been deposited in the WHO Library. Single copies may be obtained on request.

FIG. 5
 TRACHOMA PREVALENCE RATES, BY ADMINISTRATIVE DIVISION IN ORDER OF FIELD OPERATIONS AND BY EXAMINER



absorbed; only consistent trends in the same directions are demonstrated.

The difference between the two examiners in the total trachoma rates, although varying in magnitude, is sustained throughout. On the other hand, both examiners show a definite pattern of a higher active trachoma rate the further the survey progresses. However, Y's trend is far more marked than that of X, resulting in appreciable convergence of the two examiners' rates as the survey advances. This convergence is seen in all stages but in particular in Stage III. Healed trachoma shows the same trend of the prevalence rates given by the two examiners coming closer with the advance in time but to a lesser degree than in active trachoma.

To determine the effect of the level of endemicity on observer variation the weighted average of the

two examiners' active trachoma rates is taken as the index of relative endemicity. The trachoma prevalence rates (P) for the combined subsamples are calculated in terms of:

$$\frac{\text{No. of cases diagnosed by } X + Y}{\text{No. of persons examined by } X + Y} \times 100$$

referred to hereafter as P_{X+Y} ; for the individual examiners the rates are referred to as P_X and P_Y , respectively. Ranking the administrative divisions in descending order of active trachoma rate shows marked discrepancies between the ranking order given to the combined P_{X+Y} and to the individual examiner's P_X and P_Y readings (Table 6). Observer variation calculated in terms of $\frac{P_X - P_Y}{P_{X+Y}}$ shows a marked inverse correlation with the degree of

TABLE 6. RELATIVE ENDEMICITY OF ADMINISTRATIVE DIVISIONS BY EXAMINER

Time sequence	Administrative division	Active trachoma prevalence rates (%)						Difference (%)	
		Pooled data		Examiner X		Examiner Y		$ P_X - P_Y $	
		Rate P_{X+Y}	Rank	Rate P_X	Rank	Rate P_Y	Rank	$P_X > P_Y$	$P_X < P_Y$
3	Nan-tou County	13.1	1	14.4	2	11.8	5	19.3	
5	Tai-pei City	14.3	2	19.2	7	9.7	1	66.6	
8	Miao-li County	15.1	3	20.1	9	11.4	4	49.3	
2	Chang-hua County	15.4	4	16.4	4	14.5	7	12.8	
13	Tai-pei County	15.7	5	15.2	3	16.3	9		7.0
1	Ping-tung County	16.4	6	20.3	10	12.0	6	49.6	
6	Hwa-lien County	16.8	7	23.3	15	9.9	2	80.2	
9	Tai-nan County	18.9	8	20.5	11	17.1	12	18.1	
18	Hsin-chu County	19.1	9	12.9	1	24.5	17		60.6
4	Tai-chung County	19.8	10	23.2	14	16.7	10	32.9	
7	Kao-hsiung City	20.1	11	30.4	19	10.3	3	100.1	
12	Kao-hsiung County	21.0	12	21.4	12	20.5	13	4.7	
16	Tai-nan City	21.3	13	18.8	6	23.6	15		22.1
19	Tao-yuan County	21.3	14	21.5	13	21.1	14	1.8	
10	I-lan County	21.9	15	19.6	8	24.2	16		21.2
20	Yang-ming-shan Adm.	22.3	16	17.3	5	28.9	19		52.3
21	Kee-lung City	22.8	17	29.6	18	16.8	11	55.9	
22	Tai-chung City	24.6	18	35.0	20	15.9	8	77.6	
11	Chai-i County	27.3	19	27.6	17	27.0	18	2.6	
17	Tai-tung County	28.3	20	25.0	16	31.6	20		23.3
14	Yun-lin County	37.9	21	38.1	22	37.5	21	1.6	
15	Peng-hu County	41.3	22	37.2	21	45.9	22		21.0

agreement between examiners in assigning an administrative division to a position on the endemicity scale, but, conversely, bears no consistent relationship to endemicity. However, the effect endemicity may have on observer variation could have been overshadowed by the influence of time.

The county administrative divisions are formed of urban and rural communities. As mentioned above, the random ordering of field operations was limited to the administrative divisions; within any one county the *lins*, whether belonging to an urban or a rural community, were taken in the survey in the order dictated by local convenience. The time element, therefore, ceases to be a factor in determining the magnitude and direction of observer variation in the urban and rural communities within any one administrative division. Calculating the active trachoma prevalence rates for the urban and rural communities shows that the examiners agree in 14 divisions out of the total of 16 on which community has a higher rate (Table 7).

Designating the communities in every county as of relatively low or relatively high endemicity¹ and calculating the cumulative trachoma prevalence rates for each group of communities separately (Fig. 6)² shows clearly the interplay between endemicity and time as factors in determining the magnitude and direction of observer variation. Excluding Ping-tung County, a bigger observer variation in active trachoma prevalence rates is encountered in communities of relatively low endemicity. In either group of communities systematic observer variation in the active rates decreases with time but by the end of the survey is still more marked in the group of relatively low endemicity (Group RLE) than in the group with relatively high endemicity (Group RHE):

Trachoma stage	Magnitude of observer variation in terms of $\frac{P_X - P_Y}{P_X + P_Y} \times 100$ in communities of:	
	RLE	RHE
I	72.3	32.3
II	41.7	11.3
III	5.1	2.2
Subtotal I - III	16.4	5.5
IV	15.9	23.8
Total I - IV	3.9	10.6

¹ In both Ping-tung and Hsin-chu Counties, failing an agreement between the examiners on which community has the lower and which has the higher relative endemicity, the active trachoma prevalence rate calculated on the basis of pooled findings ($P_X + Y$) is taken as the index of relative endemicity.

² The table on which Fig. 6 is based has been deposited in the WHO Library. Single copies may be obtained on request.

In the RLE group of communities the active trachoma picture is a reflection of changes over time in the diagnosis of Stage III; hardly a change is noticed in the reporting of Stages I and II. On the other hand, in the RHE group of communities Stage I shares in the change over time; a marked reduction in the systematic observer variation is noted. Stage III shows not only a reduction over time in an initially relatively small observer variation but also, by the end of the survey, a reversion in the direction of the observer variation. Stage II shows little change but, compared with the RLE group, the observer variation is nevertheless appreciably lower. The diagnosis of healed trachoma does not follow the same pattern. A definite systematic variation is seen in both RLE and RHE groups of communities. Moreover, although a more marked reduction is seen in the RHE communities over time, yet observer variation is appreciably higher all the time in this same group.

The total trachoma picture (the summation of the active and healed rates, the latter predominating in the great majority of communities irrespective of examiner) shows a mild systematic variation late in the survey in the RLE group of communities. Conversely a sustained and rather marked systematic variation is noted throughout in the case of the RHE communities.

DISCUSSION

Presurvey trials

Notwithstanding the fact that the trial had the limited objective of revealing the areas of difference between the two examiners and was not at the time conceived as a study of inter-observer variation, it nevertheless demonstrated:

(1) that when the simple question is asked—is a case trachomatous or not?—a high degree of agreement is reached, comparing favourably in fact with observations in other fields of clinical medicine (Witts, 1964³); and

(2) that there is good agreement as to whether a case of trachoma is active or healed; further breakdown, however, reduces progressively the area of agreement.

Considering that the relative distribution of stages is largely a function of age, the differences

³ The book cited contains several chapters dealing with observer variation and useful references to earlier studies in this field.

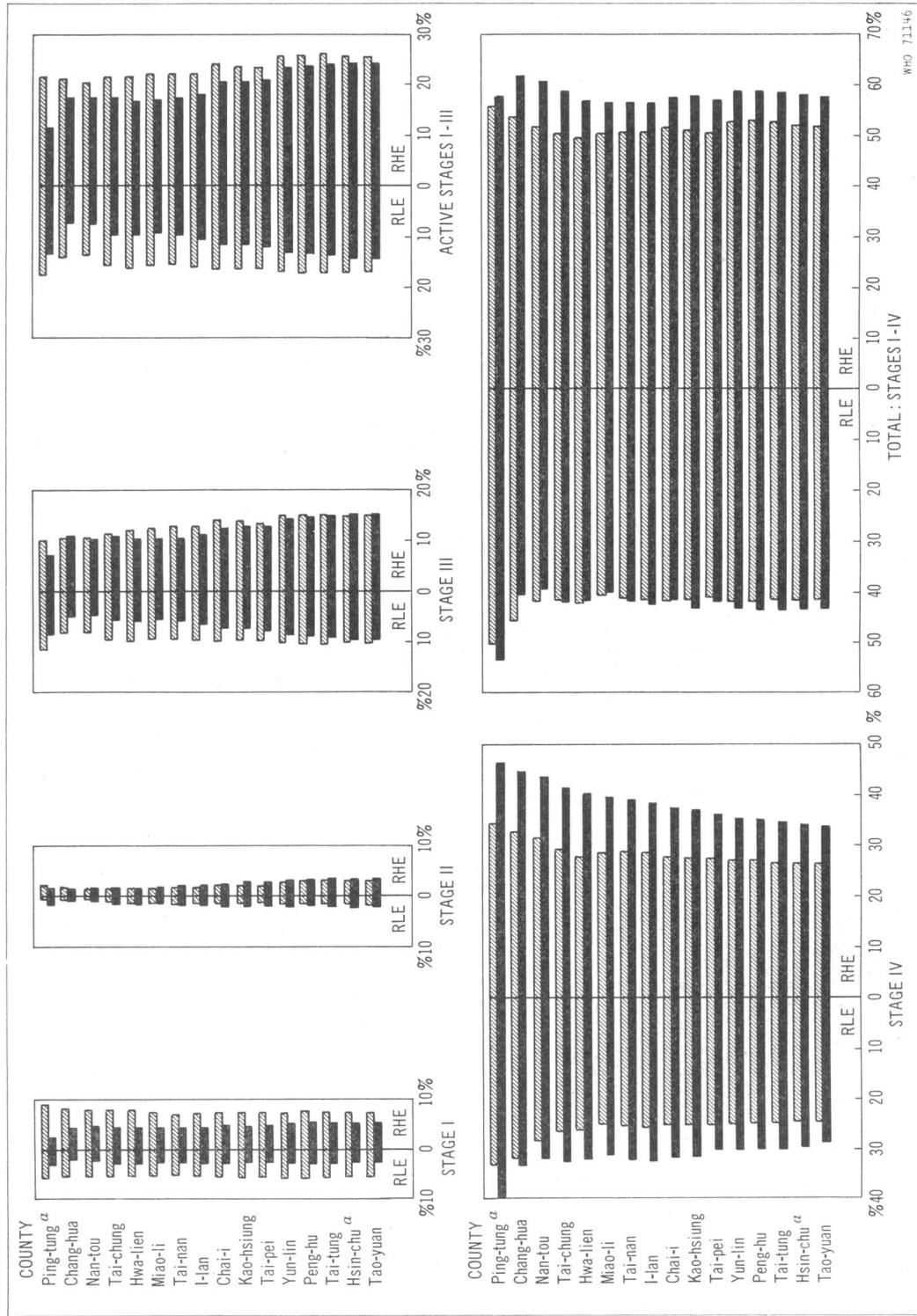
TABLE 7
RELATIVE ENDEMICITY OF URBAN AND RURAL COMMUNITIES BY EXAMINER

County administrative division	Community	Number examined			Active trachoma: stages I-III					
		X	Y	X+Y	Number			Prevalence rates (%)		
					X	Y	X+Y	X	Y	X+Y
Ping-tung	Urban	376	345	721	66	46	112	18.0	13.3	15.5
	Rural	885	765	1 650	190	87	277	21.5	11.4	16.8
Chang-hua	Urban	647	714	1 361	77	33	110	11.9	4.6	8.1
	Rural	758	972	1 730	154	211	365	20.3	21.7	21.1
Nan-tou	Urban	432	413	845	56	35	91	13.0	8.5	10.8
	Rural	291	279	570	48	47	95	16.5	16.8	16.7
Tai-chung	Urban	442	515	957	114	93	207	25.8	18.1	21.6
	Rural	480	521	1 001	100	80	180	20.8	15.4	18.0
Hwa-lien	Urban	223	210	433	51	20	71	22.9	9.5	16.4
	Rural	257	246	503	61	25	86	23.7	10.2	17.1
Miao-li	Urban	414	393	807	105	69	174	25.4	17.6	21.6
	Rural	366	413	779 ^a	42	23	65	11.4	5.6	8.3
Tai-nan	Urban	435	379	814	73	51	124	16.8	13.5	15.2
	Rural	938	906	1 844	209	169	378	22.3	18.7	20.5
I-lan	Urban	310	265	575	59	55	114	19.0	20.8	19.8
	Rural	327	337	664	66	91	157	20.2	27.0	23.6
Chai-i	Urban	538	560	1 098	102	98	200	19.0	17.5	18.2
	Rural	807	731	1 538	270	250	520	33.5	34.2	33.8
Kao-hsiung	Urban	254	283	537	44	41	85	17.3	14.5	15.8
	Rural	870	841	1 711	197	189	386	22.6	22.5	22.6
Tai-pei	Urban	904	788	1 692	132	106	238	14.6	13.5	14.1
	Rural	434	459	893	71	97	168	16.4	21.1	18.8
Yun-lin	Urban	467	517	984	115	120	235	24.6	23.2	23.9
	Rural	815	663	1 478	374	323	697	45.9	48.7	47.2
Peng-hu	Urban	75	66	141	29	37	66	38.7	56.0	46.8
	Rural	78	69	147	28	25	53	35.9	36.2	36.1
Tai-tung	Urban	131	130	261	22	38	60	16.8	29.2	23.0
	Rural	237	234	471	70	77	147	29.5	32.9	31.2
Hsin-chu	Urban	297	322	619	31	90	121	10.4	28.0	19.5
	Rural	323	396	719	49	86	135	15.2	21.7	18.8
Tao-yuan	Urban	351	293	644	59	52	111	16.8	17.7	17.2
	Rural	431	404	835	109	95	204	25.3	23.5	24.4

^a One individual in whose case the examiner could not be identified was withdrawn from the study.

FIG. 6

CUMULATIVE TRACHOMA PREVALENCE RATES, BY EXAMINER, IN COMMUNITIES WITH RELATIVELY LOW ENDEMICITY (RLE) AND RELATIVELY HIGH ENDEMICITY (RHE) WITHIN EACH COUNTY



Examiner X.

Examiner Y.

^a In both Ping-tung and Hsin-chu Counties the communities are assigned to RLE or RHE according to P_{X-I, Y}; the examiners did not agree as to which community had the higher or lower active trachoma rate.

in the diagnosis of individual stages would show up in the age-specific rates, particularly in the younger ages when a difference in the diagnosis of one stage is not compensated for by a corresponding difference in the opposite direction in the diagnosis of another stage. Because of the small number examined in this presurvey trial no breakdown by age is attempted; the effect of observer variation in age-specific rates as distinct from over-all rates is dealt with in detail in the discussion on the prevalence survey. The marked intra- and inter-observer variation in the repeat examination of "contested" cases in which agreement could not be reached in the first instance suggests that, in the field, trachoma presents a core of cases in which an agreement on the diagnosis could be reached and a crust of varying depth where the subjectivity of the observer dominates and in which reproducibility of the diagnosis is hard to obtain. The possibility of the joint examination playing a role, through over-compensation, i.e., over-diagnosis in the opposite direction, cannot be excluded.

Prevalence survey

The systematic observer variation in this survey is apparently due to the mild forms of trachoma that predominate in Taiwan and to differences in the detection and interpretation of follicles, scars and pannus when these lesions are minimal.

Pooling the two observers' findings it is found that:

- (1) in 86.1% of Tr I and 92.0% of Tr III follicular involvement was minimal;¹
- (2) 86.9% of Tr III cases and 80% of Tr IV cases had minimal cicatrization;
- (3) in 90.9% of all reported cases of pannus extension was less than 1 mm from the limbus.

In Taiwan, trachoma follicles are sometimes seen towards the angles of the upper tarsal conjunctiva without the characteristic involvement of the central area. When these atypically placed follicles are few in number and unassociated with other confirmatory signs their interpretation is conjectural. Observer differences in detecting minimal scarring is partially explained by the well-known difficulty in deciding whether a minute whitish patch represents residual infiltration or an early cicatrix. The identification of early pannus is notoriously difficult without the aid of a biomicroscope and, although both examiners in

the present study had exceptionally keen vision supplemented by $\times 10$ monocular loupe, their reporting of 0.5-mm extension of pannus must in many cases have been based on impressions rather than certainty of observation.

We may now examine some of the specific observer differences.

Follicles. In the younger ages X's excess in reporting Tr I is not compensated for by an equal excess in Y's reported Tr II or Tr III. In 98.0% of cases diagnosed as Tr III by both examiners (pooled data) follicles are the determining factor in classifying them as active trachoma; in only 96 cases (2.0%) is Tr III diagnosed on evidence of scars and active pannus in absence of conjunctival follicles. The consistent difference between the two examiners over the range 15-64 years is again in the same direction as in Tr I in the younger ages. It would seem, therefore, that X consistently reports more cases with (presumably atypical) follicles than does Y.

Scars. With regard to scars it seems that the picture is reversed: Y tends to detect the presence of minimal scarring more often than X. In the younger ages Y exceeds X in the diagnosis of Tr III and Tr IV. In the case of Tr III, Y's higher rates can be attributed, at least in part, to an excess over X in noting scars—without which these cases would have been diagnosed as Tr I or Tr II. Y's higher rates of Tr IV clearly show his tendency to record less follicles and more scars than X. In the older age-groups the excess of Y's Tr IV is not compensated for by a higher Tr III prevalence rate from X, i.e., there is an excessive number of cases with scars detected by Y beyond what could be accounted for by a possible misclassification of cases (Tr III misclassified as Tr IV, or *vice versa*, depending on who is compared with whom).

Pannus. Compared with Y, X records pannus in a higher proportion of cases irrespective of stage; the difference being most marked in stages I and II (Table 8). One may therefore assume that, at least in a proportion of cases, X when in doubt favours a diagnosis of minimal pannus while under similar circumstances Y does not.

X reported 208 and Y 221 cases which presented pannus as the only sign. However, of X's cases, 64 were diagnosed as active pannus while of Y's only 25 were regarded as active. It is interesting to note that the examiners' general trend is here maintained, X tending to diagnose more active and Y more healed lesions.

¹ In contrast to 40.3% of Tr II cases—where observer variation is least noted.

TABLE 8
FREQUENCY OF REPORTED PANNUS, BY EXAMINER,
IN PREVALENCE SURVEY

Stage	Total number of cases		Cases in which pannus is recorded			
	X	Y	X		Y	
			No.	%	No.	%
I	1 200	644	918	76.5	306	46.1
II	407	491	343	84.3	293	59.7
III	2 445	2 264	2 260	92.4	2 005	88.6
IV	4 486	5 570	4 111	91.6	4 898	87.9
Total	8 538	8 989	7 632	89.4	7 502	83.5

Influence of time

The change over time in an observer's criteria of diagnosis and classification of trachoma affects particularly stage III, which constitutes the bulk of active disease. Noting the tendency of Y to diagnose more scars than X one cannot but assume that Y is detecting more follicles as the survey progresses. The abrupt change in the pattern of observer variation in the last two divisions is hard to account for. It may be that working for a long period in the field and at last approaching the end of the trail—the two last divisions were small—one or both examiners gave way to fatigue and boredom and relaxed in his examination procedures and/or criteria of diagnosis and classification of trachoma. It is of interest to note that this abrupt change brings the observer variation to the same direction, as regards trachoma stages, as at the starting-point. However, the excess in X's active trachoma stages overcompensates for Y's higher rates of healed trachoma resulting in X's higher over-all trachoma rates.

Influence of endemicity

Concerning the relation of endemicity to observer variations two aspects are noted:

1. Observer variation affects the weighted average of the two examiners' findings and hence plays a role in determining the apparent relative endemicity of trachoma in a given community.

2. In reverse, the actual level of endemicity has an effect on the magnitude of systematic observer variation, both in the diagnosis of trachoma and in its differentiation into active and healed.

In Taiwan the higher the trachoma prevalence, the lower the proportion of mild equivocal cases

(Assaad & Maxwell-Lyons, 1967) and hence the smaller the observer variation. This being so it is presumably also true that control programmes resulting in diminished incidence, prevalence and gravity of trachoma could, over time, lead to increased inter- and intra-observer variation and to diminished reliability in assessing the situation.

CONCLUSIONS

Systematic observer variation of some degree is inevitable in trachoma surveys—especially when, in the absence of suitable laboratory diagnostic procedures, decisions must depend solely on clinical judgement under field conditions and without recourse to a biomicroscope. Variation takes the form of differences: (a) in diagnosis, as to whether a case in trachomatous or not, and (b) in assigning a case diagnosed as trachoma to one of the evolutive stages of the WHO classification.

While the over-all trachoma prevalence rates given by different examiners may closely approach each other, a breakdown by stage and age may reveal wide differences in clinical assessment. Stage I proved to be the most liable to misinterpretation.

If an examiner's diagnostic bias, whatever its nature or degree, remains constant throughout a survey the estimates of *difference* between the trachoma rates in different communities remain unbiased. The present study, however, shows a definite change in the examiner's diagnosis within the 8 months' duration of the survey; the estimated differences between communities are thus no longer unbiased.

Longitudinal or follow-up studies, which usually extend over much longer periods, may likewise be subject to variation in observer bias—thereby resulting in a biased estimate of change in trachoma rates over time in the *same* community. The present study shows, moreover, that, so far as early trachoma is concerned, systematic observer variation is inversely related to the level of endemicity. Thus a true reduction over time in trachoma rates resulting from active intervention or improvement in environment may be expected to increase an observer's bias—perhaps in the direction of overdiagnosis and understatement of the actual change in endemicity.

Systematic and non-systematic observer variation, and changes therein induced over time, cannot be elicited if only one clinician undertakes all examinations. It would be advisable, therefore, in both prevalence and follow-up studies, to have more than

one examiner in order: (a) to be able to determine the bias introduced by systematic observer variation, and thus (b) to obtain more reliable estimates of trachoma prevalence and distribution.

The conclusions reached in this critical study of data from a prevalence survey would, of course,

apply equally to the evaluation of controlled clinical trials and mass treatment operations in trachoma. Here again, the provision of two examiners may indicate important areas of unreliability in diagnosis and thereby point to better indices for evaluation.

ACKNOWLEDGEMENTS

The authors acknowledge with gratitude and admiration the immense amount of work done by the ophthalmologists, Dr I. H. Chang and Dr S. Maffei, in collecting the data on which the present study is based.

RÉSUMÉ

Il est admis que l'écart personnel d'observation est inévitable en recherche clinique. Il peut être systématique (toujours de même sens) ou non. Pour essayer de déterminer sa nature et son ampleur possible dans les enquêtes sur la prévalence du trachome, on s'est servi des données provenant d'une étude faite à Taïwan en 1960-1961. Deux ophtalmologistes particulièrement expérimentés et dûment avertis ont examiné en tout 35 607 sujets composant un échantillon au 1/250, convenablement stratifié, de la population globale de l'île. L'échantillon a été également réparti entre ces deux observateurs par sous-échantillonnage des ménages.

Les deux observateurs ont signalé des taux très voisins de prévalence de tous les cas de trachome (évolutif et cicatrisé), mais l'un d'eux a indiqué un taux plus élevé de trachome évolutif, alors que son homologue, en compensation, a trouvé un taux de trachome cicatrisé supérieur. La différence maximale entre les deux observateurs a porté sur le diagnostic de Tr I. En général l'un des observateurs a signalé plus de cas avec follicules et plus de cas à pannus que son confrère; celui-ci en revanche a rapporté un nombre supérieur de cas à cicatrices.

Étant donné que la distribution relative des stades du trachome dépend beaucoup de l'âge, l'analyse des taux par groupes d'âges met en évidence les différences de diagnostic pour chaque stade.

Le diagnostic différentiel est rendu plus difficile, à Taïwan du moins, par l'apparition en nombre relativement important de cas présentant des signes minimaux.

L'écart personnel systématique d'observation dans l'enquête étudiée présente une particularité remarquable: il diminue, et peut changer de sens, en fonction de la progression de l'enquête dans le temps. Cette évolution est également influencée par le niveau local d'endémicité; en général plus la prévalence du trachome évolutif est élevée,

plus les cas sont graves et la proportion de cas bénins douteux faible, si bien que l'écart personnel d'observation diminue et décline ensuite encore au cours du temps. Pour l'appréciation du trachome cicatrisé parmi les mêmes groupes de population, cet écart a atteint au début des valeurs notablement plus fortes, mais diminue néanmoins en fonction du temps.

Cette étude permet de conclure que dans les enquêtes sur le trachome où, jusqu'à présent, les épreuves de diagnostic en laboratoire ne sont pas d'un grand secours et où l'appréciation de l'allure de l'épidémie est uniquement basée sur les manifestations cliniques de la maladie, il faut s'efforcer de déterminer le biais introduit par l'écart personnel systématique d'observation. Pour ce faire, il est indispensable que l'enquête soit confiée à deux observateurs au moins et qu'ils étudient de préférence des sous-échantillons superposés.

Les études longitudinales ou suivies peuvent de même être faussées par l'évolution du diagnostic dans le temps. L'appréciation des variations de la prévalence du trachome dues à une intervention active peut ainsi se trouver biaisée, très probablement dans le sens d'une sous-estimation.

Les conclusions de cette étude critique des données recueillies au cours d'une enquête sur la prévalence s'appliquent également à l'évaluation des essais cliniques contrôlés, des opérations de traitement de masse et d'autres types d'intervention active. Pour toutes les études sur le terrain concernant le trachome, il est conseillé de faire appel à deux cliniciens au moins afin a) d'être en mesure de déterminer l'erreur introduite par l'écart personnel systématique d'observation, et ainsi b) d'obtenir des estimations plus fidèles de la prévalence, de la distribution et de leurs variations dans le temps.

REFERENCES

- Assaad, F. A. & Maxwell-Lyons, F. (1966) *Bull. Wld Hlth Org.*, **34**, 341-355
- Assaad, F. A. & Maxwell-Lyons, F. (1967) *Amer. J. Ophthal.*, **63**, 1327-1355
- Attiah, M. A. H., El Kholy, A. M. & Omran, A. R. (1962) *Bull. ophthal. Soc. Egypt*, **55**, 395
- Bobb, A. A., Jr (1966) *Amer. J. Ophthal.*, **61**, 776
- Cochran, W. G. (1960) *Sampling techniques*, New York, Wiley
- MacCallan, A. F. (1908) *Ophthalmoscope*, **6**, 857
- WHO Expert Committee on Trachoma (1952) *Wld Hlth Org. techn. Rep. Ser.*, **59**
- WHO Expert Committee on Trachoma (1962) *Wld Hlth Org. techn. Rep. Ser.*, **234**
- Witts, L. J., ed. (1964) *Medical surveys and clinical trials*, London, Oxford University Press
- Yerushalmy, J., Harkness, J. T., Cope, J. H. & Kennedy, B. R. (1950) *Amer. Rev. Tuberc.* **61**, 443-464
-