INTERNATIONAL JOURNAL OF LEPROSY

Volume 59, Number 1
*Printed in the U.S.A.*

# Inter- and Intra-Laboratory Variation in the Reporting of Skin Smears in Leprosy[1,2]

Betsy Abraham and Annaiah Cariappa[3]

The need for monitoring the results of skin-smear examinations obtained in leprosy control programs has been recognized ([4, 5]). Studies on the variation in the reports of smears read by peripheral laboratories in control programs and re-read by base laboratories are few ([1, 8]). Such information would define the limits of variation, and would provide program managers necessary information about the expected variation in results during the course of monitoring skin smears.

Smears collected in the field and read at a peripheral laboratory situated in the control area of Schieffelin Leprosy Research and Training Centre (SLR&TC), Karigiri, India, were re-read 43 km away at the base laboratory situated at SLR&TC. The inter- and intra-observer variations in the base laboratory were also studied in order to define variation in the reporting of smears under ideal conditions, and to provide reference values against which inter-laboratory variation may be interpreted.

## MATERIALS AND METHODS

Skin smears were collected by two technicians under field conditions as part of the leprosy control activities of the Branch of Epidemiology and Leprosy Control, SLR&TC. These technicians had been trained earlier in all aspects of skin smears at the base laboratory. Their experience after training ranged between 1 and 14 years approximately.

Smears were taken from a minimum of four sites per patient: the right earlobe, the forehead, the chin, the gluteal region for males and the front of the thigh for females. These four sites are called "routine sites." Smears from these four sites are taken from all new patients, regardless of clinical classification, as baseline information. In addition, in suspected paucibacillary patients, smears are taken from active lesions; these are called smears from selective sites. In the control area, 80% to 90% of the patients are clinically classified as borderline tuberculoid or tuberculoid leprosy.

Smears were fixed immediately, and staining and grading were completed within 48 hr by the same technicians at the peripheral laboratory situated in the control area. The staining method used was the modified Ziehl-Neelsen method. Smears were graded according to Ridley's logarithmic scale ([6]). A minimum of 25 fields, when positive, and the entire smear when negative, are examined in all cases.

Routinely, all smear slides collected during a month are submitted to the base laboratory at the end of the month. At the base laboratory, 10% of the slides received each month are randomly selected for cross-checking which is completed within a day or two of receipt of the slides at the base laboratory. For the purposes of this study, the results obtained during the period 1985–1989 (4 years) were analyzed. During this period, a total of 7938 slides had been received and a total of 875 slides (approximately 10%) had been re-read at the base laboratory.

All slides were read blind at the base laboratory by two technicians who had between 9 and 18 years of experience in the smear technique. Results from the control area laboratory were entered alongside the results from the base laboratory only after the latter laboratory had completed examination. A report of 1+, when obtained by one technician at the base laboratory, was always counter-checked by the other.

Inter- and intra-observer variations among the two technicians at the base lab-

[1] Presented in part at the 15th Biennial Conference of the Indian Association of Leprologists, Visakhapatnam, Andhra Pradesh, India, November 1987.

[2] Received for publication on 7 May 1990; accepted for publication in revised form on 17 September 1990.

[3] B. Abraham, M.B.B.S., Ph.D., Research Fellow; A. Cariappa, M.B.B.S., M.D., Research Pathologist, Schieffelin Leprosy Research and Training Center, Karigiri 632106, India.

TABLE 1. *Inter-laboratory variations.*

| Results from | | Slides[a] | | Smears[b] | |
|---|---|---|---|---|---|
| Base lab | Control area lab | No. | % | No. | % |
| Negative | Negative | 788 | 90% | 3166 | 93% |
| Positive | Positive | 48 | 5.5% | 155 | 4.5% |
| Negative | Positive (1+) | 30 | 3.5% | 71 | 2% |
| Positive (1+) | Negative | 9 | 1% | 18 | 0.5% |

[a] One slide denotes 1 patient.
[b] Four smears on 1 slide; sometimes 3 or less when only selective sites taken.

oratory were studied as follows: 54 individual smears (11 slides) that had been collected routinely at the base hospital and were not part of the slides received from the control area, were given to both technicians, three times each on three different occasions. The smears were read under "ideal" conditions: Two smears from the batch of 54 were given to each technician per day. These smears were always given at the beginning of the day, to exclude the possibility of error due to fatigue. The original reports ranged from 0 to 6+ (0 = 13 smears; 1+ = 7; 2+ = 6; 3+ = 17; 4+ = 7; 5+ = 2 and 6+ = 2). The smears had been reported at the base laboratory up to 2 weeks prior to the experiment. The identification number etched at one end of the slide was covered with a thick white card, and the slides were coded by one of the authors before examination. The code was broken at the end of the experiment and code numbers were matched with identification numbers. It was ensured that each smear was re-graded within a month of its collection and staining, in order to overcome the possible effect of fading of stains.

The value of kappa (K) ([2]), a measure of agreement, was calculated for the results of the experiment on inter- and intra-observer variations.

## RESULTS

**Inter-laboratory variation.** Table 1 shows that 90% of the slides were reported negative by both laboratories. Therefore, the majority of patients from whom smears were collected were considered negative by both laboratories. Only 5.5% were reported positive by both laboratories. A small percentage (3.5%) was reported positive by the control area laboratory and negative by the base laboratory; in an even smaller percentage (1%), the reverse was true; all except one report showed a positivity of 1+. When individual smears were compared, the percentages were approximately similar (Table 1).

When positive slides (those with one or more smears positive) were checked for agreement, only 8.3% (Table 2) showed agreement. The most common finding was a difference of 1+. A difference of 2+ was rare. However, comparison of individual smears in the same set of slides revealed that there was 55% agreement. A difference of 1+ was found in 43%. This would indicate that different laboratories may agree on the grading of individual smears, but that agreement on sets of four smears may be rare.

**Intra- and inter-observer variation in base laboratory.** Figures 1 and 2 display the intra-observer variations among the two technicians routinely engaged in crosschecking smears. Each technician read the same smear three times, and 54 individual smears were read. When the results of all smears (positive and negative) were considered, both technicians could reproduce results with no difference in grading most of the time: technician 1 = 75% of the time, technician 2 = 80% of the time (data not shown). The values of kappa (K) calculated for technician 1 were 0.84, 0.81, and 0.74 for the first vs

TABLE 2. *Inter-laboratory variations in positives.*

| Grading difference | Slides | | Smears | |
|---|---|---|---|---|
| | No. | % | No. | % |
| Nil | 4 | 8.3% | 85 | 55% |
| 1+ | 40 | 83.3% | 67 | 43% |
| 2+ | 4 | 8.3% | 3 | 2% |

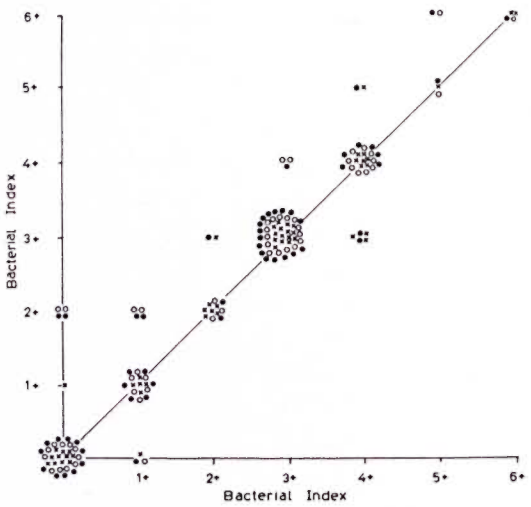


FIG. 1. Intra-observer variation of technician 1: 54 smears, each reported three times. Variation/agreement between reports 1 and 2 (x); reports 2 and 3 (O); reports 1 and 3 (●).

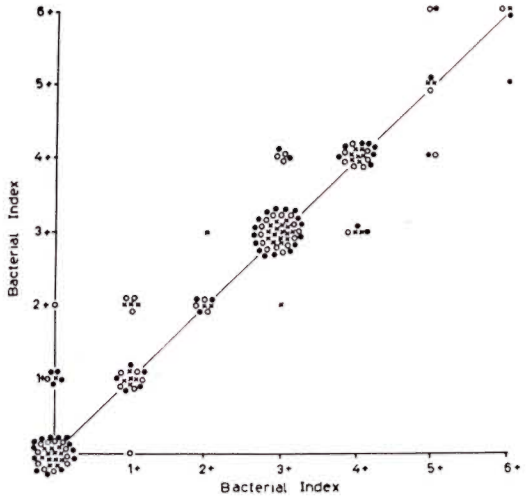


FIG. 3. Inter-observer variation of technician 1 vs 2: 54 smears, each reported three times by each technician. Variation/agreement of report 1 (x); report 2 (O); report 3 (●).

second, second vs third, and first vs third reports, respectively; for technician 2 the values were 0.91, 0.78, and 0.79, respectively. All values except one are greater than 0.75. Values ≥ 0.75 are taken to represent excellent agreement beyond chance ([2]). The intra-observer agreement for positive smears was 80% for technician 1 and 83% for technician 2 (Table 3).

Inter-observer variation is seen in Figure 3. Three sets of reports on 54 individual smears were plotted. Again, agreement was fairly common: both technicians agreed 66% of the time (data not shown). The values of K were 0.79, 0.73, and 0.76 for the first, second, and third sets of reports, respectively. The inter-observer agreement for positive smears between technicians 1 and 2 was 80% (Table 3).

Figures 1 to 3 also show that variations were unrelated to the grade of the smears.

When positive slides instead of individual positive smears were compared (Table 3), technician 1 reproduced results 54% of the time and technician 2, 45.5% of the time.

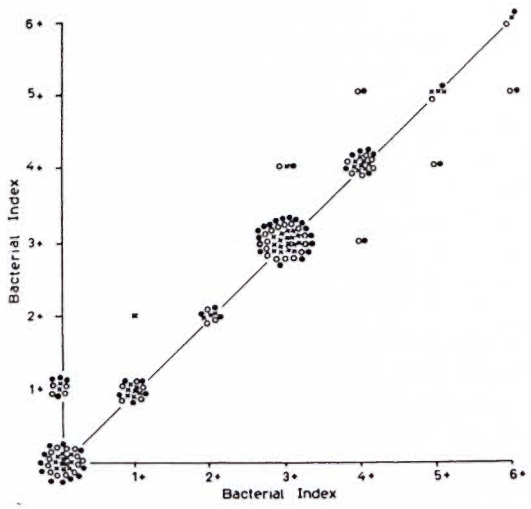


FIG. 2. Intra-observer variation of technician 2 (see Fig. 1 legend).

TABLE 3. *Intra- and inter-observer variations in positives.*[a]

| Technician | Grading difference | Slides[b] | | Smears[b] | |
|---|---|---|---|---|---|
| | | No. | % | No. | % |
| | | Intra-observer | | | |
| 1 | Nil | 6 | 54.5% | 33 | 80% |
| | 1+ | 5 | 45.5% | 8 | 20% |
| 2 | Nil | 5 | 45.5% | 34 | 83% |
| | 1+ | 6 | 54.5% | 7 | 17% |
| | | Inter-observer | | | |
| 1 vs 2 | Nil | 4 | 36.4% | 33 | 80% |
| | 1+ | 7 | 63.6% | 8 | 20% |

[a] 41/54 smears.
[b] Mean of 3 reports.

Both technicians agreed 36.4% of the time; the remainder in each case varied by 1+.

## DISCUSSION

The aim of this paper is to define the extent of variation present in the grading of skin smears when the same smears are read by a laboratory in a control area and by a base laboratory. This variation was then compared with the values obtained under "ideal" conditions from the study of inter- and intra-observer variation in the base laboratory.

There was hardly any variation between laboratories in the reporting of negative slides and smears. Vettom and Pritze describe similar results ([8]). However, a few slides reported positive (1+) by the base laboratory were reported negative by the control area laboratory (Table 1). It is important to define this number, since smear-negative patients would receive a shorter duration of therapy compared to smear-positive patients. As this paper shows, only 1% of the patients would have received inappropriate therapy in the sample studied.

Inter-laboratory variation in the reporting of positive slides was quite common. There was agreement in only 8.3% of positive slides (Table 2). The extent of disagreement or variation was between 1+ and 2+; the most common variation was 1+. The experiment studying inter-observer variation in the base laboratory shows that it is possible to report positive slides with total agreement between technicians in 36% of the slides (Table 3). The extent of the variation in the remaining 64% was 1+. When individual positive smears were compared, the agreement in the inter-laboratory study was 55% (Table 2), and in the intra-laboratory study (inter- and intra-observer variation) was 80% (Table 3). The inter-laboratory results suggest that complete agreement in the reporting of skin-smear slides is not a feasible proposition. This is reinforced by the finding that even under "ideal" conditions in the base laboratory, the intra-observer agreement in the reporting of positive slides was only between 45% and 55% and the inter-observer agreement was only 36% (Table 3). It appears, therefore, that while almost total agreement may be achieved when reporting negative slides, it is prudent to expect a variation of 1+ among positive slides.

Our results are different from those obtained by de Rijk, *et al.* ([1]). They observed variations greater than 1+ in 21% of their sample; some were $\geq$ 3+. Their results of inter-observer variation also show a much higher variation than seen in our study. Even though comparison of both sets of results is not strictly valid because of different sampling methods used, it is interesting to note that variations greater than 1+ were not unusual in de Rijk, *et al.*'s study ([1]).

In light of the results obtained by us and de Rijk, *et al.* it appears, as stated earlier, that a variation of 1+ in the reporting of positive slides is to be expected between laboratories reporting skin smears in a field program.

Vettom and Pritze suggest that the reasons for this variation in the reporting of skin smears may be a) the method of grading, which counts only a limited number (25) of fields, and b) nonuniform distribution of bacilli ([8]). A logical extension of this argument would be to count all fields in positive smears and to try to declump organisms. However, such procedures would increase the time, effort, and complexity of the technique to unacceptable levels, especially under field conditions. The success of laboratory tests in any field program depends to a large extent on the degree of simplicity without compromising too much of the quality of reporting. While the scanning of all fields when the smear is positive may be theoretically possible in regions where paucibacillary leprosy is common, it will be impossible in areas where multibacillary leprosy is endemic. Further, declumping of *M. leprae* is difficult even under the best of laboratory conditions, and it is unrealistic to attempt it in the field. In any case, changes in the technique would require meticulous standardization.

In retrospect, the redefinition of paucibacillary (PB) and multibacillary (MB) leprosy by the WHO ([9]) vis-a-vis the bacterial index (BI) seems appropriate. Earlier, PB patients were those with a BI of < 2+ at all sites ([10]). Our study shows that a variation of 1+ is present in the reporting of approximately 92% of all positive slides. Hence, the recent WHO recommendation

([9]) that all smear-negative patients be classified as PB and all smear-positive patients as MB seems pragmatic. However, even with this redefinition, the problem of misclassification remains, although to a slightly lesser degree. Two scenarios for misclassification can be described: a) A patient whose smear slide is reported as 1+ positive by the base laboratory is reported as negative by the field laboratory and is treated for PB leprosy. b) A patient whose smear slide is reported as negative by the base laboratory is reported as 1+ positive by the field laboratory and is treated with the MB regimen. Of these two examples, the former should be of greater significance to program managers since the aim of a field program is the rapid bacterial sterilization of the patient population. In our study only 1% of patients were reported 1+ positive by the base laboratory and negative by the field laboratory. We feel that this percentage may be acceptable to managers of a field program. On the other hand, 3.5% of the patients were reported negative by the base laboratory and 1+ positive by the field laboratory. In this group the primary aim of bacterial sterilization is met although it is a moot point whether time, money, materials and manpower have been wasted and patients overtreated. When the new definition of PB and MB leprosy ([9]) is applied to this study, the percentage of misclassification is 4.5% (1% + 3.5%). When the old definition ([10]) is applied, 5% of the patients (44 out of 875 slides, Tables 1 and 2) would have been misclassified. In areas unlike ours where MB leprosy is common, the percentage of misclassification would be higher with the old definition. This is because, as shown in our study, a variation of 1+ among positives is inherent in the technique.

Quality control (QC) of all aspects of the skin-smear technique is essential for the successful implementation of leprosy control programs. It must be remembered that the reporting of skin smears is the last event in a long chain of events, which includes the selection of sites in patients and the various aspects of collection, fixing, and staining. Therefore, any deviation in the technique in any of these aspects would also be reflected in the final report. (The reader is referred to reviews that address these issues[3, 7].) In the few published reports on QC of smears ([1, 8]), different methods have been used to analyze the variation in reporting. We feel that three simple guidelines are all that are required by base laboratories to monitor the reporting of smears in peripheral laboratories: a) positive slides may vary only up to 1+; b) negative slides must agree; c) the percentage of slides reported as 1+ positive in the base laboratory and as negative in the control area laboratory should be low (approximately 1%).

## SUMMARY

This paper defines the variations in the reporting of skin smears between a base and field laboratory in a leprosy control program. Ten percent of all slides read by the field laboratory in a control area were re-read by the base laboratory. There was almost no variation in the reporting of negative slides, but a variation of 1+ was present in approximately 92% of positive slides. Thus, there was agreement in approximately 8% of positive slides. This paper also defines the variations in the reporting of positive slides under "ideal" conditions by describing the results of a study on intra- and inter-observer variations among technicians at the base laboratory. There was between 45% and 55% agreement within observers and about 36% agreement between observers. The results of both studies are compared. Simple guidelines are derived to monitor the reporting of skin smears in leprosy control programs.

## RESUMEN

Este trabajo analiza las variaciones en los reportes de los resultados de extendidos de linfa cutánea entre un laboratorio base y un laboratorio "de campo," en un programa de control de la lepra. El 10% de todas las laminillas leídas por el laboratorio de campo en un área controlada fueron releídas en el laboratorio base. Aunque casi no hubieron variaciones en los reportes de las laminillas negativas, en aproximadamente el 92% de las laminillas positivas se encontró una variación de 1+. Asi, la concordancia fue de aproximadamente el 8% en las laminillas positivas. El trabajo también describe las variaciones en los reportes de las laminillas positivas bajo condiciones "ideales" al analizar los resultados de un estudio sobre las variaciones intra- e inter-observadores, efectuado en el laboratorio base. La concordancia fue del 45% al 55% dentro de los observadores y de aproximadamente el 36% entre los

observadores. Los resultados de ambos estudios se compararon y de aquí se derivaron lineamientos simples para valorar los reportes sobre los extendidos de linfa cutánea en los programas de control de la lepra.

## RÉSUMÉ

Cet article décrit les variations dans les résultats des frottis cutanés rapportés par un laboratoire central et un laboratoire périphérique dans un programme de lutte contre la lèpre. Dix pourcents de toutes les lames par le laboratoire périphérique d'une région furent relues au niveau du laboratoire central. Il n'y avait pratiquement pas de différence dans les résultats rapportés pour les lames négatives, mais une différence de 1+ était présente pour environ 92% des lames positives. Il y avait donc accord pour 8% des lames positives. Cet article décrit également les variations dans la notification de lames positives dans des conditions "idéales" en détaillant les résultats d'une étude sur la variabilité inter- et intra-observateurs parmi les techniciens du laboratoire central. Il y avait un accord intra-observateurs variant de 45% à 55%, et un accord inter-observateurs d'environ 36%. Les résultats de ces deux études sont comparés. Des directives simples en sont déduites pour contrôler les résultats des frottis cutanés rapportés au niveau des programmes de lutte contre la lèpre.

**Acknowledgments.** The technicians at the base laboratory are Mr. K. Vilvanathan and Mrs. Shanthi Yesupadam. The technicians at the laboratory in the control area are Mr. S. John and Mr. Y. Inbanathan. We thank Dr. K. Jesudasan, Head, Branch of Epidemiology and Leprosy Control, for permitting analysis of the data and for helpful discussions. The encouragement given by Dr. C. J. G. Chacko, Head, Branch of Laboratories, is appreciated.

## REFERENCES

1. DE RIJK, A., NILSSON, T. and CHONDE, M. Quality control of skin smear services in leprosy programmes: preliminary experience with inter-observer comparison in routine services. Lepr. Rev. **56** (1985) 177–191.
2. FEINSTEIN, A. R. The measurement of interrater agreement. In: *Clinical Epidemiology. The Architecture of Clinical Research.* New York: Igaku-Shoin/Saunders, 1985, pp. 212–236.
3. GEORGIEV, G. D. and MCDOUGALL, A. C. The bacteriological examination of slit-skin smears in leprosy control programmes using multiple drug therapy: a plea for radical changes in current operational methodology. Indian J. Lepr. **59** (1987) 373–385.
4. GEORGIEV, G. D. and MCDOUGALL, A. C. Skin smears and the bacterial index (BI) in multiple drug therapy leprosy control programs: an unsatisfactory and potentially hazardous state of affairs. Int. J. Lepr. **56** (1988) 101–104.
5. INDIAN DIRECTORATE GENERAL OF HEALTH SERVICES, LEPROSY DIVISION. Report on second independent evaluation of National Leprosy Eradication Programme. New Delhi: Ministry of Health and Family Welfare, 1987, pp. 27–29.
6. RIDLEY, D. S. Bacterial indices. In: *Leprosy in Theory and Practice.* Cochrane, R. G., ed. Bristol: John Wright and Sons Ltd., 1959, pp. 371–372.
7. SEHGAL, V. N. and JOGINDER. Slit-skin smear in leprosy. Int. J. Dermatol. **29** (1990) 9–16.
8. VETTOM, L. and PRITZE, S. Reliability of skin smear results: experience with quality control of skin smears in different routine services in leprosy control programmes. Lepr. Rev. **60** (1989) 187–196.
9. WHO EXPERT COMMITTEE. Sixth report. Geneva: World Health Organization, 1988, p. 15. Tech. Rep. Ser. 768.
10. WHO STUDY GROUP. Chemotherapy of leprosy for control programmes. Geneva: World Health Organization, 1982, p. 240. Tech. Rep. Ser. 675.