J. Commun. Dis. 42(3) 2010 : 191-199

Assessment of X-ray Readers in TB Prevalence Surveys

Gopi PG, Vasantha M*, Muniyandi M**

(Received for Publication: Oct. 2009)

Abstract

A prevalence study was carried out a rural community in Tiruvallur district in Tami Nadu to standardize the method of assessing an X-ray reader in tuberculosis (TB) prevalence surveys by means of different measures of agreement between the reader and a Standard Reader (SR).

The exercise on assessing the X-ray readers was carried out on two occasions; one involving three trainee readers (R_1, R_2, and R_3), and the other involving one trainee reader (R_4). The extent of agreement was estimated using Kappa statistics (K), over-diagnosis, under-diagnosis, crude agreement and prevalence adjusted bias adjusted kappa (PABAK).

The overall performance of readers R_1, R_2 and R_3 was not satisfactory in terms of K (21, 34 and 14%) in the first assessment. The K, over-diagnosis and under-diagnosis were estimated to be 61, 28 and 4% for R1, 63, 18 and 4% for R_2 and 58, 31 and 5% for R_3 in the final assessment. This suggested that R_2 performed well compared to the other two readers. The K was 68% for R_4 in the first assessment. As the over-diagnosis was to the extent of 40%, the trainee reader underwent one more assessment. The K was 64% which was as good as before, but there was no improvement in the over-diagnosis (43.5%) in the second assessment.

Based on the performance, only one reader (R_2) was certified as qualified for X-ray reading in the first occasion while the reader (R_4) assessed in the second occasion was not qualified. These findings were subject to the inherent variation in the SR's readings against which the readers were assessed.

Keywords : TB, X-ray reader, Standard reader, Agreement, Over- diagnosis, Under-diagnosis.

* Tuberculosis Research Centre (ICMR), Chennai
Correspondence to: Dr. M. Muniyandi, Health Economist, Tuberculosis Research Centre (ICMR), Mayor V.R.Ramanathan Road Chetput, Chennai–600 031 India. Ph. 044-28369573
Email : mmuniyandi@yahoo.com

INTRODUCTION

In TB prevalence surveys, chest radiography is one of the screening tools for diagnosis of tuberculosis (TB). An X-ray reader who has the recognized qualifications to undertake the reading of the film reads the film, and identifies the presence of any abnormality that requires further investigation of sputum specimens for diagnosis of TB by smear and culture. For this, the reader should have written guidelines indicating the types of abnormalities. No chest radiographic pattern is absolutely typical of TB though certain configurations have traditionally been viewed as highly suggestive of TB. Chest radiography can undoubtedly be very helpful in localizing abnormalities in the lung but to establish its etiology, further investigations are required and in TB bacteriological examination will confirm the diagnosis. Over dependence on X-ray for diagnosis often overlooks the inherent limitations associated with its interpretation. The observer error determines the reliability of X-ray diagnosis to a great extent. Besides this, it is difficult to assess the activity of the lesion and to determine the etiology on X-ray alone. Chest X-ray (CXR) being a two-dimensional picture conceals almost 20-30% of the lung field because of overlapping structures. This further limits its interpretation. Hence, purely radiological criteria cannot give satisfactory evidence of TB in individual patients unless investigated further. In spite of the limitations described above, chest radiography is used as a survey tool for prevalence surveys.

In TB prevalence surveys, 70 mm photo-fluorogram posteroanterior (PA) film, Mass Miniature Radiograph (MMR), is recommended for diagnosis of TB. The usefulness of chest radiography is determined largely by the reader's ability to detect abnormal shadows and interpret them correctly. This implies not missing or under reading (under-diagnosis) abnormalities suggestive of TB and not over-reading (over-diagnosis) normal as abnormal. The ability may vary not only from one reader to another (inter-reader variation) but also by the same reader (intra-observer variation).[1] Several studies from India and elsewhere have shown that there is a tremendous amount of over-diagnosis and inter and intra reader variation in interpreting X-ray findings.[2] In public health terms, false positive diagnosis will result in inefficient use of resources, and false negative diagnosis may pose a threat to public health through spread of TB.[3] Moreover, the performance of CXR depends on the quality of film, which in turn depends on the functioning of CXR machine, the reagents and the developing process. To minimize variation in reading, the film is given to two independent readers. In case of discrepancy between the first two readers a third reading is obtained from an 'umpire' reader. This is to pick out people with abnormal X-ray suggestive of TB as read by at least two X-ray readers.

There has always been a need to train and assess fresh X-ray readers in reading MMR films in surveys. The objective of the study was to standardize the methods of assessing X-ray reader by means of different measures of agreement between the reader and a SR.

MATERIALS AND METHODS

In community surveys, all persons eligible for X-ray are directed to a central place where X-ray technician takes X-ray of all participants.

The processed X-ray films were read by two independent readers for further investigation. Assessment of the new X-ray readers is also done using the MMR films used in the survey.

Procedure for assessing X-ray readers in reading chest X-ray films: First, the X-ray reader needs to be trained in reading films. The qualified reader (SR) gives initial training to

the X-ray trainee reader. Initially, they are given intensive training with the help of the manual and other materials in categorizing each photoflurogram as normal, non-TB, possible or probable TB. This includes explaining to the trainee reader the purpose of reading the films, the abnormalities and the different classifications such as, lung pathology other than TB, inactive TB, possibly active TB and probably active TB and how to code it in the X-ray result sheet with abnormality marked in the lung figure using the manual for X-ray reading and sputum collection. The trainee reader goes through the film, tries to identify the existing abnormalities and then compares with the readings of the SR. The trainee reader continues this exercise with subsequent films and becomes more familiar in reading the films. The trainee reader may view more films and discuss with the SR in case of any doubts in coding the results (annexure). The trainee reader should read adequate number of films during the training period of about a week till the trainee becomes confident in reading the films independently before he is assessed. For each person, it is essential that the film is correctly labeled with the survey unique number and that the quality of the film is good.

Pre-assessment : After the initial training, the trainee reader reads independently about 2000 films without the help of the SR or his reading using the guidelines for classification of the films. These readings are compared with SR's readings by different kinds of measures of agreement like Kappa (K), over-diagnosis, under-diagnosis and prevalence adjusted bias adjusted kappa (PABAK) as given below. If the performance is satisfactory then the trainee reader goes for a final assessment. In case of disagreement with the SR, the trainee reader should identify the area where he disagreed and clarify with the SR. The SR should assist the trainee in improving his skill in reading the films correctly without missing any abnormality or picking up any normal as abnormal.

Final assessment : The trainee reader reads another set of films independently and gives his readings. If the inter-reader agreement is satisfactory the reader will be given the same set of rolls for his intra-reader agreement.

In community surveys, the purpose of reading the MMR film is to decide whether the corresponding person is eligible for sputum collection (code C or D or digit code ≥5 suffixed with A or B as given in the annexure). So, we condense the tabulation in a dichotomous classification as follows.

Trainee reader	SR		Total
	E	NE	
E	A	B	a+b
NE	C	D	c+d
Total	a+c	b+d	a+b+c+d (=N)

E = Eligible for sputum collection (abnormal) based on X-ray reading

NE = Not eligible for sputum collection (normal) based on X-ray reading

The extent of agreement may be measured by the following measures

Crude agreement (P_0)=(a+d)/N: Observed agreement between the two readers

Over diagnosis = (b/(a+b) : The proportion of abnormal as identified by the trainee reader that are really normal (< 20%)

Under diagnosis=(c/(c+d) : The proportion of normal identified by the trainee reader that are really abnormal (< 5%).

$K=(P_0-P_c)/(1-P_c)$: The extent of agreement between the two readers after adjusting for chance expected agreement, where $\mathbf{P_0}$ is the crude agreement and $\mathbf{P_c}$ is the proportion of agreement expected by chance.[4,5] Performance is assessed as K<0.20, poor agreement; 0.21-0.40, fair agreement; 0.41-0.60, moderate agreement; 0.61-0.80 good agreement; and 0.81-1.00 very good agreement (K>=0.6).

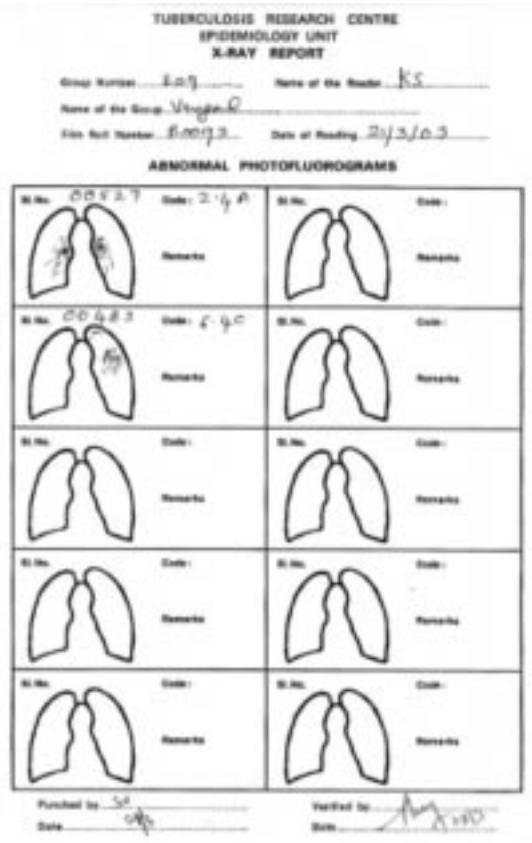
PABAK = $2P_0-1$: An index of agreement

between two observers that adjusts kappa for differences in the prevalence of positives and negatives and for bias between two observers.[6]

If the agreement is within the limits as given in the parenthesis against the measures (K, over-diagnosis and under-diagnosis) given above, the trainee reader may be certified as the X-ray reader. If the trainee reader is not through in the assessment, he needs further training and goes for another assessment.

In this study, the exercise on assessing the readers was carried out on two occasions; one involving three trainee readers (R_1, R_2 and R_3) and the other involving one trainee reader (R_4). In the first occasion all three trainee readers were given training in the basic concept of reading the X-ray film, the procedure of recording the abnormality, if any, on an X-ray form (Fig.) and the relevant code (Annexure). All readings of X-ray films were recorded in the form, which contains group number, individual number, film roll number, date of X-ray exposure, date of X-ray reading and the name of X-ray reader. After initial training the extent of agreement was estimated. The films they disagreed in the reading were discussed and clarified. After they became confident they were asked to read another batch of X-ray films and the results were compared with that of the SR. The films were again given to the same readers to assess the internal consistency. In the second occasion, R_4 was asked to read the X-rays after he acquainted himself with the reading in comparison with the SR's reading. After the trainee satisfied with the performance, a set of films was read independently. The readings were compared with that of the SR and the extent of agreement between the two was estimated.

Data were scrutinized and entered twice in order to ensure accuracy, corrected for discrepancy and missing information. Data analysis was carried out using SPSS software 13.0 version.


TUBERCULOSIS RESEARCH CENTRE
EPIDEMIOLOGY UNIT
X-RAY REPORT

ABNORMAL PHOTOFLUOROGRAMS

Figure 1 : X-ray coding sheet (see Annexure)

RESULTS

First occasion: In the pre-assessment, the readers read 1510 X-ray films. The results of the pre-assessment of the three readers R_1, R_2 and R_3 are given in Table 1 (a, b, c). It was observed that the extent of agreement beyond chance agreement (K) was 21, 34 and 14% for R_1, R_2 and R_3 respectively. The over-diagnosis was to the extent of 77, 58 and 82% respectively. The under- diagnosis was to the extent of 4 % for R_1 and R_2; and 5% for R_3. The PABAK was 79, 87, and 80% for R_1, R_2 and R_3 respectively. The overall performance of all readers was not satisfactory in terms of over diagnosis.

Table 1a. First assessment between trainee reader R_1 and SR

Trainee reader R_1	**SR**		
	Eligible	**Not-Eligible**	**Total**
Eligible	28	96	124
Not-Eligible	**59**	**1284**	**1343**
Total	**87**	**1380**	**1467**
Kappa statistics	= 21%		
Crude agreement	= 90%		
Over- diagnosis	= 77%		
Under- diagnosis	= 4%		
PABAK	= 79%		

Table 1b. First assessment between trainee reader R_2 and SR

Trainee reader R_2	**SR**		
	Eligible	**Not-Eligible**	**Total**
Eligible	**29**	**40**	**69**
Not-Eligible	59	1354	1413
Total	88	1394	1482
Kappa statistics	= 33%		
Crude agreement	= 93%		
Over- diagnosis	= 58%		
Under- diagnosis	= 4%		
PABAK	**= 87%**		

Table 1c. First assessment between trainee reader R_3 and SR

Trainee reader R_3	**SR**		
	Eligible	**Not-Eligible**	**Total**
Eligible	18	82	100
Not-Eligible	68	1285	1353
Total	86	1367	1453
Kappa statistics	= 14%		
Crude agreement	= 90%		
Over- diagnosis	= 82%		
Under- diagnosis	= 5%		
PABAK	**= 79%**		

Note : As technically inadequate X-ray films read by either trainee reader or SR were excluded for analysis, the number of X-ray films was not equal in each pair.

The extent of agreement in the final assessment is given in Table 2 (a,b,c). In this assessment the readers read 1864 films. The K was estimated to be 61, 63 and 58% respectively for R_1, R_2 and R_3 respectively. The

Table 2a. Final assessment between trainee reader R_1 and SR

Trainee reader R_1	**SR**		
	Eligible	**Not-Eligible**	**Total**
Eligible	**89**	**35**	**124**
Not-Eligible	65	1670	1735
Total	154	1705	1859
Kappa statistics	= 61%		
Crude agreement	= 95%		
Over- diagnosis	= 28%		
Under- diagnosis	= 4%		
PABAK	= 89%		

Table 2b. Final assessment between trainee reader R_2 and SR

Trainee reader R_2	**SR**		
	Eligible	**Not-Eligible**	**Total**
Eligible	**83**	**18**	**101**
Not-Eligible	70	1682	1752
Total	153	1700	1853
Kappa statistics	= 63%		
Crude agreement	= 95%		
Over- diagnosis	= 18%		
Under- diagnosis	= 4%		
PABAK	**= 91%**		

Table 2c. Final assessment between trainee reader R_3 and SR

Trainee reader R_3	**SR**		
	Eligible	**Not-Eligible**	**Total**
Eligible	**82**	**37**	**119**
Not-Eligible	**69**	**1660**	**1729**
Total	**151**	**1697**	**1848**
Kappa statistics	= 58%		
Crude agreement	= 94%		
Over- diagnosis	= 31%		
Under- diagnosis	= 4%		
PABAK	**= 89%**		

over diagnosis was to the extent of 28, 18 and 31% respectively, while the under-diagnosis was maintained at 4% by all readers. PABAK was 89, 91, and 89% for R_1, R_2 and R_3 respectively. It could be observed that R_2 performed well compared to the other two readers. The agreement was satisfactory in terms of K, over-diagnosis and under-diagnosis. The internal consistency was also satisfactory for the R_2 as observed from Table 3.

Table 3 : Internal consistency for reader R_2

Trainee reader	SR		
	Eligible	**Not-Eligible**	**Total**
Eligible	**77**	**26**	**103**
Not Eligible	**22**	**1734**	**1756**
Total	**79**	**1760**	**1859**
Kappa statistics	= 75%		
Crude agreement	= 97%		
Over- diagnosis	= 25%		
Under- diagnosis	= 1%		
PABAK	**= 95%**		

Second occasion : The extent of agreement between the R_4 and the SR on the second occasion is shown in Table 4 (a, b). A total of 2756 X-ray films were read by both SR and R_4 for the first assessment. The R_4 read 2605 films as normal and 151 as abnormal suggestive of TB against 2649 normal and 107

Table 4a : First assessment between trainee reader and SR

Trainee reader	SR		
	Eligible	**Not-Eligible**	**Total**
Eligible	**90**	**61**	**151**
Not-Eligible	17	2588	2605
Total	107	2649	2756
Kappa statistics	= 68%		
Crude agreement	= 97%		
Over- diagnosis	= 40%		
Under- diagnosis	= 0.7%		
PABAK	**= 94%**		

Table. 4b : Second assessment between trainee reader and SR

Trainee reader	SR		
	Eligible	**Not-Eligible**	**Total**
Eligible	**105**	**81**	**185**
Not-Eligible	28	3574	3602
Total	133	3655	3788
Kappa statistics	= 64%		
Crude agreement	= 97%		
Over- diagnosis	= 44%		
Under- diagnosis	= 0.8%		
PABAK	**= 94%**		

abnormal by SR. The K was estimated to be 68%. Over-diagnosis of the reader was to the extent of 40% and under-diagnosis was 0.7%. PABAK was 94% (crude agreement 97%). Here, even though the agreement between the two readers in terms of K was good, over diagnosis was more than expected.

In the second assessment, 3788 X-ray films were read by the trainee reader (Table 5). The K between the two readers in the assessment was 64% as good as before; over-diagnosis was 43.5%, under- diagnosis 0.7% and PABAK 94%.

DISCUSSION

The study findings showed that on the first occasion, the performance was poor for all readers in the first assessment. In the second assessment, the second reader performed well in terms of the measures agreement between the SR and trainee reader whereas in the second occasion even though the agreement was good in terms of K, over- diagnosis was very high; 40 and 44% in the first and second assessments. The workload for sputum collection would be very high and hence the trainee reader was not qualified to read the X-ray films. Since over-diagnosis is inherent in X-ray, the cut-off at 20% for over-diagnosis was arbitrarily decided allowing inherent variability in the interpretation in the X-ray.

This means there is a 20% increase of work load in sputum collection. This appears to be reasonable as this would lead to examination of sputum from more individuals for diagnosis of TB. Still, we can have a lower value of over-diagnosis for better performance. But it should be borne in mind that we assess the performance of the trainee reader in comparison with the SR whose reading was considered as the gold standard. This may not be always true as there could be variation in his reading also. In the second occasion of our study, over-diagnosis for the trainee reader was 40% in the first assessment. So, he underwent for one more assessment. Again, over-diagnosis was more than 40% in the second assessment. This showed that the reader disagreed with SR in reading X-ray films as before.

The Russian study[3] highlighted the highest levels of agreement among radiologists when compared with TB specialist and respiratory specialists. The level of experience (years of work in the specialty) influenced agreement on presence of abnormalities and cavities. The study concluded that population screening for TB may be less optimal due to limited agreement on interpretation of chest X-rays. In a study[7] conducted in South Africa, the K between two readers was 0.69 for abnormalities consistent with TB and suggested that chest radiography may be much more useful as a screening tool for TB. However, the authors emphasized the need to evaluate the reading methodology by more readers and different study settings. Another study[8] conducted in Nairobi, Kenya to establish the performance of chest X-ray showed that when cost of treatment was considered CXR followed by ZN microscopy was more cost effective and recommended the introduction of a scoring system, clinical conferences and a system of chest X-ray quality control to contribute to improved diagnostic performance. In our set up we follow the former procedure of chest radiography followed by Fluorescent microscopy in our community surveys.

It is observed in our community surveys that 4-8% of the population is abnormal on CXR suggestive of tuberculosis. While reading the films prevalence index (PI) defined as PI= (a-d) / N is large resulting in low value of K. Similarly, for large value of bias index (BI) defined as BI= (b-c) / N, kappa is higher than when the BI is low or absent. The K coefficient is influenced by prevalence index PI and BI. So, K should be interpreted in the light of these two indices. The kappa coefficient that is adjusted for these two indices is referred to as PABAK and its use is critical because the effects of bias and prevalence on the magnitude of kappa are themselves informative and should not be adjusted and thereby disregarded.[9] Thus, the PABAK could not be considered to generate a value for kappa that does not relate to the situation in which the original ratings were made. Therefore, the PABAK coefficient on its own is uninformative because it relates to a hypothetical situation in which no prevalence or bias effects are present. However, if PABAK is presented in addition to, rather than in place of, the obtained value of K, its use may be considered appropriate because it gives an indication of the likely effects of prevalence and bias alongside the true value of K.[10]

In prevalence surveys, the X-ray unit is usually mounted onto mobile van (vehicle) which is used to carry out radiological examination. Taking the unit to difficult terrain without passable roads may pose many problems. In such cases, it may be possible to establish a center and all persons will be directed to this centre for taking the X-ray. This introduces a potential problem in that some participants may not wish to take the time to go for the examination and special

arrangements must be made to transport the patients. Bringing back the films to the main centre for processing, arranging for independent reading and preservation of the films for future reference may be difficult. To avoid all the above problems, the Digital X-ray is preferred in prevalence surveys. However, a standardized procedure has not yet been defined for using the Digital X-ray.

Our study demonstrated the procedure for assessing the X-ray readers using the measures of agreement like K, over-diagnosis and under-diagnosis. The assessment of X-ray readers is subject to limitation due to the observer error in the interpretation of chest radiograph. A small proportion of people without TB are subjected to unnecessary sputum examination due to over reading of films. This, however, can be minimized by intensive training initially and then assessing the readers for intra and inter reader variation. Our exercise warrants for further exploratory studies using innovative methods for inter and intra observer agreement.

ACKNOWLEDGMENTS

The authors thank the staff of Statistics division and Electronic Data Processing Department of Epidemiology Unit for scrutiny and computerization of data pertaining to this study. The guidance rendered by Dr. C. Kolappan and Dr. K. Sadacharam in this endeavour is gratefully acknowledged. The authors appreciate the field staff for undertaking the chest radiography of all subjects included from the survey area for this exercise. The authors are grateful to all the X-ray readers involved in reading the X-ray films required for this exercise. The study was supported in part by the World Health Organization (WHO) with financial assistance provided by United States Agency for International Development under the Model Direct Observation Treatment-Short course (DOTS) project.

ANNEXURE

X-ray reading Code:

0.0 Normal (not to be recorded by readers)

(11) (11) Lost, not read

0.(11). Technically inadequate

1. Extra respiratory

1.1. Cardiac

1.2. Vascular

1.3. Bony abnormalities (eg., scoliosis)

2. Respiratory, definitely extra pulmonary

2.1. Very dense spot or spots in hilar region (calcification)

2.2. Obliterated Costo-phrenic angle and/or pleural scar and/or pleural calcifications

2.3. Evidence of chest surgery

2.4. Enlarged mediastinal and/or hilar glands

2.5. Basal-parietal opacity, indicative of pleurisy with effusion (in any area)

2.6. Pneumothorax or hydro pneumothorax

2.7. Special pathology not specified above

3-9 Opacity or opacities in lung fields

3. Very dense and very well demarcated (eg. calcifications)

4. Dense and well demarcated (eg., fibrosis)

5. Special patterns

5.1. Uniformly dense, round opacity, single or multiple (eg., cyst)

5.2. Atelectasis

5.3. Consolidation

5.4. Less dense opacity combined with cardiac abnormality (6 and 1.1 both present in one person)

6. Less dense opacity, or less well demarcated (eg., infiltration)

7. Ill-demarcated or doubtful cavity

8. Well-demarcated cavity or cavities, each less than 4 cm (less than 6mm on the small film)

9. At least one well-demarcated cavity more than 4 cm (more than 6mm on the small film)

In case of multiple lesions, only the most serious is recorded. For the reading 3, 4, 6, 7, 8 and 9 above, use a second digit from the following :

Total extent of opacities	Location		
	Single opacity	More opacities in one lung only	Both Lungs
Less than one square centimeter (1.5 mm. sq. on the small film) or linear bands	1	2	3
Less than one sixth of total area of lung fields	4	5	6
More than one sixth of total area of lung fields	7	8	9
Small spots, widely disseminated in both lungs	–	–	0

For a first digit reading of 1-9, use the following alphabets: as the third code (except calcifications)

A – Other lung pathology than TB

B – Probably TB but inactive

C – Probably TB, possibly active

D – Probably TB and active

REFERENCE

1. Santadevi T. The Scientific basis of DOTS, TB control in India (Chapter 4).

1. Toman's tuberculosis. Case detection, treatment, and monitoring- questions and answers, Second edition (Chapter 12), WHO, Genewa, 2004.

2. Balabanova Y, coker R, Fedorin I, Zakharova S, Plavinskij S, Krukov N, Atun R, Drobiewski F. Variability in interpretation of chest radiograph among Russia clinician and implication for screening programmes: observational study. *BMJ*. **331**, 2005: 379.

3. Fleiss JL. *Statistical methods for rates and proportion*. 2nd edition. New York, NY: John Wiley & Sonns Inc; 1981.

4. C.D.J. Holman, I.R. James, P.J. Heenen, L.R. Matz, J.B. Blackwell, G.R.H. Kelshall, A. Singh & R.E.J. Ten Seldam. An improved method of analysis of observer variation between pathologists. (1982) *Histopathology* **6**, 581-589.

5. Ted Byrt, Janet Bishop & John B. Carlin. Bias, Prevalence and Kappa. Clin Epidemiol, **46**(5), 1993: 423-429.

6. Van Cleeff MRA, Kivihya Ndugga LE, Meme H, Odhian JA, Klatser PR. The role and performance of chest X-ray for the diagnosis of tuberculosis: A cost-effectiveness analysis in Nairobi, Kenya. *BMC infectious Diseases* 2005, **5**: 111.

7. Den Boon S, Bateman E.D, Enarson D.A, Borgdorff M.W, Verver S, Lombard C.J, Irusen E, Beyers N, White N.W. Development and evaluation of a new chest radiograph reading and recording system for epidemiological surveys of tuberculosis and lung disease. *Int J Tuberc Lung Dis* 2005; **9**(10): 1088-1096.

8. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *J Clin Epidemiol.* 2000; **53**: 499-503.

9. Julius Sim, Chris C Wright. The Kappa Statistics in reliability studies: Use, interpretation, and sample size requirements. *Physical Theraphy*. **85**(3), 2005.