Inter-Observer Variability in the Detection and Interpretation of Chest X-Ray Anomalies in Adults in an Endemic Tuberculosis Area

Boniface Moifo¹,²*, Eric Walter Pefura-Yone¹,³,⁴, Georges Nguefack-Tsague¹,⁵, Marie Laure Gharingam¹, Jean Roger Moulion Tapouh², André-Pascal Kengne⁶, Samuel Nko’o Amvene¹,²

¹Faculty of Medicine and Biomedical Sciences (FMBS), The University of Yaoundé I, Yaoundé, Cameroon
²Department of Radiology and Radiation Oncology, FMBS, Yaoundé, Cameroon
³Department of Internal Medicine and Specialties’, FMBS, Yaoundé, Cameroon
⁴Service of Pneumology A, Jamot Hospital of Yaoundé, Yaoundé, Cameroon
⁵Department of Public Health, FMBS, Yaoundé, Cameroon
⁶South African Medical Research Council & University of Cape Town, Cape Town, South Africa

Email: *bmoifo@yahoo.fr, pefura2002@yahoo.fr, nguefacktsague@yahoo.fr, mclor01@yahoo.fr, tapouh@yahoo.fr, apkengne@yahoo.com, nkoo_as@yahoo.com

Received 25 June 2015; accepted 9 September 2015; published 11 September 2015

Copyright © 2015 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).

http://creativecommons.org/licenses/by/4.0/

Abstract

Purpose: To assess the inter-observer agreement in reading adults chest radiographs (CXR) and determine the effectiveness of observers in radiographic diagnosis of pulmonary tuberculosis (PTB) in a tuberculosis endemic area. Methods: A quasi-observational study was conducted in the Pneumology Department of Yaoundé Jamot Hospital (Cameroon) from January to March 2014. This included six observers (two chest physicians, two radiologists, two end-training residents in medical imaging) and 47 frontal CXRs (4 of diffuse interstitial lung disease, 6 normal, 7 of lung cancers, 7 of bacterial pneumonia, 23 of PTB). The sample size was calculated on the basis of an expected 0.47 Kappa with a spread of 0.13 (α = 5%, CI = 95%) for six observers and five diagnostic items. The analysis of concordance was focused on the detection of nodules, cavitary lesions, pleural effusion, adenomegaly and diagnosis of PTB and lung cancer. These intervals of kappa coefficient were considered: discordance (<0.0), poor agreement (0.0 - 0.20), fair (0.21 - 0.40), moderate (0.41 - 0.60), good (0.61 - 0.80), excellent (>0.81). Results: The average score for the detection of caverns was the highest (58.3%) followed by that of the correct diagnosis of tuberculosis (49.3%). Pneumologists had the highest proportions of correct diagnosis of tuberculosis (69.6% and 73.9%) and better inter-observer agreement (k = 0.71) for PTB diagnosis. Observers were
more in agreement for the detection of nodules (0.32 - 0.74), adenomegalies (0.43 - 0.69), and for the diagnosis of cancer (0.22 - 1) than for the diagnosis of tuberculosis (0.19 - 0.71). Disagreements were more frequent for the detection of pleural effusions (~0.08 - 0.73). Conclusion: The inter-observer agreement varies with the type of lesions and diagnosis. Pneumologists were most effective for the diagnosis of pulmonary tuberculosis. Observers were more in agreement for the detection of nodules and the diagnosis of cancer than for the diagnosis of pulmonary tuberculosis.

Keywords
Inter-Observer Variability, Concordance, Pulmonary Tuberculosis, Nodules, Caverns, Lung Cancer, Chest Radiography, Kappa

1. Introduction
Chest X-ray (CXR) is the most prescribed radiography in developing countries. It plays a major role in management of many thoracic diseases [1]. It is widely used in the preoperative workup and screening [2] [3], and it provides evidence for prescription of many chest CT-scans [1] [4].

The complexity of CXR image is source of great variability in the detection and interpretation of pulmonary anomalies between readers [1] [5]-[7]. Several studies have been conducted to assess the variability in the diagnosis of pulmonary tuberculosis (PTB) [8]-[10], of pneumoconiosis [2] [3], of lung cancers [4] [11], of pneumonia [8] [12] [13] and of lung nodules [7]. CXR is an important tool in the diagnosis of pulmonary tuberculosis [10] [14] [15]. In developing countries, where CXR is usually the only available or accessible chest imaging test and where PTB is endemic [16] [17], inter-observer variability in the interpretation of CXR has not been studied.

The purpose of this study was to assess the concordance in reading adult CXRs between radiologists, pneumologists and senior residents in medical imaging and to determine their effectiveness in suggesting diagnosis of pulmonary tuberculosis among other chest diseases.

2. Materials and Methods
It was a cross-sectional quasi-observational study carried out in Yaounde (capital city of Cameroon) from January to March 2014, including six observers and 47 CXRs.

2.1. Selection of Observers
Six observers, all working in university-affiliated hospitals, selected by convenience agreed to participate in the study: two pneumologists of six and 14 years of experience, two radiologists of two and six year’s experience, and two senior residents in medical imaging. Readers were categorized as “radiologist”, “pneumologists” and “residents”. Pneumologists were from the department of pneumology B of the Yaounde Jamot Hospital (YJH). Arbitrarily, observer designated by “1” was the one who had the best score for the detection and diagnosis in its category. The inter-observer agreement was calculated between readers 1 and 2 for each category, and between reader 1 of a giving category and reader 1 of the other category.

2.2. Selection of Chest Radiographs
Technically adequate frontal chest radiographs of patients 25 years old and above were selected in the department of pneumology A of the YJH, the highest referral and treatment center for respiratory diseases in Yaounde and its neighborhoods. All abnormal CXRs had a definitive diagnosis of concerned disease by suitable means (e.g. PTB confirmed by positive sputum smear).

Radiographs of patients with pulmonary tuberculosis (PTB) were chosen among the first 50 PTB patients hospitalized in the service during the year 2013. Normal CXRS or including diseases other than tuberculosis were consecutively selected from the files of patients treated in outpatient pneumonia. Informatics treatments were performed to cancel name of patients and date of examination on all the CXR images.
2.3. Reading of Radiographs

The consensus interpretation was obtained by reviewing all selected CXRs by a group consisting of one radiologist and one pneumologist (nine years experience each) and one senior resident in radiology. Members of this committee did not participate as readers to the study. Consensus interpretation determined for each case basic radiographic lesions and radiological diagnosis. The following were selected for analysis of concordance: two types of parenchymal lesions (16 cases of nodules and 12 cases of cavitary lesions), 6 CXRs with pleural effusion and 6 others with hilar or mediastinal adenomegaly, 23 cases with the diagnosis of PTB and 7 of lung cancer.

A total of 47 radiographs were selected for this study: 4 of diffuse infiltrative lung disease, 6 normal, 7 of lung cancer, 7 of bacterial pneumonia and 23 of pulmonary tuberculosis. Reading of CXRs by the six observers was performed on the same computer and using a report form with part for description of detected lesions and part for radiological diagnosis. Reading time was not limited, and each observer chose its convenient time to read the same 47 CXRs.

2.4. Data Collections and Analysis

The sample size [18]-[22] was calculated using the “KappaSize” R Version 2.13.0 statistical software. Based on expected Kappa of 0.47 ± 0.13 [8], and a type I error of 0.05, the minimum sample size was 46 radiographs for six observers and five diagnostic possibilities. The data were entered and analyzed using SPSS 17 software. These Kappa (K) intervals and thresholds [23] were used to measure inter-observer agreement: discordance (<0.0), low (0.0 - 0.20), poor (0.21 - 0.40), moderate (0.41 - 0.60), good (0.61 - 0.80), excellent (>0.81). For each observer, the score of detection of a giving lesion was the number of correctly detected lesion over the total number of that lesion detected during consensus reading.

This study was approved by the Ethics Committee of the Faculty of Medicine and Biomedical Sciences and the administrative authorities of the Yaounde Jamot Hospital.

3. Results

The most common lesion was pulmonary nodules (16/47). Figure 1 shows anomalies and diagnosis on which the Kappa coefficients of agreement were calculated.

The performances in detection and diagnosis of CXRs anomalies for each observer are shown in Table 1. The average score of correct results was 42.3% with variable proportions between different observers, and for the same reader from one lesion or diagnosis to another. The radiologist 1 had the highest average score of correct results (53.5%) with excellent detection of pleural effusions. The average score for the detection of caverns was the highest (58.3%). Pneumologists had the best proportions of correct diagnosis of tuberculosis (69.6% and 73.9%).

Kappa coefficients of inter-observer agreement in the detection of elementary lesions between different observers are shown in Table 2.

The highest Kappa coefficient was found in the agreement between the residents for the detection of pleural effusions (k = 0.73) and between radiologist and pneumologists in the detection of nodules (k = 0.74). Observers were more in agreement for the detection of nodules and adenomegalies than for the detection of caverns and pleural effusions, with frequent disagreement in the detection of pleural effusions. The inter-observer agreement...
Table 1. Proportion of correct detected lesion and correct diagnosis for each observer.

<table>
<thead>
<tr>
<th>Observers</th>
<th>ADP (n = 6)</th>
<th>Pleural eff. (n = 6)</th>
<th>Cavern* (n = 12)</th>
<th>Micronodule (n = 16)</th>
<th>Lung cancer (n = 7)</th>
<th>PTB (n = 23)</th>
<th>Mean score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resident 1</td>
<td>16.7%</td>
<td>50.0%</td>
<td>58.3%</td>
<td>12.5%</td>
<td>57.1%</td>
<td>39.1%</td>
<td>38.9%</td>
</tr>
<tr>
<td>Resident 2</td>
<td>16.7%</td>
<td>33.3%</td>
<td>58.3%</td>
<td>6.3%</td>
<td>57.1%</td>
<td>34.8%</td>
<td>34.4%</td>
</tr>
<tr>
<td>Radiologist 1</td>
<td>50.0%</td>
<td>100%</td>
<td>50.0%</td>
<td>43.7%</td>
<td>42.8%</td>
<td>34.8%</td>
<td>53.5%</td>
</tr>
<tr>
<td>Radiologist 2</td>
<td>16.7%</td>
<td>16.7%</td>
<td>66.7%</td>
<td>62.5%</td>
<td>14.3%</td>
<td>43.5%</td>
<td>36.7%</td>
</tr>
<tr>
<td>Pneumologist 1</td>
<td>33.3%</td>
<td>50.0%</td>
<td>41.7%</td>
<td>56.2%</td>
<td>57.1%</td>
<td>73.9%</td>
<td>52.0%</td>
</tr>
<tr>
<td>Pneumologist 2</td>
<td>16.7%</td>
<td>0%</td>
<td>75.0%</td>
<td>37%</td>
<td>28.6%</td>
<td>69.6%</td>
<td>37.9%</td>
</tr>
<tr>
<td>Mean score</td>
<td>25.0%</td>
<td>41.7%</td>
<td>58.3%</td>
<td>36.4%</td>
<td>42.8%</td>
<td>49.3%</td>
<td>42.3%</td>
</tr>
</tbody>
</table>

Table 2. Kappa coefficient (CI 95%) of inter-observer agreement in the detection of elementary lesions.

<table>
<thead>
<tr>
<th>Observers</th>
<th>Adenomegaly</th>
<th>Cavern</th>
<th>Pleural effusion</th>
<th>Micronodule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resident 1/resident 2</td>
<td>0.69 (0.36 - 1.02)</td>
<td>0.26 (−0.08 - 0.60)</td>
<td>0.73 (0.37 - 1.09)</td>
<td>0.69 (0.47 - 0.91)</td>
</tr>
<tr>
<td>Radiologist 1/radiologist 2</td>
<td>0.55 (0.20 - 0.89)</td>
<td>0.25 (−0.09 - 0.59)</td>
<td>−0.08 (−0.14 - −0.02)</td>
<td>0.32 (0.01 - 0.64)</td>
</tr>
<tr>
<td>Pneumologist 1/pneumologist 2</td>
<td>0.61 (0.30 - 0.91)</td>
<td>0.45 (0.10 - 0.79)</td>
<td>0.34 (−0.05 - 0.73)</td>
<td>0.53 (0.25 - 0.81)</td>
</tr>
<tr>
<td>Resident 1/radiologist 1</td>
<td>0.43 (0.05 - 0.80)</td>
<td>0.09 (−0.22 - 0.40)</td>
<td>−0.08 (−0.14 - −0.02)</td>
<td>0.64 (0.40 - 0.87)</td>
</tr>
<tr>
<td>Resident 1/pneumologist 1</td>
<td>0.69 (0.36 - 1.01)</td>
<td>0.50 (0.14 - 0.85)</td>
<td>0.23 (−0.22 - 0.68)</td>
<td>0.47 (0.14 - 0.68)</td>
</tr>
<tr>
<td>Radiologist 1/pneumologist 1</td>
<td>0.55 (0.20 - 0.89)</td>
<td>0.25 (−0.09 - 0.59)</td>
<td>−0.07 (−0.12 - −0.01)</td>
<td>0.74 (0.41 - 0.98)</td>
</tr>
</tbody>
</table>

ranged from poor to good ($k = 0.32$ to 0.74) for the detection of nodules, and moderate to good ($k = 0.43$ to 0.69) for the detection of adenomegalies.

Kappa coefficients of inter-observer agreement in the diagnosis of pulmonary tuberculosis and bronchopulmonary cancer among readers are shown in Table 3.

The agreement for the diagnosis of tuberculosis was higher among pneumologists ($k = 0.71$). Observers were more consistent for the diagnosis of cancer than for that of tuberculosis. The inter-observer agreement was excellent ($k = 1$) between resident and pneumologist for the diagnosis of lung cancer and good between pneumologists ($k = 0.71$) for the diagnosis of tuberculosis.

4. Discussion

This study shows that the agreement between observers varies with the type of lesion and diagnosis. Observers were more in agreement for the detection of nodules and adenomegalies (ADP). Disagreement was most frequent regarding the detection of pleural effusions. Observers more agreed for the diagnosis of cancer and for that of tuberculosis.

Cascade et al. [6] in a study of competence of chest and nonchest radiologists in interpreting chest radiographs founded no difference in clinically important missed diagnoses among chest radiologists, but a statistically significantly higher rate of seemingly obvious misdiagnoses for nonchest specialty radiologists. When evaluating the reliability and validity of chest radiographs in the diagnosis of tuberculosis by 25 physicians of varying qualifications, Kumar et al. [15] in Nepal founded the overall sensitivity and specificity of CXR of 78% and 51% respectively, and a poor agreement between the best physician and the best radiologist. They concluded of an unsatisfactory sensitivity and specificity of chest x-rays in the diagnosis of pulmonary tuberculosis.

In establishing the performance of chest X-ray (CXR) in all suspects of tuberculosis (TB), a study by van Cleeff et al. [14] showed 89% agreement ($K = 0.75$) for the combined scores “TB” or “no-TB”.
Table 3. Kappa coefficient (CI 95%) of inter-observer in the diagnosis of tuberculosis and lung cancer.

<table>
<thead>
<tr>
<th>Observers</th>
<th>Pulmonary tuberculosis</th>
<th>Lung cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resident 1/resident 2</td>
<td>0.60 (0.25 - 0.94)</td>
<td>0.72 (0.23 - 1.20)</td>
</tr>
<tr>
<td>Radiologist 1/radiologist 2</td>
<td>0.57 (0.24 - 0.89)</td>
<td>0.66 (0.13 - 1.18)</td>
</tr>
<tr>
<td>Pneumologist 1/pneumologist 2</td>
<td>0.71 (0.40 - 1.01)</td>
<td>0.22 (~0.197 - 0.63)</td>
</tr>
<tr>
<td>Resident 1/radiologist 1</td>
<td>0.25 (~0.15 - 0.65)</td>
<td>0.72 (0.23 - 1.2)</td>
</tr>
<tr>
<td>Resident 1/pneumologist 1</td>
<td>0.34 (0.08 - 0.59)</td>
<td>1</td>
</tr>
<tr>
<td>Radiologist 1/pneumologist 1</td>
<td>0.19 (~0.08 - 0.46)</td>
<td>0.72 (0.23 - 1.20)</td>
</tr>
</tbody>
</table>

For the detection of ADP, inter-observer agreement was good (K = 0.61) between pneumologists and moderate (K = 0.55) between radiologists. In a similar study conducted in 2010 in London, Abubakar et al. [24] reported a poor agreement between both pneumologists and radiologists. The frequent association of ADP with pulmonary tuberculosis and the high prevalence of tuberculosis in our environment could explain this difference.

The average score for the detection of caverns was the highest (58.3%). This can be justified by the fact that in an endemic TB area, our observers are used to see the caverns that are very common in pulmonary tuberculosis in the tropics [16] [17] [25] [26]. However, the agreement in the detection of caverns ranged from poor to moderate (K = 0.25 to 0.50). It could be that the observers were in disagreement on some caverns. In addition, the detection of these depends on their location, size, and content, and wall thickness. Balabanova et al. [10] had found the similar inter-observer agreement between pneumologists and radiologists in Russia.

Disagreements were common in the detection of pleural effusions in contrast to the results from Shinsaku et al. [5] in Japan and Abubakar et al. [24] in London who found a moderate to excellent level of agreement. The predominance of low abundance pleural effusions in our sample (n = 5/6) and their association with other abnormalities could explain these disconcordances.

Radiologists had a poor agreement in the detection of nodules (K = 0.32). Dawson et al. [27] founded similar results (K = 0.32) between radiologists in South Africa. Anna Ralph et al. [28] observed low concordance (K = 0.12) between radiologists in Australia. In our study, observers were generally more consistent for the detection of nodules than for the detection of caverns even if individual performances were better in the detection of caverns. Both lesions are very common in pulmonary tuberculosis [25] [26].

Pneumologists had good agreement (K = 0.71) in the diagnosis of tuberculosis, while radiologists had moderate agreement (K = 0.57). This difference could be explained by the experience of pneumologists (6 and 14 compared to 2 and 6 years for radiologists) but also by their specific activities allowing them more experience in PTB patients. Indeed, in our setting, almost all TB patients converge to pneumologists who would on average seen much more chest radiographs of tuberculosis than radiologists. Dawson et al. reported a good agreement between radiologists in South Africa in 2010 [27]. Abubakar et al. [24] observed a moderate agreement for both radiologists and pneumologists. Balabanova et al. [10] had reported a poor agreement between pneumologists and moderate between radiologists. However, in our study, the pneumologists because of their specific activities were of risk of “hindsight bias”, especially since there were 23/47 cases of PTB. The two pneumologists were from the department B while the cases where selected in the department A to avoid “memory bias”.

In general, our observers were more consistent for the diagnosis of lung cancer than for that of tuberculosis. This could be justified by evocative radiological aspect of selected cases of cancer, and the highly variable radiological presentation of pulmonary tuberculosis.

In this study, the total number of cases was relatively low even if the distribution of anomalies and diseases, as well as observers recalled the normal daily exercise. This is to our knowledge, one of the few studies of inter-reader agreement in the interpretation of CXRs in Africa. It backs in the saddle the difficulties of interpretation of current imaging test regardless of experience and qualification of the observer, and it encourages more interdisciplinary collaboration.

The limitations of this study are the lack of use of a standardized interpretation grid and variable experience of readers (years of practice).
5. Conclusion

The inter-observer agreement varies with the type of lesion and diagnosis. Pneumologists were most effective for the diagnosis of tuberculosis. Observers were more in accord for the detection of nodules and the diagnosis of cancer than for the detection of pleural effusions and diagnosis of tuberculosis. The use of a standardized interpretation scheme is recommended to improve detection and reading concordance between different observers.

Declaration of Interest

The authors declare to have no competing interest in relation to this article.

References


http://apps.who.int/iris/bitstream/10665/75938/1/9789241564502_eng.pdf


