A Study of Variability in the Interpretation of Chest X-rays in the Detection of Lung Cancer

ABRAHAM M. LILIENTHAL AND BERNARD KORDAN

Department of Chronic Diseases, Johns Hopkins University School of Hygiene and Public Health, Baltimore, Maryland, and the Department of Radiology, Loma Linda University School of Medicine, Los Angeles, California

Summary

As part of the American Cancer Society-VA Cooperative Pilot Study for the evaluation of radiologic and sputum cytologic screening in the early detection of lung cancer, it was possible to obtain some estimate of the variability in the interpretation of chest X-rays by radiologists. With respect to interindividual variability, the positive readings of one radiologist will be called positive by another radiologist in about 50-60% of instances and the negative readings in about 90-95% of instances. Similar degrees of agreement were found when a radiologist read a series of X-rays twice. This degree of variability is not unexpected and indicates the need for having 2 radiologists read X-rays independently in surveys in order to maximize the yield of suspect cases.

The American Cancer Society-VA Cooperative Pilot Study was initiated in 1958 in order to compare the relative efficacy of semiannual sputum cytologic and radiologic screening in the early detection of lung cancer and to determine whether they would be practical for mass screening of a nonpatient population group. The general plan of the study was for 6 VA domiciliary establishments to take chest X-ray films and to collect sputa semiannually from all male members of the domiciles. Two posterior-anterior films (14 x 17 inches) in full respiration with a tube shift between exposures were taken on each participant for each screening examination. All X-rays were subjected to 3 reading; after first being read in the domicile, they were read independently by 2 radiologists in the radiology center. Slides were prepared from sputum specimens and sent for screening to 1 of 4 participating cytology centers. Screening was done by techni-

1 The abbreviations used are: SN, suspect neoplasm; OSPA, other significant pulmonary abnormality; CV, cardiovascular abnormality; NSA, nonsignificant abnormality; NEG, negative; TU, technically unsatisfactory.

Initially, no provision was made on the cards reporting the X-ray screening findings to identify the individual radiologist making the report. However, since each radiologist completed and sent in an individual report card for each of his readings, it was possible to compare these 2 reports without individual identification. After the study had been in progress, it became evident from a preliminary analysis that there were differences in interpretation between individual radiologists and that there was and would continue to be a turnover of the radiologists reading the X-rays. It was, therefore, considered desirable to identify the individual radiologist by a code number on the report card.

The procedure described permitted an analysis of variability between individual radiologists. However, due to chance and accidental circumstances, a certain small number of films were read twice independently by the same reader, allowing an analysis of variability within individuals. In presenting the analysis of variability, we shall first present the results with respect to interindividual variability and then intraindividual variability.

The comparisons of the readings of the 2 radiologists during the early phase of the study, when they were not individually identified, are presented in Table 1; clearly, variability of interpretation was present. If the technically unsatisfactory readings are disregarded, the % of times in which both readers agreed is 67%. From the viewpoint of screening for abnormal pulmonary findings, it would seem reasonable to group the SN1 and OSPA categories into 1 and the other categories into another, thereby making 2 groups. The results, when this is done, are presented in...
Comparisons of Interpretations of Chest X-ray Films by 2 Radiologists in Early Phase of Study When Radiologists Were Not Identified

<table>
<thead>
<tr>
<th>Chest X-ray Interpretation by Another Radiologist</th>
<th>No. of individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>120</td>
</tr>
<tr>
<td>SN and OSPA</td>
<td>870</td>
</tr>
<tr>
<td>CV, NSA, and NEG</td>
<td>1260</td>
</tr>
</tbody>
</table>

For abbreviations, see footnote to Table 1.

Underlined numbers represent agreement by both radiologists.

\[
\text{% Agreement} = \frac{\text{Total of numbers in diagonals (underlined)}}{\text{Total No. of satisfactory films}}
\]

\[
\text{% Agreement} = \frac{45 + 750 + 301 + 249 + 2703}{6135} = 67.0
\]

### Table 3

Summaries of Comparisons of Readings by 2 Radiologists Using Various Measures of Agreement

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>No. of pairs of films</th>
<th>5 Categories</th>
<th>2 Categories</th>
<th>Proportion of readings read “positive” by one reader compared to the other</th>
<th>Proportion of readings read “negative” by one reader compared to the other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>1. One reader vs. other reader</td>
<td>6,135</td>
<td>67.0</td>
<td>85.1</td>
<td>69</td>
<td>62</td>
</tr>
<tr>
<td>2. A vs. B</td>
<td>19,795</td>
<td>65.1</td>
<td>89.4</td>
<td>63</td>
<td>65</td>
</tr>
<tr>
<td>3. A vs. I</td>
<td>29</td>
<td>58.6</td>
<td>82.8</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>4. A vs. C</td>
<td>1,295</td>
<td>62.2</td>
<td>85.1</td>
<td>82</td>
<td>48</td>
</tr>
<tr>
<td>5. A vs. D</td>
<td>558</td>
<td>59.5</td>
<td>86.9</td>
<td>24</td>
<td>49</td>
</tr>
<tr>
<td>6. A vs. E</td>
<td>8,183</td>
<td>59.0</td>
<td>89.2</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td>7. A vs. F</td>
<td>310</td>
<td>31.6</td>
<td>87.1</td>
<td>45</td>
<td>35</td>
</tr>
<tr>
<td>8. A vs. G</td>
<td>1,438</td>
<td>55.2</td>
<td>90.5</td>
<td>68</td>
<td>54</td>
</tr>
<tr>
<td>9. A vs. H</td>
<td>521</td>
<td>55.9</td>
<td>89.4</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>10. B vs. I</td>
<td>213</td>
<td>76.1</td>
<td>89.2</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>11. B vs. C</td>
<td>388</td>
<td>60.3</td>
<td>80.9</td>
<td>90</td>
<td>33</td>
</tr>
<tr>
<td>12. C vs. H</td>
<td>256</td>
<td>68.8</td>
<td>91.8</td>
<td>57</td>
<td>64</td>
</tr>
<tr>
<td>13. A vs. A</td>
<td>69</td>
<td>55.1</td>
<td>91.3</td>
<td>66</td>
<td>50</td>
</tr>
<tr>
<td>14. B vs. B</td>
<td>26</td>
<td>46.2</td>
<td>80.8</td>
<td>66</td>
<td>33</td>
</tr>
</tbody>
</table>

* First reader refers to 1st reader mentioned in Column 1 and likewise with respect to 2nd reader.

* Pos., positive; Neg., negative.

Table 2. In this instance, as expected, the % agreement is increased to 85.1%, which seems quite high.

This high frequency of agreement partially results from the fact that a large proportion of the readings are in the “negative” (CV, NSA, and NEG) group, and agreement with respect to...
Variability in Interpretation of X-rays

TABLE 4
Comparison of Interpretation of Chest X-ray Films by Radiologists A and B

<table>
<thead>
<tr>
<th>CHEST X-RAY INTERPRETATION BY RADIOLOGIST A (No. of individuals)</th>
<th>CHEST X-RAY INTERPRETATION BY RADIOLOGIST B (No. of individuals)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>OSPA</td>
<td>CV</td>
</tr>
<tr>
<td>61</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>70</td>
<td>1320</td>
<td>63</td>
</tr>
<tr>
<td>19</td>
<td>151</td>
<td>1322</td>
</tr>
<tr>
<td>25</td>
<td>407</td>
<td>43</td>
</tr>
<tr>
<td>28</td>
<td>157</td>
<td>91</td>
</tr>
<tr>
<td>TU</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>203</td>
<td>2053</td>
</tr>
</tbody>
</table>

* Abbreviations as in footnote to Table 1.
* Underlined numbers represent agreement by both radiologists.

% Agreement = Total of numbers in diagonals (underlined) / Total No. of satisfactory films

% Agreement = 61 + 1320 + 1322 + 1716 + 8475 = 12,894 / 20,033 = 65.1

% Agreement = Total of numbers in diagonals (underlined) / Total No. of satisfactory films

% Agreement = 1467 + 16,232 = 17,699 / 19,795 = 89.4

negative readings contributes considerably to the total % agreement. Another and perhaps preferable way of measuring variability is to determine what is the % of positive readings by 1 radiologist that are also read positive by the other radiologist and similarly for negative readings. For example, if one compares the readings of Radiologists A and B, the following percentages can be determined:

(1) No. of positive readings by both Radiologist A and B / No. of positive readings by Radiologist A

(2) No. of positive readings by both Radiologist A and B / No. of positive readings by Radiologist B

(3) No. of negative readings by both Radiologist A and B / No. of negative readings by Radiologist A

(4) No. of negative readings by both Radiologist A and B / No. of negative readings by Radiologist B

These were computed for the comparisons made in Table 2 and they are presented as the first comparison in Table 3 in Columns 5-8. We note that for this measure of variability there is good agreement with respect to a negative reading—of the order of 89-92%—but a moderate degree of agreement with respect to positive readings—of the order of 62-69%.

Similar types of comparisons were made for those readings when the radiologists were identified. A specific example is presented in Tables 3 and 4. These individual comparisons are also summarized in Table 3. On the whole, the degree of agreement is quite similar for each of these comparisons, with fairly good agreement with regard to negative readings and only a moderate degree of agreement with respect to positive readings.

Estimates of intraindividual variability are summarized as comparisons 13 and 14 in Table 3; they represent only a small number of readings and are limited to 2 radiologists. It is noteworthy that the degree of agreement of 1 reader with himself is no greater than when he is compared with another radiologist.

The extent of both interindividual and intraindividual variability was not unexpected, as was mentioned earlier. It appears to be of approximately the same order of magnitude reported in other studies of variability in the interpretation of chest X-ray films. These results clearly justify the original decision of having 2 radiologists read the X-rays independently and to provide diagnostic follow-up for those individuals whose X-rays were regarded as being suspicious by either of 2 radiologists.

Acknowledgments

This study was supported in part by grants (Grant No. L-31) from the American Cancer Society, the Veterans Administration, and by a USPHS career program award (K6-GM-13,901) from the National Institute of General Medical Sciences.

Reference

A Study of Variability in the Interpretation of Chest X-rays in the Detection of Lung Cancer

Abraham M. Lilienfeld and Bernard Kordan