Implications of Mathematical Misconception in Medical Auditing: The common errors of using grouped data in medical audit of screening mammography

WEI-TENG WANG1,2  HUAY-BEN PAN1,2  HUEI-LUNG LIANG1,2  CHEN-PIN CHOU1,2  HUNG-JU LI3  KAM-FAI WONG3

Department of Radiology1, Kaohsiung Veterans General Hospital, Kaohsiung, Taiwan
National Yang-Ming University2, Taipei, Taiwan
Institute of Statistics3, National University of Kaohsiung, Kaohsiung, Taiwan

ABSTRACT

We reviewed the grouped data in medical audit of screening mammography in Taiwan in 2009. With numerical examples presented, we discussed the implications of our result for the calculation and interpretation of common misleading of using grouped data instead of individual data. We concluded that it is essential to monitor the audited data from an individual interpreter as the basic unit rather than to use the hospital as a representative grouped data unit. Strictly asking for the volume maintains at the certain level and flatting interpretative level of population-based screening mammography by education through a quality assurance may minimize the impact.

In the medical auditing, we usually use the grouped data of a hospital as the representative unit for the demonstration of screening mammographic performance [1]. As grouped data can provide more convenient and simple description, it is often adopted by government departments to inform the public by publishing the percentage. Many statisticians believe that the mainstream public should be informed that the counter-intuitive results are possible [2, 3]. Careful interpretation or analysis is necessary to avoid biased conclusion.

For a screening mammography, a medical audit is the most convincing way to demonstrate the successful practice [4]. The data of medical audit [5, 6] such as positive predictive value (PPV) or recall rate of each screening institute can be disclosed through Health Promotion Bureau’s publications so that the public can form an objective opinion on each hospital’s performance and to recognize overall performance of the program [7]. High PPV1 of screening mammography indicates a good performance of screening radiologists, whereas high recall rate may increase false positive cases, which will inevitably decrease PPV1, increase resource waste and patients’ anxiety [8]. The optimal recall rate of screening mammography has been debated with wide ranges (%-15%) reported in the literature [9, 10]. Since recall rate is a simple indicator for monitoring the quality of screening, screening hospitals are usually asked to comply with the recall rate below 10% under the recommendation of ACR (American College of Radiology) in a certain range of PPV1 (5-10%) [5].

The concept of average is based on homogeneity of distribution. Therefore some questions about mathematical misconception in medical auditing may appear. Can we conclude that an institute with a recall rate higher than 10% has poorer quality than those of recall rates below 10% even if both of them have similar positive predictive value 1 and cancer detection rate? In this paper we try to propose some real results to show that it is essential to monitor the audited data from an individual interpreter as the basic unit rather than to use the hospital as a representative grouped data unit.
MATERIALS AND METHODS

Data Sources
Data on participants in a population-based mammography screening program were obtained from the Bureau of Health Promotion, which coordinates cancer screening in Taiwan. In 2009, there were 245,250 consecutive mammography interpreted by 293 identified screening radiologists in 140 screening facilities in Taiwan. The data from three different hospitals were selected for comparison in this study.

The confidentiality of all research subjects had been protected according to the principles set forth in the Declaration of Helsinki.

Assessment, Definition and Reference standard (by ACR) [5, 6]
Each screening mammogram is categorized by using Breast Imaging Reporting and Data Systems (BI-RADS) derived from the ACR’s criteria. Mammograms assessed as BI-RADS 1, 2, or 3 are defined as negative and those assessed as BI-RADS 0, 4 and 5 are defined as positive.

A mammogram with positive findings are defined as true positive (TP) or false positive (FP); Negative findings are defined as true negative (TN) or false negative (FN) based on whether or not the screenee appeared on a list in the National Cancer Registry within the following 12 months of their examinations.

Recall rate is defined as the proportion of individual those assessed as BI-RADS 0, 4 and 5. In addition, recall rate can be rewritten as (TP + FP) / (TP + FP + TN + FN). Recall rate is suggested less than 10% by benchmark of ACR.

Figure 1. A histogram was adopted to present the frequency of hospital in different recall rate
The recall rates histogram of screening mammography was formed by 140-screening facilities, which examined 245,250 screenees in Taiwan in 2009. The overall recall rate was 9.6%. In the subgroup of 7–10%, that were grouped by 31 facilities including Hospital A and Hospital C. In the subgroup of 10–13%, that were grouped by 33 facilities including Hospital B.

Table 1. The grouped- and detailed data of the three sampled hospitals

<table>
<thead>
<tr>
<th>Institution</th>
<th>Recall rate (%) (Recall/Screenees)</th>
<th>PPV1 (%)</th>
<th>CDR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hospital</td>
<td>Radiologist</td>
<td>Hospital</td>
</tr>
<tr>
<td>A</td>
<td>10.0 (100/997)</td>
<td>7.86 (36/458)</td>
<td>5.00 (5/100)</td>
</tr>
<tr>
<td></td>
<td>*14.52 (35/241)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>11.2 (196/1752)</td>
<td>7.41 (6/81)</td>
<td>8.79 (83/944)</td>
</tr>
<tr>
<td>C</td>
<td>9.1 (186/2040)</td>
<td>4.74 (35/738)</td>
<td>8.36 (32/383)</td>
</tr>
</tbody>
</table>

Asterisk: recall rate > 10%
Positive predictive value 1 (PPV1) is defined as the proportion of positive screening mammograms (BI-RADS categories 0, 4, or 5), calculated as (TP) / (TP + FP). Reference standard of PPV1 is 5-10%.

Cancer detection rate (CDR) is defined as the number of breast cancer in initial positive assessments, calculated as (TP) / (TP + FP + TN + FN). CDR is 0.2-1% in benchmark.

By above definitions, PPV1 = Cancer detection rate (CDR) / Recall rate.

**Data analysis**

We grouped the recall rate of screening hospitals into subgroups by scaling the histogram with 3% increment in the intervals. Three hospitals with acceptable PPV1 (5-10%) and CDR (0.2-1%) were selected for demonstration. In addition, the three selected hospitals, respectively, contained at least 3 identified screening radiologists. The detailed data of selected hospitals were disclosed in Table 1.

**Statistical Analysis**

All statistical analyses were performed with a SPSS program for Windows (version 12.0). The proportion of recall rate, PPV1 and CDR were recorded to present the performance of the mammography screening. A histogram was adopted to present the frequency of hospital in different recall rates. Chi-square test was used to detect the difference of proportion between the two hospitals, and also to compare whether the recall rates of radiologists within a hospital were different. A p-value < 0.05 was considered significant.

**RESULT**

The overall recall rate, PPV1 and CDR of mammographic screening in Taiwan in 2009 was 9.60%, 4.04% and 0.44%, respectively. The histogram of recall rate of the 140 screening institutions with 3% increment in the intervals was listed as Figure 1.

We compared the grouped data of hospital-A, hospital-B and hospital-C in which three of them have recall rates closed to 10% with acceptable PPV1 (5-10%) and CDR (0.2-1%). In addition, there were no statistical different of the recall rate (p=0.11), the PPV1 (p=0.81) and the CDR (p=0.95), respectively, among the three hospitals.

Although there was no significant statistically difference in the recall rate, PPV1 and CDR among the three selected hospitals, if we use the grouped data of a hospital as the representative unit, we may consider that hospital-C had the best performance due to its lowest recall rate (9.1%), which is below the benchmark that ACR recommended (<10%). On the other hand, hospital-B had the worst performance due to its highest recall rate (11.2%), which is out of the ACR recommended benchmark.

However, inspecting more relevant information in table 1, one radiologist in hospital-C had very low recall rate (4.74%). This radiologist may have some false negative cases, which is most horrible in the screening. Except this member mentioned above, the remaining recall rate of hospital-C is 11.6% (recall/screenees = 151/1302), even more higher than that of hospital-B (11.2%) and hospital-A (10.0%). Hence we cannot consider that hospital-C had the best performance by only using the recall rate of hospital as a representative grouped data unit. Judging the individual interpreter as the basic unit in Table 1, the performance of hospital-B and hospital-A may be better than the performance of hospital-C.

Moreover, it were statistical significant that the recall rates of the radiologists within the hospitals were different (p=0.02 in hospital-A, p<0.01 in hospital-B and p<0.01 in hospital-C). This performance heterogeneity among the screening radiologists within the grouped data may confound the expressed data result.

**DISCUSSION**

Theoretically, PPV1 has been generally thought to represent the "necessary balance" between sensitivity and specificity. But if only the lesions with a high likelihood of being malignant are recalled, the high specificity could elevate PPV1 despite the low sensitivity. Recall rate or PPV1 should not be used as the sole indicator for practice optimization because it does not take into account the number of cancers that would be missed at that sensitivity. [11].

The grouped data in table 1 was the real data obtained from the website of population-based breast cancer screening mammography in Taiwan in 2009 [7]. According to the recommendations of ACR and bureau of health promotion of Taiwan, it seemed that hospital-C had the best performance in the three sampled hospitals. But looking into the details of grouped data and relevant information, the differences of the recall rate, PPV1 and CDR were of no statistical significance. Meanwhile, the widest performance heterogeneity among the screening radiologists occurred in hospital-C, thereafter, we could not make the conclusion that hospital-A or -B has less performance than that of hospital-C. Inspecting more relevant information in table 1, other than grouped data only, may reflect the real performance of each screening radiologists.

Specific interpretive goals and benchmarks of screening mammography could be set and developed within the radiological professional societies with oversight by Mammography Quality Standards Act (MQSA) of USA [12] and Taiwan [13]. It is essential to monitor the audited data from an individual interpreter rather than to use the hospital as a representative unit. For example, the recall rate of a sampled hospital with two mammography-screening radiologists were 20% and 0%, respectively, the overall recall rate (10%) of this hospital would be within an
recommended range (5-10%), thus confounding the interpretation of audited results. Indeed, in comparison of the data between different countries also have the same consideration [14]. For logistic reasons, a medical audit is an internal feedback for radiologists in mammography interpretation. Thus many mammography screening programs do not provide the details in a population-based auditing. However, at least the situation should be kept in mind when we interpret the results just in percentage. Surely, homogeneity of sample distribution i.e. flattening interpretative level of population-based screening mammography by education through a quality assurance may minimize the impact. From the viewpoint of auditing, considering the whole performance of medical audit would be more detached than just focus on a sole indicator.

The main limitation of our study is that the selected examples were simplistic. Although the data was from the real data of nationwide screening mammography performed in 2009. It was just preferred for the explanations of misconception not reflect in general.

We conclude that it is essential to monitor the audited data from an individual interpreter as the basic unit rather than to use the hospital as a representative grouped data unit. Strictly asking for the volume maintains at the certain level and flattening interpretative level of population-based screening mammography by education through a quality assurance may minimize the impact.

ACKNOWLEDGMENTS

This work was supported by a grant from the Bureau of Health Promotion (No. 9910027C), Department of Health, Executive Yuan, Taipei, Taiwan. We thank participating mammography facilities, and radiologists for the data they provided for this study.

REFERENCES

7. Website of Bureau of health promotion of Department of health of Taiwan. The cancer screening programme. Available at https://www.screening.tw/ accessibility verified July 30 2011