Quality control for histological grading in breast cancer: an Italian experience

Controllo di qualità sulla valutazione del grading istologico nel carcinoma mammario: un’esperienza italiana

ITALIAN NETWORK FOR QUALITY ASSURANCE OF TUMOUR BIOMARKERS (INQAT) GROUP*

Key words
Quality control • Histological grading • Breast cancer • Intraobserver reproducibility • Interobserver reproducibility

Parole chiave
Controllo di qualità • Grading istologico • Carcinoma mammario • Riproducibilità entro-osservatore • Riproducibilità tra osservatori

Summary
Nonostante i criteri diagnostici per la valutazione del grading istologico nel carcinoma mammario siano stati meglio definiti dal metodo Nottingham/Tenovus, risultati soddisfacenti in termini di riproducibilità sono ancora difficili da ottenere.
Scopo di questo studio è stato la validazione di un protocollo metodologico da utilizzarsi in programmi di valutazione esterna di qualità. Abbiamo quantificato la riproducibilità nella lettura del grading e delle sue componenti evidenziando un livello di accordo entro e tra laboratori non del tutto soddisfacente. Per quanto riguarda le categorie di grading, la classe G2 ha fornito il contributo minore all’accordo osservato. Fra le variabili componenti, la meno riproducibile è risultata il polimorfismo nucleare seguita da conta mitotica e formazione di tubuli. In conclusione, i nostri risultati suggeriscono la necessità di sviluppare programmi di controllo qualità come processi dinamici che possano aiutare a comprendere le ragioni di alcune prestazioni insoddisfacenti ed implementare le necessarie azioni correttive.

Introduction

One of the fundamental concepts of histopathology is that the morphological appearance of tumors may be correlated with the degree of malignancy. In particular for breast cancer, morphological assessment of the degree of differentiation, evaluated with the histological grading method, has been shown in several studies to provide useful prognostic information.1-8

Even though the introduction of the Nottingham/Tenovus classification has improved and better defined the criteria to assess the histological grading of breast carcinoma,9 good results in terms of reproducibility and reliability are hard to obtain.9 Nevertheless, histological grading of breast carcinoma is commonly used by clinicians as prognostic information on which to base therapeutic decisions.11-15. It is therefore important that a good level of intra- and interobserver reproducibility be sought in the determination of histological grade. Quality control programs can be considered a valid tool to reach this goal.16

We initiated an external quality control program in Italy, focused on the performance of a group of pathologists, to validate the development of an operative and methodological procedure to be possibly adopted by institutions and organizations engaged in external quality control programs.

A quality control program requires a good level of organization that can be achieved by the assignment of responsibilities for the planning and management of ea-
ch of its steps. To accomplish this, our program was conceived as an interactive relationship between two main types of expertise, the first represented by an experienced pathologist responsible for all the bioclinical aspects related to histological grading in breast cancer and the second by a group of biostatisticians responsible for study design and statistical analysis. Once the general purposes of the project had been established, we decided to evaluate the reproducibility of the analytical determination of histological grade. In addition, we investigated the influence of the three component variables (nuclear pleomorphism, tubule formation and mitotic count) on the assessment of histological grade. The present paper refers to the Quality Control Program performed in Italy in 2003.

**Study features**

We documented and formalized the purposes and methods of the quality control program in a manual that was approved by all participants before the start of the study. This document included the rationale of the study, a standard procedure for evaluation of the slides, and detailed instructions for each step of the study. Careful and functional management of the practical aspects is as important as appropriate study planning. For this reason we designed the quality control program as a dynamic feedback process built on the continuous interaction between participants and investigators. The data produced by each participant were returned to the Biostatistics Unit where they were entered into the database and periodically checked for completeness and conformity to the measurement scales. After the database had been closed, the results of the statistical analyses were reported to the participants. Such pure reporting of the data, together with a lab-specific form including the results provided by each pathologist, allowed the participants to compare their own activity with that of the others and possibly improve their performance. Thirteen pathologists, listed in the appendix, participated in the project; they pertained to the three major types of institutions where histological grading is assessed (research institutions, university departments and hospitals). Each participant provided 20 slides from their own practice, which were pooled; from this pool we selected 20 master slides and 10 reserve slides to be used in case of damage. Selection of the slides was based on stratified randomization so that the number of slides of each grading class, G1, G2 and G3, reflected the proportion of each class found in routine practice. Since, as known, the diagnostic reproducibility is heavily affected by the quality standards of slides, the selected specimens were subjected to the attentive approval of three “assessors” among the participating pathologists, who were unanimously put forward for this task by the participants. Only those slides judged as well manufactured were added to the pool to be rounded among the participating pathologists. We organized two rounds of distribution of the 20 selected slides so that participants could examine the same slides on two independent occasions with an average interval of five months between the two. A new code was attributed to the slide at the second round to make the determination blind.

**Statistical methods**

The following aspects of reproducibility were investigated:
(i) intra-observer reproducibility;
(ii) reproducibility between each observer and the reference standard;
(iii) interobserver reproducibility;
(iv) contribution of each category to the observed reproducibility.

As the classification criteria adopted in slide assessment for grading and its component variables involve a categorical-ordinal scale, the reproducibility related to (i), (ii) and (iii) was evaluated by computation of the weighted kappa statistic ($K_{w}$). This makes it possible to adjust the observed agreement for chance by making allowance for the relative seriousness of disagreement (i.e. the distance between the categories). $K_{w}$ values lie between zero (absence of agreement) and 1 (absolute agreement). As previously, the observed values of $K_{w}$ were considered satisfactory if equal to or greater than 0.80. As regards (iv), kappa category-specific statistics ($K_{cs}$) and their weighted averages (Cohen’s kappa statistic, $K_{c}$) were estimated by jointly considering all participants.

<table>
<thead>
<tr>
<th>Kappa value (range)</th>
<th>Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.00</td>
<td>Disagreement</td>
</tr>
<tr>
<td>0.00-0.20</td>
<td>Slight agreement</td>
</tr>
<tr>
<td>0.21-0.40</td>
<td>Fair agreement</td>
</tr>
<tr>
<td>0.41-0.60</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.61-0.80</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>0.81-1.00</td>
<td>Almost perfect agreement</td>
</tr>
</tbody>
</table>
QUALITY CONTROL FOR HISTOLOGICAL GRADING IN BREAST CANCER

Each $K_w$ value was interpreted in a qualitative manner on the basis of the Landis and Koch classification criteria \(^{20}\) (Tab. I).

**Results**

**GRADING**

(i) Intra-observer reproducibility
A median $K_w$ value of 0.60 (range: 0.38-0.96) was found for the intra-observer reproducibility. Figure 1 illustrates the concordance pattern observed. In this graph all the $K_w$ values are plotted as blank diamonds; of the two horizontal lines the upper corresponds to the satisfactory $K_w$ threshold ($K_w \geq 0.80$) and the lower to the median $K_w$ value. Only three observers reached a satisfactory $K_w$ and only one was near satisfactory ($K_w = 0.73$).

(ii) Comparison with the reference values
The results of the comparison with the reference values are represented in Figure 2 for the first determination (panel A) and the second determination (panel B). In this figure all the $K_w$ values are plotted as black diamonds; also here the two horizontal lines refer to the satisfactory threshold and the median value.

Concordance analysis showed an unsatisfactory level of reproducibility with a median $K_w$ of 0.43 (range: 0.22-0.64) for the first determination and a median $K_w$ of 0.46 (range: 0.24-0.72) for the second determination. Between the first and second determinations increased $K_w$ values were observed for seven laboratories whereas for two laboratories the values decreased.

(iii) Interobserver reproducibility
Table II reports the $K_w$ values related to all possible pairwise comparisons (first determination upper half, second determination lower half). A rather unsatisfactory level of interobserver reproducibility was found for most observers with a median $K_w$ of 0.44 (range:

<p>| Tab. II. $K_w$ values of all pairwise comparisons: first determination (upper part) and second determination (lower part) – GRADING. |
|---|---|---|---|---|---|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>Observer</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.21</td>
<td>0.38</td>
<td>0.38</td>
<td>0.26</td>
<td>0.26</td>
<td>0.80</td>
<td>0.45</td>
<td>0.36</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.44</td>
<td>1.00</td>
<td>0.22</td>
<td>0.35</td>
<td>0.38</td>
<td>0.77</td>
<td>0.60</td>
<td>0.25</td>
<td>0.42</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>0.41</td>
<td>0.65</td>
<td>1.00</td>
<td>0.40</td>
<td>0.59</td>
<td>0.58</td>
<td>0.62</td>
<td>0.79</td>
<td>0.45</td>
<td>0.52</td>
</tr>
<tr>
<td>4</td>
<td>0.11</td>
<td>0.54</td>
<td>0.39</td>
<td>1.00</td>
<td>0.36</td>
<td>0.67</td>
<td>0.79</td>
<td>0.70</td>
<td>0.46</td>
<td>0.59</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>0.55</td>
<td>0.75</td>
<td>0.24</td>
<td>1.00</td>
<td>0.50</td>
<td>0.54</td>
<td>0.39</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.48</td>
<td>0.41</td>
<td>0.48</td>
<td>0.34</td>
<td>0.38</td>
<td>1.00</td>
<td>0.53</td>
<td>0.41</td>
<td>0.44</td>
<td>0.55</td>
</tr>
<tr>
<td>7</td>
<td>0.32</td>
<td>0.70</td>
<td>0.77</td>
<td>0.63</td>
<td>0.36</td>
<td>0.48</td>
<td>1.00</td>
<td>0.22</td>
<td>0.44</td>
<td>0.63</td>
</tr>
<tr>
<td>8</td>
<td>0.31</td>
<td>0.57</td>
<td>0.75</td>
<td>0.46</td>
<td>0.41</td>
<td>0.48</td>
<td>0.73</td>
<td>1.00</td>
<td>0.37</td>
<td>0.69</td>
</tr>
<tr>
<td>9</td>
<td>0.32</td>
<td>0.53</td>
<td>0.48</td>
<td>0.60</td>
<td>0.27</td>
<td>0.34</td>
<td>0.64</td>
<td>0.42</td>
<td>1.00</td>
<td>0.44</td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>0.34</td>
<td>0.34</td>
<td>0.44</td>
<td>0.42</td>
<td>0.48</td>
<td>0.33</td>
<td>0.32</td>
<td>0.46</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Fig. 1. Intra-observer reproducibility for Grading.

![Fig. 1](image1)

Fig. 2. Reproducibility between each observer and the reference standard for Grading.

![Fig. 2](image2)

![Tab. II](image3)
for the first determination and a median $K_w$ of 0.46 (range: 0.11-0.77) for the second.

(iv) Contribution of each category to the overall unweighted agreement

As can be seen in Table III, for both determinations the most important contribution to the overall unweighted agreement (corresponding to a $K_c$ of 0.29 and 0.27, respectively) is due to class G1, which shows moderate agreement with a $Kcs$ of 0.49 (first determination) and 0.43 (second determination). By contrast, the G2 class appears to provide the poorest contribution to the observed overall agreement.

**COMPONENT VARIABLES**

As reported in Table IV, when the role played by the three components of grading (tubule formation, nuclear pleomorphism and mitotic count) was investigated, a kind of hierarchy in reproducibility emerged with tubule formation in top position, mitotic count in the middle and nuclear pleomorphism at the bottom. This hierarchy was found for all three investigated aspects: intra-observer agreement, comparison with the reference values (Tab. IV) and interobserver agreement (data not shown). Finally, as reported in Table V, for each component the Score 1 class showed the most important contribution to the overall reproducibility, followed by Score 3. As expected, the most problematic class was Score 2.

**Discussion**

Morphological assessment of the degree of differentiation evaluated with the histological grading method has been shown to provide useful prognostic information in breast cancer $1^\text{-}8$, even when the diagnosis was made by different pathologists $11^\text{-}21$ and based on different criteria without well-defined guidelines $12$. Following the introduction of standardized grading criteria, partially based on semiquantitative evaluation, there has been a substantial improvement in the interobserver reproducibility related to the determination of this histological parameter $3$.

Histological grading is used in association with the conventional clinicopathological parameters (TNM classification) to define prognostic indexes that are widely accepted and utilized to identify different prognostic groups of breast cancer patients $11^\text{-}13$. However, the lack of good reproducibility among pathologists in the grading of breast cancer excludes any interchangeability of determinations made by different laboratories, consequently limiting the reliability and clinical usefulness of a parameter thought to be important for the management of breast cancer.

Quality control programs can be a valid tool to monitor the performance of pathologists and improve the reproducibility between laboratories. Our program was developed as a simple and useful proposal for institutions engaged in quality assessment of tumor biomarkers. A good quality control program for any biological marker should cover two main aspects: 1) it should monitor the actual practice of determinations throughout the country; 2) it should provide suggestions to help understand the reasons for the possible unsatisfactory performance of some laboratories and assistance in planning corrections. Whereas it is easy to find papers dealing with the first aspect, the second has been scarcely addressed. Our program was designed as an attempt to fill this gap.
As far as we know, our external quality control program is the first to be undertaken on a nationwide scale in Italy. Our findings, which appear to be in good agreement with those of some groups but not others, show a less than optimal level of agreement for both intra-observer (Figs. 1, 2) and interobserver reproducibility (Tab. II). However, it should be noted that the studies whose results are not in agreement with ours differ from ours in study planning, study design and number of pathologists involved. Although it has not yet been defined which level of reproducibility ought to be considered clinically "acceptable", we feel that our interobserver results do not appear to be satisfactory and further efforts are needed to better understand the reason for the unsatisfactory performance of some laboratories and to implement the necessary corrective actions.

In the light of our results, the steps we propose to be taken are the following:
1. very precise standardized protocols should be adopted;
2. workshops should be organized in which the problems of the individual laboratories are compared and discussed;
3. training sessions should be held to discuss and hopefully solve these problems and bring the performance into uniformity so that the results of different laboratories will become interchangeable;
4. quality control programs should be consistently implemented.

References

18 Italian Network for Quality Assurance of Tumor Biomarkers (INQAT) Group. Interobserver reproducibility of immunohistochemical HER-2/neu evaluation in human breast cancer: the real-
Appendix

* INQAT Group Co-workers and affiliations

Writing Committee (in alphabetical order)

M.E. Cortese (Biostatistics Unit) – Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano
E. Marubini (Biostatistics Unit) – Università di Milano, Milano
A. Paradiiso (Project Coordinating Unit) – Istituto Oncologico, Bari
J. Rosai (Reference Unit) – Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano
L. Saragoni (Reference Unit) – Ospedale “Morgagni-Pierantoni”, Forlì
P. Verderio (Biostatistics Unit) – Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano

Pathologists participating in the Grading Quality Control Program (in alphabetical order)

M. Aldi – Ospedale per gli Infermi, Faenza
L. Andreini – Imola AUSL, Imola
S. Bianchi – Università di Firenze, Firenze
P. Cossu Rocca – Università di Sassari, Sassari
G. De Rosa – Ospedale Mauriziano, Candiolo
C. Giardina – Università di Bari, Bari
G. Lanzanova – Ospedale Civile, Ravenna
F. Marandino – Istituto Regina Elena per lo Studio e la Cura dei Tumori, Roma
F. Marzullo – Istituto Oncologico, Bari
F. Nuzzo – Presidio Ospedaliero, Cesena
N. Ravarino – Ospedale Mauriziano, Turin
M. Ricci – Ospedale degli Infermi, Rimini
A. Zito – Istituto Oncologico, Bari
