Reproducibility of HPV DNA Testing by Hybrid Capture 2 in a Screening Setting

Intralaboratory and Interlaboratory Quality Control in Seven Laboratories Participating in the Same Clinical Trial

Francesca Maria Carozzi, PhD,1 Annarosa Del Mistro, MD,2 Massimo Confortini, PhD,1 Cristina Sani, MD,1 Donella Puliti, MD,1 Rossana Trevisan,2 Laura De Marco, PhD,3 Anna Gillio Tos, PhD,3 Salvatore Girlando, MD,4 Paolo Dalla Palma, PhD,4 Antonella Pellegrini, PhD,3 Maria Luisa Schiboni, PhD,5 Paola Crucitti, MD,6 Paola Pierotti, MD,6 Alberta Vignato, MD,7 and Guglielmo Ronco, PhD8

Key Words: Screening; Cervical cancer; Human papillomavirus; HPV; Hybrid Capture 2; HC2; Quality control; Reproducibility; Accuracy

DOI: 10.1309/84E5WHJQHK83BGQD

Abstract

Within a large Italian randomized trial on new technologies for cervical cancer screening involving 7 laboratories with different levels of experience, an intralaboratory and interlaboratory quality control program for human papillomavirus (HPV) DNA testing by Hybrid Capture 2 (HC2; Digene, Gaithersburg, MD) was implemented. To monitor the hybridization and detection steps, target samples containing purified, concentration-defined, HPV DNA were introduced in each test run. Only 3 of 1,024 showed a mistake in a positive vs negative classification with a 1 relative light unit (RLU)/positive control specimen (PC) ratio cutoff. To monitor the preanalytic steps (particularly denaturation), blinded specimens (33 collected in PreservCyt [Cytyc, Boxborough, MA] and 36 in Specimen Transport Medium [STM, Digene]) were centrally prepared, divided into aliquots, and sent to each laboratory. The multiple-rater κ scores for negative (<1 RLU/PC), low-positive (1 to <11 RLU/PC), and high-positive (≥11 RLU/PC) samples, respectively, were 0.91, 0.60, and 0.69 with PreservCyt and 0.93, 0.87, and 0.90 with STM. Our data showed high reliability and reproducibility with HC2, with κ values higher for STM than ThinPrep (Cytyc) samples.

Human papillomavirus (HPV) infection is the major cause of most cervical cancers and cervical intraepithelial neoplasia worldwide.1-3 Persistent HPV infection by high-risk types is a necessary (although not sufficient) cause for the development of cervical cancer. Consequently, there is strong interest in the use of HPV testing for cervical cancer screening.

By using Hybrid Capture (HC) Systems (Digene, Gaithersburg, MD) or different polymerase chain reaction (PCR) protocols, HPV testing has been evaluated in cervical screening as an adjunct to the Papanicolaou smear or alone as the only screening test. By combining the results of several studies,4 HPV testing has been shown to have higher sensitivity but lower specificity than conventional cytologic testing in detecting high-grade lesions; this is due mainly to the recognition of transient, clinically silent HPV infections as well. The most relevant result of these studies is the high negative predictive value (approaching 100%) attained by the combination of cytologic and HPV testing, whereas the major drawback of this modality is the increased number of women who need to be referred for colposcopy.

The HC2 HPV DNA test (Digene) is a commercially available microplate assay approved by the US Food and Drug Administration for the detection of HPV DNA and consists of a liquid-phase hybridization by the use of 2 RNA probe mixtures. The HC2 probe mix B detects a group of 13 cancer-associated HPV types: 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, and 68, whereas probe mix A detects low-risk types. Previous studies have demonstrated that the HC2 assay is sensitive and specific for detecting HPV DNA from cervical specimens,5 and it has good agreement with PCR assays.6,7

HC2 seems robust, but to consider HPV testing for cervical cancer screening programs, issues such as reproducibility...
between different testing centers need further attention. Indeed, to achieve quality reproducible results, it has been stated that “in the absence of a formal proficiency testing program, the laboratory must develop its own external program (eg, through interlaboratory exchange) or devise some system for internal proficiency testing.” Specimen processing includes critical steps, and lack of optimization potentially can give rise to false-positive results; thus, good reproducibility of blinded clinical samples in the laboratories might warrant the high reliability of trial results.

The reproducibility of HC2 depends on factors influencing the sensitivity and specificity of the reaction: collection, storage, transport and processing of biologic samples, target concentration, presence of reaction inhibitors, occurrence of nonspecific hybridization; and lack of optimization in one or more process phases. In particular, samples collected in fixative-containing solutions (which allow the preparation of cytologic slides) require an additional processing phase to be suitable for HC2 testing. This step, known as sample conversion and necessary to eliminate the fixative, is critical for optimal DNA denaturation and has not been evaluated extensively so far.

In the frame of a large, multicenter, randomized, controlled trial on New Technologies for Cervical Cancer (NTCC study) screening conducted in Italy, we implemented an intra-laboratory and interlaboratory quality control program to monitor reproducibility and accuracy in different steps of the assay in the 7 participating laboratories. The NTCC study involves about 100,000 women aged 25 to 60 years, coming for a new screening round within 9 organized screening programs, who were randomized in 2 arms for 2 periods (2 phases in which recruitment lasted about 1 year each).

During the first phase (HPV testing associated with cytologic testing), HPV testing was performed on samples collected in a fixative-containing transport medium, whereas in the second phase (HPV testing alone), the samples were collected in the Specimen Transport Medium (STM) provided by the HC2 manufacturer (Digene). We studied the performance of the HC2 test with both transport media.

Materials and Methods

The HPV testing was performed in 7 laboratories related to the screening programs and involved in the Italian trial. Four were diagnostic cytopathology laboratories with no experience in molecular virology tests; they are hereafter referred to as laboratory 1 through laboratory 7 in a randomized manner. A uniform and fully detailed protocol was prepared and used throughout the study in each laboratory.

Cervical specimens collected in PreservCyt (ThinPrep, Cytyc, Boxborough, MA) were processed for thin-layer slide preparation. The residual cell suspension was used to perform the HC2 HPV DNA test; samples containing less than 4 mL were considered inadequate to avoid false-negative results; 4 mL of each sample was processed with the Sample Conversion Kit (Digene) according to the manufacturer’s instructions.

Cervical specimens collected in STM were tested directly with HC2. HC2 was performed using only probe set B according to the manufacturer’s instructions. In October 2002, a modification in the protocol (ie, distribution of the probe mix after distribution of samples) was introduced in all laboratories as suggested by the manufacturer’s specialist.

Results were expressed as the ratio of light emission from a sample (relative light units [RLU]) to the average of 3 concurrently tested positive control specimens containing 1 pg/mL of HPV-16 DNA. The recommended positivity threshold of 1 RLU/PC (equivalent to 5,000 viral copies per test well) was used as a cutoff, and all samples with an RLU/PC ratio of 1 or more were considered positive.

Quality Control Protocols

Intralaboratory Quality Control

A panel of 6 synthetic target samples with a defined concentration of purified HPV DNA in a range from 0 to 150 pg/mL (HC2 HPV DNA test panel, Digene) was included in HC2 test runs as internal control samples. Each target had an expected value (RLU/PC) on the basis of HPV type and concentration [Table I]. In each laboratory, the targets were denatured at the same time, and 1-run aliquots were prepared and stored frozen at –20°C. In each test run, every laboratory introduced 2 target samples, 1 negative and 1 positive, for the high-risk probe set.

Interlaboratory Quality Control

During the first phase of the NTCC study, 4-mL aliquots of 3 blinded clinical PreservCyt samples were sent monthly from the laboratory in Florence, Italy, to the other centers. During the second phase of the NTCC study, 100-µL aliquots of 3 blinded clinical samples collected in STM were sent monthly from the laboratory in Florence to the other centers. The blinded samples were used as external control samples and processed in the same way as the NTCC study samples.

Reproducibility for Samples With RLU/PC Near the Cutoff

During the first phase of the NTCC study, samples collected in PreservCyt with an RLU/PC between 1.00 and 4.00 obtained during the period June through July 2002 were retested by using the residual aliquot of the conversion used for the initial testing and a newly converted aliquot. During the second phase of the NTCC study, samples collected in STM with an RLU/PC between 0.90 and 0.99 obtained during the entire second phase were retested in each center.
**Statistical Analyses**

Data from each laboratory were centralized and processed at the data center in Florence. Each laboratory provided the HPV DNA testing results for external control and DNA target samples, according to the HC2 protocol.

For external quality control, chance-adjusted agreement between laboratories was computed. Data were classified in 2 groups (positive and negative with a cutoff of 1 RLU/PC) and in 3 groups (negative, <1 RLU/PC; low-positive, 1 to <11 RLU/PC; and high-positive, ≥11 RLU/PC). With both categorizations, an unweighted $\kappa$ was computed for each laboratory by comparing its result with the most frequently reported result for each sample (majority result). In addition, the overall agreement for the 7 laboratories was assessed by computing $\kappa$ values for multiple-rater analysis.\(^{10}\) Conventionally, the $\kappa$ values 0.00 to 0.20, 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80, and 0.81 to 1.00 are considered indicative of slight, fair, moderate, substantial, and almost perfect agreement, respectively.\(^{11}\)

### Results

**Intralaboratory Quality Control**

A total of 1,024 HPV DNA targets were analyzed as internal control samples, 421 during the first phase and 603 during the second phase of the NTCC study. The number of targets tested by laboratories 1 through 7, respectively, was 214, 192, 101, 87, 69, 264, and 97. The number of targets tested by each laboratory was proportional to the number of samples processed in the trial.

The results from all laboratories provided an RLU/PC ratio out of the expected range in 19 (1.86%) of 1,024. Only 3 (0.29%) were truly the wrong result (positive vs negative or negative vs positive).

All erroneous results were obtained during the first phase in 3 laboratories. Laboratories 1 and 5 obtained a low-positive result (1.82 and 1.11, respectively) from targets 1 and 2, respectively, instead of a negative result; in laboratory 6, target 3 gave a negative result (0.19) instead of the expected positive result with a ratio of more than 2. Table 1 gives the mean and SD for each DNA target obtained in the 7 laboratories.

**Table 1** Intralaboratory Quality Control: Results of DNA Test Panel Targets

<table>
<thead>
<tr>
<th>Target</th>
<th>HPV Type/Concentration (pg/mL)</th>
<th>Probe Mix B</th>
<th>Expected Result (RLU/PC)</th>
<th>Laboratory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Negative/0 &lt;1</td>
<td>0.27 ± 0.09</td>
<td>0.21 ± 0.09</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6/5</td>
<td>0.28 ± 0.10</td>
<td>0.24 ± 0.10</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>16/5</td>
<td>3.87 ± 0.53</td>
<td>4.09 ± 0.56</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>18/20</td>
<td>12.89 ± 2.08</td>
<td>13.74 ± 2.05</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>43/50</td>
<td>0.25 ± 0.06</td>
<td>0.28 ± 0.12</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>56/150</td>
<td>106.00 ± 19.90</td>
<td>100.66 ± 20.00</td>
<td>6</td>
</tr>
</tbody>
</table>

HPV, human papillomavirus; PC, positive control specimen; RLU, relative light units.

* Data are given as the mean ± SD (95% confidence interval) in an RLU/PC ratio (see the text).

**Interlaboratory Quality Control**

During the first phase of the NTCC study, 33 blinded clinical PreservCyt samples were prepared; the number of external quality control samples tested by each laboratory was proportional to the number of clinical samples tested and to the time necessary for recruitment of the women. Laboratories 2 through 5 tested all 33 blinded samples, laboratory 1 tested 32, laboratory 3 tested 28, and laboratory 7 tested 31 samples.
26, laboratory 6 tested 30, and laboratory 7 tested 24. During the second phase of the NTCC study, 36 blinded samples collected in STM were prepared: laboratories 2, 3, 5, and 6 tested all 36 samples, laboratory 7 tested 15, and laboratories 1 and 4 tested 24.

Figure 2 and Figure 3 show the distributions of the results (expressed as logarithmic values of RLU/PC ratios) of external quality control samples constituted by 33 ThinPrep and 36 STM clinical samples in 7 laboratories. The κ values for each laboratory against the majority result, using a positive-negative classification, are given in Table 2. Values were always more than 0.90, except for laboratory 5, which showed a κ value of 0.88. When distinguishing between low- and high-positive HC2 results Table 3, κ values in general were still high, but values near 0.60 were observed for 3 laboratories with ThinPrep samples, whereas higher values were observed with STM samples. In a multiple-rater analysis Table 4, clearly higher κ values were obtained for low-positive and high-positive with STM than with ThinPrep samples.

Reproducibility of Results on Samples With an RLU/PC Near the Cutoff

We retested 82 PreservCyt samples showing RLU/PC values in the range between 1.00 and 4.00. Overall, by using the residual aliquots of the first conversion, a positive result was confirmed in 18% (9/51) of the clinical samples with RLU/PC values from 1.00 to 1.99, in 57% (8/14) of the samples with RLU/PC values from 2.00 to 2.99, and in 71% (12/17) of those with values between 3.00 and 4.00. The same results were obtained by using the residual aliquots of the first conversion or the newly converted aliquot (data not shown).

At retesting of the routine STM samples collected during the second NTCC phase and showing an RLU/PC result between 0.90 and 0.99, confirmation of a negative result was obtained in 45 (73%) of 62 specimens, whereas 17 samples (27%) gave a low-positive result.

Discussion

An intralaboratory and interlaboratory quality control program was implemented within a large multicenter randomized Italian trial on cervical screening, in which the HC2 kit was used by 7 laboratories to detect high-risk HPV DNA. The setting involved many laboratories with different previous experiences (in some cases no experience) processing a high number of samples. Therefore, it approached the conditions of a field application of HPV testing in mass screening for cervical cancer. Three different studies were performed, including controls aiming to monitor processing and testing steps. Overall, the data showed high reproducibility in all 7 sites.
The Intralaboratory Quality Control study was designed to monitor the accuracy of HC2 hybridization and detection steps in each laboratory. As internal control samples, purified HPV DNA samples of defined concentrations (provided as a commercial panel of 6 targets) were included in the test runs. More than 99% of the panel samples tested were within the expected range; only 3 of 1,024 gave a truly wrong result. These data indicate high accuracy. It is interesting that the most recent version of the HC2 kit contains an HPV high-risk quality control sample similar to HPV DNA target 3 used in our study.

This kind of quality control samples, however, being purified, cell free, and independent of the conversion step, cannot exactly mimic the clinical samples. Thus, the Intralaboratory Quality Control study was designed to also monitor the sample conversion and denaturation steps by using clinical specimens collected in PreservCyt or STM.

We found excellent agreement beyond chance for a positive-negative classification with PreservCyt and STM samples. We obtained overall $\kappa$ values of 0.95 for ThinPrep and 0.96 for STM samples. High $\kappa$ values also were obtained when the results were stratified as negative, low-positive, and high-positive. As shown in Table 4, $\kappa$ values were lower for PreservCyt than for STM control specimens and for the low-positive category (1 to <11 RLU/PC) than for the negative and high-positive sample. Indeed, STM control samples showed high reproducibility in every category ($\kappa$ values between 0.87 and 0.93), whereas the control samples collected in PreservCyt showed values of 0.60 and 0.69 for the low- and high-positive categories, respectively. In this regard, it must be taken into account that the circulated samples consisted of aliquots obtained by splitting, which likely could result in variability in cell numbers. Uneven distribution of target sequences is more critical for samples containing low amounts of HPV DNA, and the samples are likely to be more homogeneous with STM than with PreservCyt as a consequence of their composition and intended use. PreservCyt is aimed primarily at long-term preservation of cellular morphologic features, whereas STM is a specific DNA preservative.

Moreover, the conversion procedure, necessary to eliminate the fixative reagent contained in PreservCyt and allow complete DNA denaturation, is a critical step in sample processing because traces of undenatured DNA might give false-positive results. Indeed, since October 2002, even the HC2 manufacturer’s procedure has been modified (ie, distribution of probe mix after distribution of samples) to minimize this risk and optimize the denaturation step. Few data have been published regarding the reproducibility of the HC2 test and the influence of the collection medium type; in line with our data, Peyton et al. found a higher level of nonspecific detection for samples collected in PreservCyt than for those collected in STM.

The reproducibility for low-positive samples study was performed by retesting clinical specimens collected in PreservCyt with RLU/PC values in the range 1.00 to 4.00. In our experience, major reproducibility problems were observed for the specimens with RLU/PC values from 1.00 to 1.99; only 18% (9/51) of them were confirmed as positive. As a general rule, higher HC2 test results correspond to higher amounts of HPV DNA present in the specimen. The

---

**Table 3**

Intralaboratory Quality Control

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>HC2 Results (RLU/PC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>ThinPrep</td>
</tr>
<tr>
<td>1</td>
<td>0.92</td>
</tr>
<tr>
<td>2</td>
<td>0.93</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>0.93</td>
</tr>
<tr>
<td>5</td>
<td>0.93</td>
</tr>
<tr>
<td>6</td>
<td>0.93</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
</tr>
</tbody>
</table>

HC2, Hybrid Capture 2; PC, positive control specimen; RLU, relative light units.

$^*$ Unweighted $\kappa$ value of each laboratory in reference to the modal value (ie, the most frequently reported HC2 result) when the data were categorized as negative (<1 RLU/PC), low-positive (1 to <11 RLU/PC), and high-positive (≥11 RLU/PC [see the text]). For proprietary information, see the text.

**Table 4**

Intralaboratory Quality Control

<table>
<thead>
<tr>
<th>Hybrid Capture 2 Result</th>
<th>ThinPrep Samples</th>
<th>STM Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>0.91</td>
<td>0.93</td>
</tr>
<tr>
<td>Low-positive</td>
<td>0.60</td>
<td>0.87</td>
</tr>
<tr>
<td>High-positive</td>
<td>0.69</td>
<td>0.90</td>
</tr>
<tr>
<td>Overall</td>
<td>0.74</td>
<td>0.90</td>
</tr>
</tbody>
</table>

PC, positive control specimen; RLU, relative light units.

$^*$ $\kappa$ values for multiple raters. Data were categorized as negative (<1 RLU/PC), low-positive (1 to <11 RLU/PC), and high-positive (≥11 RLU/PC [see the text]). For proprietary information, see the text.
RLU values, however, can be influenced by other biologic factors, such as cross-hybridization with HPV types whose specific probes are not included in the mix and presence of more than 1 HPV type; in these circumstances, the hybridization efficiency might be less consistent.

A study by Castle et al, in which 1,072 specimens (collected in PreservCyt) initially tested in 4 clinical centers were retested by a quality control laboratory, found a k value of 0.84 for negative vs positive test results (92% in exact agreement). Discordant results were obtained in about 8% of the paired tests, and it is interesting that the vast majority of the results negative at retesting had an RLU/PC ratio between 1.00 and 3.00 in the first test, a result in line with our results and the observations by de Cremoux et al. In this latter study, of 1,075 PreservCyt specimens tested, 77 were considered false-positive results because they were negative at retesting; these had RLU/PC values between 1.00 and 3.00 at initial testing.

In a previous study, Castle et al tested 1,734 STM specimens twice with the HC2 test. Of these, 92 gave a negative result in one and a low-positive result in the other HC2 assay and were considered borderline; 68% of these specimens were HPV DNA negative by PCR with L1 consensus primers for the same types tested for by HC2, suggesting that most of these specimens might truly be negative for HPV DNA or have very low viral loads.

Random variation of RLU values is obviously more influential on the classification as positive or negative for values near the cutoff. The advisability of applying a different cutoff depends on its effect in terms of sensitivity and specificity for intraepithelial lesions. The NTCC study is expected to provide data about this. Another point to consider is that low reproducibility near the cutoff was found for the clinical samples but not for the standard DNA with a low concentration of purified HPV DNA (target 3). This suggests that clinical samples are better quality control samples than purified viral DNA.

Our study findings stress the importance of using standardized protocols and proficiency panels for the comparability of results among studies. We demonstrated high reliability of the HC2 assay; the use of a common and detailed protocol was instrumental in obtaining these excellent results. Efforts to minimize borderline results, however, are still needed to optimize the use of the HC2 test in cervical cancer screening programs.

References


Supported by the Italian Ministry of Health, Rome, Italy (Special Project: “Evaluation of New Technologies of Cervical Cancer Screening”); the European Union, Brussels, Belgium (Europe Against Cancer); and local public support (Piedmont Regional Government, Turin; “Maggiore” Hospital, Bologna; Veneto Regional Government, Venice; Trento Local Government, Emilia-Romagna Regional Government, Bologna).

Address reprint requests to Dr Carozzi: Centro per lo Studio e la prevenzione Oncologia, Viale Volta 171 50131 Florence, Italy.