JCM
Journals.ASM.org

Interlaboratory Reproducibility and Proficiency Testing within the Human Papillomavirus Cervical Cancer Screening Program in Catalonia, Spain

R. Ibáñez,[a] M. Félez-Sánchez,[a,b] J. M. Godínez,[b] C. Guardià,[c] E. Caballero,[d] R. Juve,[e] N. Combalia,[f] B. Bellosillo,[g] D. Cuevas,[h] J. Moreno-Crespi,[i] L. Pons,[j] J. Autonell,[k] C. Gutierrez,[l] J. Ordi,[m] S. de Sanjosé,[a,b,n] I. G. Bravo[a,b]

Infections and Cancer Unit, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain[a]; Infections and Cancer, Bellvitge Institute of Biomedical Research (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain[b]; Primary Care Clinical Laboratory Barcelonés Nord y Vallés Oriental, Catalan Institute of Health, Badalona, Spain[c]; Microbiology Department, University Hospital Vall d'Hebron, Catalan Institute of Health, Barcelona, Spain[d]; Primary Care Clinical Laboratory Bon Pastor, Catalan Institute of Health, Barcelona, Spain[e]; Pathology Department, UDIAT-CD, University Hospital Parc Tauli, Universitat Autònoma de Barcelona, Barcelona, Spain[f]; Pathology Department, Hospital del Mar, Barcelona, Spain[g]; Pathology and Molecular Genetics Department, University Hospital Arnau de Vilanova, Catalan Institute of Health, Lleida, Spain[h]; Pathology Department, University Hospital of Girona Dr, Josep Trueta, Infections and Cancer Unit, Catalan Institute of Oncology, Girona, Spain[i]; Pathology Department, IISPV, URV, Hospital de Tortosa Verge de la Cinta (HTVC), Catalan Institute of Health, Tortosa, Spain[j]; Pathology Department, Hospital Consortium of Vic, Vic, Barcelona, Spain[k]; Clinical Laboratory ICS Tarragona, Molecular Biology Section, University Hospital of Tarragona Joan XXIII, Catalan Institute of Health, Tarragona, Spain[l]; Pathology Department, CRESIB–Hospital Clínic, University Barcelona, Barcelona, Spain[m]; CIBER Epidemiology and Public Health, Barcelona, Spain[n]

In Catalonia, a screening protocol for cervical cancer, including human papillomavirus (HPV) DNA testing using the Digene Hybrid Capture 2 (HC2) assay, was implemented in 2006. In order to monitor interlaboratory reproducibility, a proficiency testing (PT) survey of the HPV samples was launched in 2008. The aim of this study was to explore the repeatability of the HC2 assay's performance. Participating laboratories provided 20 samples annually, 5 randomly chosen samples from each of the following relative light unit (RLU) intervals: <0.5, 0.5 to 0.99, 1 to 9.99, and ≥10. Kappa statistics were used to determine the agreement levels between the original and the PT readings. The nature and origin of the discrepant results were calculated by bootstrapping. A total of 946 specimens were retested. The kappa values were 0.91 for positive/negative categorical classification and 0.79 for the four RLU intervals studied. Sample retesting yielded systematically lower RLU values than the original test ($P < 0.005$), independently of the time elapsed between the two determinations (median, 53 days), possibly due to freeze-thaw cycles. The probability for a sample to show clinically discrepant results upon retesting was a function of the RLU value; samples with RLU values in the 0.5 to 5 interval showed 10.80% probability to yield discrepant results (95% confidence interval [CI], 7.86 to 14.33) compared to 0.85% probability for samples outside this interval (95% CI, 0.17 to 1.69). Globally, the HC2 assay shows high interlaboratory concordance. We have identified differential confidence thresholds and suggested the guidelines for interlaboratory PT in the future, as analytical quality assessment of HPV DNA detection remains a central component of the screening program for cervical cancer prevention.

Persistent infection with oncogenic human papillomaviruses (HPVs) is necessary for the development of invasive cervical cancer (CC) (1–3). CC screening based on cervical cytology has been instrumental in decreasing the incidence and mortality associated with CC in those countries with high screening coverage rates (4). However, the infectious etiology of this disease leads to the suggestion that the detection of the DNA of HPVs responsible for cellular transformation might provide a strong predictive marker for the early detection of women at risk (5). Worldwide studies have established that tests for the presence of HPV DNA show a higher sensitivity than cytology for the detection of high-grade cervical intraepithelial lesions (6). Despite this demonstrated higher sensitivity, the application to CC screening programs is still not widespread (6–10).

Currently, one of the most widely used tests for the detection of HPV DNA is the Digene Hybrid Capture 2 (HC2) assay (Qiagen, Gaithersburg, MD), approved by the U.S. Food and Drug Administration (FDA) in 2003 for use in clinical settings. Large prospective cohort studies and randomized controlled trials have proved that this assay has high clinical sensitivity (90 to 95%) for the detection of high-grade cervical intraepithelial neoplasia (CIN) (6). The HC2 procedure does not include a PCR amplification step and uses a cocktail of probes designed to detect 13 HPV types classified as carcinogenic (groups 1 and 2A: HPV16, -18, -31, -33, -35, -39, -45, -51, -52, -56, -58, -59, and -68) by the International Agency for Research on Cancer (11).

In the past few years, several tests for detection of DNA or transcripts of HPVs have been developed (12, 13). A number of these tests have been proposed to be applied for CC screening. Beyond sensitivity and specificity, an additional critical factor before the introduction of a new test as a screening technique is the need for high reproducibility under the diverse conditions that the different clinical laboratories may face (14). The current guide-

Received 13 January 2014 Returned for modification 7 February 2014
Accepted 19 February 2014
Published ahead of print 26 February 2014
Address correspondence to I. G. Bravo, igbravo@iconcologia.net.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JCM.00100-14.

doi:10.1128/JCM.00100-14

lines for HPV tests propose a lower confidence bound of not less than 87% for both intralaboratory reproducibility and interlaboratory agreement (14). Good reproducibility of blinded clinical samples in laboratories might further guarantee the reliability of the results (15).

In Catalonia in northeast Spain, a screening protocol, including HPV testing for selected indications, was implemented in 2006 (16). Women became eligible for the carcinogenic HPV testing if they belonged to any of the three following categories: (i) an inadequate screening history, i.e., women aged older than 40 years with no history of cytology in the previous 5 years, (ii) an incident diagnosis of atypical squamous cells of undetermined significance (ASC-US); or (iii) the first follow-up visit after a surgical conization. During the 2006-2012 period, 116,970 tests for carcinogenic HPVs have been performed within the CC screening activities in the public health sector in Catalonia. In order to monitor interlaboratory reproducibility, a proficiency testing (PT) survey of the HPV samples was launched in 2008, covering the 12 laboratories involved in the screening activities.

The aim of the present study was to evaluate the interlaboratory reproducibility and error rate of performance of the HC2 assay on cervical specimens among the 12 reference laboratories participating in the HPV screening in Catalonia during the period 2008-2011. Additionally, we aimed to investigate the nature and origin of the discrepant results in order to provide guidelines for the assessment of interlaboratory concordance thresholds in the future.

MATERIALS AND METHODS

Design of external proficiency testing program. All tests (original and PT assays) were performed using the HC2 assay with the "high-risk" probe pool only, following the manufacturer's instructions. According to the FDA-approved guidelines, the threshold for positivity is the HC2 assay response against a control sample containing 1.0 pg ml^{-1} HPV DNA, roughly equivalent to 5,000 genomic copies. A sample was considered positive if it rendered a relative light unit (RLU) ratio of $\geq$1 with respect to the positive control.

The 12 laboratories for HPV screening in Catalonia (Hospital Universitari Dr. Josep Trueta, Consoci hospitalari de Vic, Hospital Universitari Joan XXIII, Hospital del Mar, Hospital Universitari de Bellvitge–Institut Català d'Oncologia [ICO], Hospital Clínic, Hospital Universitari Vall d'Hebron, Hospital Universitari Verge de la Cinta, Hospital Universitari Arnau de Vilanova, Consorci sanitari Parc Tauli, and the Laboratoris d'Atenció Primària Doctor Robert and Bon Pastor) participated in the PT program. The laboratory at the ICO was designated by the health department of the government of Catalonia as the reference laboratory for PT purposes. The overall project was approved by the ethics committee of the ICO/Infections and Cancer, Bellvitge Institute of Biomedical Research (IDIBELL). All information regarding the identification of patients was anonymized before analysis.

The PT was conducted annually and blindly. Between 2008 and 2011, each participating laboratory delivered 20 samples to the reference laboratory. The PT of the reference laboratory itself was performed by one of the 12 laboratories. For PT, the laboratories provided the residual aliquot of the original samples. Eleven of the participating laboratories collected the samples in specimen transport medium (STM). One laboratory used liquid cytology-based screening, and in this case, buffer conversion was performed prior to retesting, according to the manufacturer's guidelines. Samples were kept frozen at −20°C before delivery. As far as possible, laboratories were requested to deliver samples for PT within the 3 months after the initial testing. Nevertheless, the possible effect of time elapsed between the original test and the PT was also assessed. In order to ensure a uniform coverage distribution of positive and negative HC2 values, the laboratories were asked to provide annually five randomly chosen samples of each of the following RLU intervals: <0.5, 0.5 to 0.99, 1 to 9.99, and ≥10.

Concordance analysis. HPV samples were collated and analyzed to allow for interlaboratory comparisons. Paired HC2 test results were categorized as original negative/PT negative, original positive/PT negative, original negative/PT positive, and original positive/PT positive based on the 1.0 RLU cutoff.

Cohen's kappa statistics were used to determine the level of agreement between the original and the PT categories of results (positive/negative and RLU intervals). The kappa statistic is a measure of interrater agreement, which tests the interrater agreement by chance as the null hypothesis (17). Generally, a kappa score between 0.8 and 1 is considered excellent agreement, values between 0.61 and 0.8 are considered substantial agreement, values between 0.41 and 0.6 are considered moderate agreement, values between 0.21 and 0.4 are considered fair agreement; and values between 0 and 0.2 are considered slight agreement (18).

The correlations between the original and PT HC2 readouts and the correlations between changes in the signals in both readouts and the time elapsed between both tests were assessed by linear regression using SPSS Statistics 17.0 (SPSS Statistics/IBM, Chicago IL). A paired Wilcoxon and Mann-Whitney test implemented in R (http://www.r-project.org/) was used to test the null hypothesis of the median difference between the original RLU values and the values for PT to be not different from zero.

For all statistical analyses, a P value of <0.05 was considered significant for rejection of the null hypothesis.

Bootstrapping analysis. In order to estimate the limits for the expected number of discrepancies between the original results and the PT results, a bootstrapping analysis was done. One thousand "virtual labs" were generated by bootstrapping among the pooled samples, allowing for replacement. All of these virtual labs were constructed with the same data structure that was requested for the PT, i.e., 20 samples in total, 5 samples from each of the intervals defined above. We also performed the same analyses in another hypothetical scenario, with 40 samples per virtual lab, 10 from each interval. The total numbers of discrepancies and concordances and Cohen's kappa statistics were computed for each of the virtual labs. The distribution of discrepancies was fitted to a Poisson model, and the 95% confidence interval (CI) was calculated based on the fitted model.

We explored further the differential repeatability of the technique for different values of the RLU variable. Each of the intervals, 0.5 to 0.99 and 1.0 to 10 RLU, was divided in 0.1-unit subintervals. For each subinterval, 1,000 virtual labs were created by drawing random samples with replacement from the original pooled samples. The size of each virtual lab was equal to the original number of samples within the subinterval. The mean percentage of discrepancies was calculated for each of the defined subintervals. The 2.5% and 97.5% quantiles of the 1,000 bootstrap samples were used to define the lower bound and the upper bound of the 95% CI. In all cases, bootstrapping analyses were performed using in-house Perl scripts.

RESULTS

During the period 2008-2011, a total of 946 specimens were retested for PT within the framework of the CC screening program in Catalonia, Spain. The results of these paired tests were used to study interlaboratory reproducibility.

Comparison between original and PT paired tests showed high correlation. The distribution of the samples selected for PT was chosen to show a flat distribution of RLU values around the clinical cutoff value. The precise distribution of the samples tested in the PT and the comparison with the distribution of the general population are given in Table 1. In the general population in which the HC2 assay was used to assess the presence of DNA of oncogenic HPVs (n = 46,949), the distributions of positive and negative results for the period 2006-2009 were 22% and 78%, respectively. In contrast, in the PT samples, these distributions

TABLE 1 Data from the general population compared to data from the reference HPV laboratories used for the HPV proficiency testing

Category of results	Screening general population (*n* [%])[a]	Proficiency testing (*n* [%])[b]
HPV negative	36,600 (78)	452 (47.8)
HPV positive	10,349 (22)	494 (52.2)
Total	46,949 (100)	946 (100)
RLU[c] <0.5	34,580 (73.7)	253 (26.7)
0.5 ≤ RLU < 1	2,020 (4.3)	199 (21)
1 ≤ RLU < 10	2,563 (5.5)	246 (26)
RLU ≥10	7,786 (16.6)	248 (26.2)

[a] Data were obtained from the HPV test performed by screening the general population in the period 2007-2009.
[b] Data were obtained from the reference HPV laboratories for the HPV proficiency testing during the period 2008-2011.
[c] RLU, relative light units.

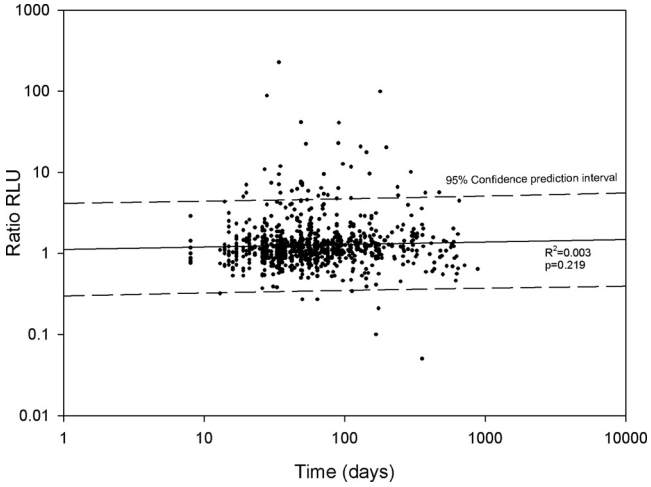

FIG 2 Scatterplot of relative light unit (RLU) ratios between the original and proficiency tests as a function of time elapsed between the two tests. Results are expressed as logarithmic values of the ratio of RLU and time in days. The correlation coefficient obtained for linear regression was 0.003 ($P = 0.219$), and therefore no loss in signal can be attributed to the time elapsed between consecutive tests.

were 52.2% and 47.8%, respectively. Regarding the classification by categories of RLU values, in the HPV-tested general population, the central intervals spanning 0.5 to 9.99 RLU comprised only 9.8% of all samples, while 47% of the PT samples were located in these intervals. These differences arise from the focus of the PT program itself, the aim of which is to evaluate the interlaboratory reliability of the technique and not necessarily match the distribution of the RLU values in the general population.

A comparison of the RLU values for the HC2 test by the 12 tested laboratories and for the repeated test is shown in Fig. 1. The overall correlation coefficient for the original and the PT values was 0.95 ($P < 0.05$), ranging between 0.88 and 0.97 for each of the individual laboratories (data not shown). Correlation restricted to samples interpreted as positive (RLU of ≥1) in both the original test and in the PT retest increased up to 0.97 ($P < 0.05$) (data not shown). The PT retests rendered systematically lower RLU values than those reported by the original laboratories, with a pairwise median decrease of 9% in signal. This median difference was significantly different from zero, as determined by a paired Wilcoxon and Mann-Whitney test ($P < 0.005$).

Time elapsed between original test and PT did not influence reproducibility. The mean time elapsed between the first test and the PT reading was 90.5 days, with a median of 53.0 days and a range from 8 to 885 days. To determine whether time elapsed between consecutive assays on the same sample could influence the differences between the RLU values reported by the HC2 assay, the correlation between the original and the PT readout values was assessed (Fig. 2). No significant correlation was found ($P = 0.83$),

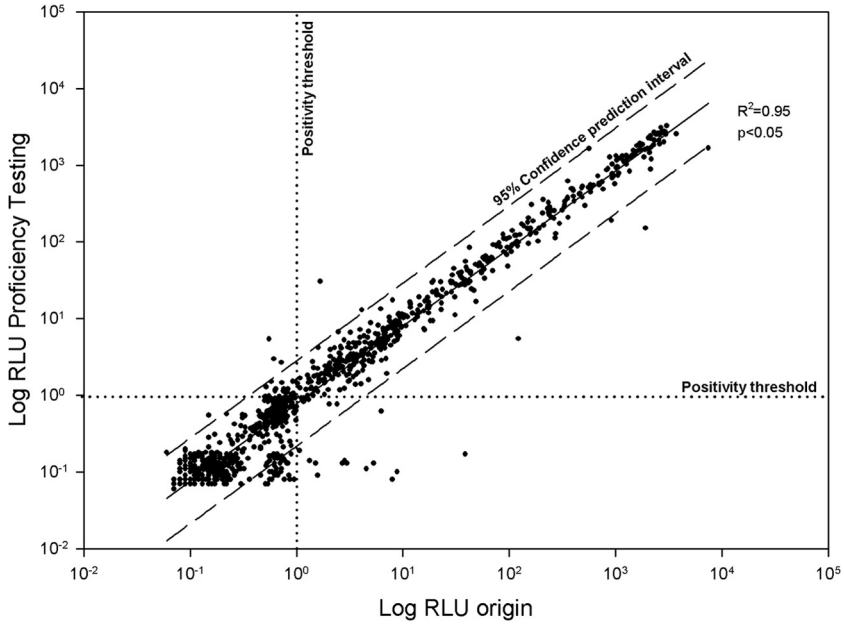

FIG 1 Scatterplot of the relative light unit (RLU) values for the results of the HC2 test at the original laboratory and for the paired proficiency test. Results are expressed as logarithmic values of RLU reads. The correlation coefficient for linear regression was 0.95 ($P < 0.05$).

TABLE 2 Comparison of tested laboratories and HPV proficiency testing laboratory paired HC2 test results[a]

Proficiency testing of HC2 assay results	HC2 assay results for original samples				Total	Analysis of agreement	
	Negative		Positive			Kappa	*P*
Negative	433		25		458	0.91	
Positive	19		469		488		0.00
Total	452		494		946		
	RLU < 0.5	0.5 ≤ RLU < 1	1 ≤ RLU < 10	RLU ≥ 10			
RLU < 0.5	243	81	11	1	336		
0.5 ≤ RLU < 1	10	99	13	0	122	0.79	
1 ≤ RLU < 10	0	19	213	7	239		0.00
RLU ≥10	0	0	9	240	249		
Total	253	199	246	248	946		

[a] HC2, hybrid capture 2.

indicating that the time interval between consecutive tests did not result in a decrease of the RLU readout value.

All laboratories exhibited high agreement with the PT results. In order to assess the agreement between the two tests, Cohen's kappa values were calculated for the positive/negative categorical classification and the RLU intervals (Table 2). Paired comparisons between the original and PT analyses for positive/negative results rendered an overall excellent agreement (kappa = 0.91). The kappa values for the individual laboratories ranged between 0.84 and 0.97, being above 0.90 for eight (66.7%) of them (see Table S1 in the supplemental material). The laboratory using liquid cytology-based samples did not behave differently from the rest of laboratories tested, which used STM as the collecting medium. The total number of discordant results was 44 (4.6% of the total samples). A slightly asymmetric distribution of discordant tests was observed, with more discordant samples being positive in the original test and negative in the PT than *vice versa* (25 versus 19, respectively), although this difference was not significant (*Z* score = 1.28, $P = 0.20$). The vast majority of discrepant results (97.7%) were reported in samples with RLU values between 0.5 and 10.0 in the original tests. Remarkably, 13 of them showed a large difference between the paired tests: 10 positive samples with an original RLU value between 2.0 and 9.0 produced negative results between 0.08 and 0.77 in the PT. In contrast, only two samples with original negative results (RLU values of 0.61 and 0.72) produced positive results (RLU values of 2.97 and 2.98, respectively) in the PT retest. Only one sample was an outlier following Tukey's criterion (19), showing an original result of 38.57 and a 0.17 result in the PT.

The reproducibility of the tests in each of the RLU intervals here defined showed an overall kappa value of 0.79 with values ranging from 0.70 to 0.84 for each of the individual laboratories (see Table S2 in the supplemental material). The observed concordance in the extreme intervals, i.e., RLU value of <0.5 and RLU value of ≥10, was higher than that in the intervals around the cutoff, i.e., RLU value between 0.5 and 10, with overall agreement values of 96.4% and 70.1%, respectively.

Exploring new alternatives in the design of the PT. Finally, we explored different alternatives for future improvement of the PT design. First, we aimed to provide a threshold for identifying significant deviations from the maximum expected number of discrepancies, under the current sampling scheme for PT. Limits for a discrepancy report were determined under two scenarios: 20 samples analyzed for PT per laboratory with 5 samples per each interval (i.e., the one currently implemented) and 40 samples per laboratory with 10 samples for each interval (which could be implemented depending on the budget). We generated, by bootstrapping with replacement from the original data, 1,000 data sets that corresponded to virtual labs, using the same stratified structure of samples (Fig. 3). The results were fitted to a Poisson distribution and the accumulated probability for the detection of the corresponding number of discrepant results was calculated. We aimed to find the number of discrepant results that could be expected to appear by chance with a probability of <5% (Fig. 4). Using the stratified structure of samples evaluated in the present study ($n = 20$, five samples of each interval), the probability of having two or more discrepancies was 7.5%, whereas the probability of three or more discrepancies was 1.7%. If 40 samples in total were collected, the probability of finding three or more discrepancies was 13.1%, while retrieving four or more discrepancies presented a probability of 4.7%. Thus, under the present PT structure with retesting of 20 samples per laboratory, the analyses should be labeled as significantly discordant when PT yields three or more discrepant samples. If 40 samples were to be retested using the same structure, the flag for significant discordance should be raised for sets with four or more discrepant samples.

Finally, we explored whether the results could be used to improve the PT design, by modifying the boundaries for sample stratification. We focused on the readout values with lower reproducibility, identified above by means of kappa values as the area around the positivity threshold. This gray area comprised two of our RLU categories (0.5 to 0.99 and 1 to 9.99). By using bootstrapping analyses as described above, we explored the expected percentage of discrepancies for both RLU categories (Fig. 4). The results indicated that in the 0.5 ≤ RLU < 1 interval the expected number of samples showing discrepant results did not experience significant changes, remaining flat around 13.7% (95% CI, 5.5 to 28.6%). In the 1 ≤ RLU < 10 interval, we observed a significant monotonic decrease (P for trend < 0.001) in the percentage of samples showing discrepant results, starting from 25.9% (95% CI, 14.8 to 38.9%) for values in the 1 ≤ RLU < 2 interval and stabilizing after the RLU ≥5 interval to reach 9.7% (95% CI, 5.7 to

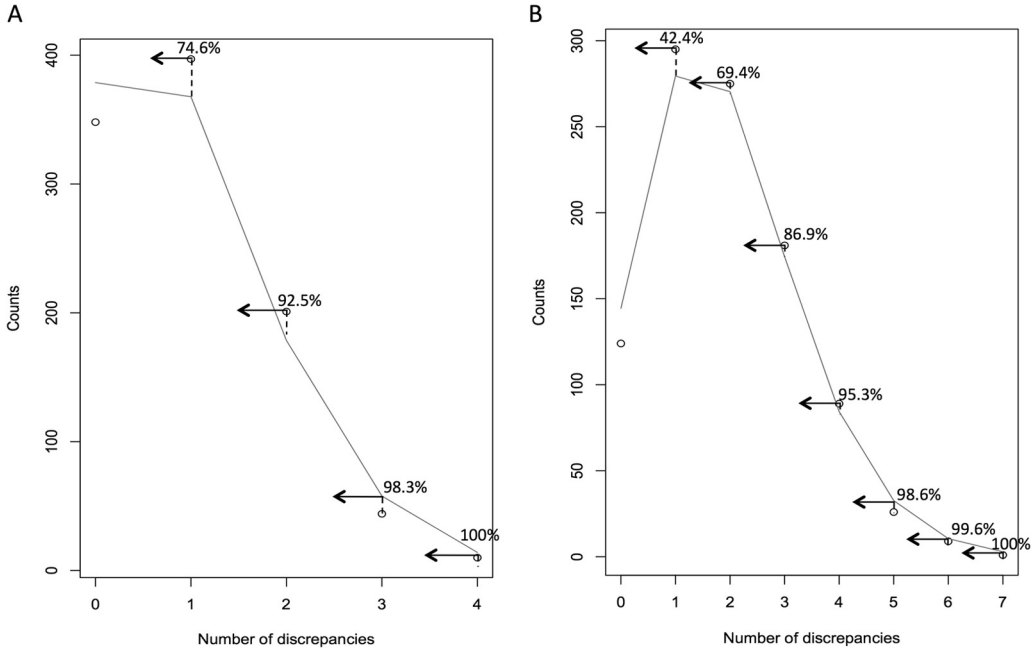

FIG 3 Distribution of numbers of clinically discrepant results between the original laboratory and the proficiency test (circles) and fit to a Poisson distribution (gray line). The accumulated probability for the detection of the corresponding number of discrepancies is shown above each point. By bootstrapping with replacement of the all samples, 1,000 virtual labs were generated, and the numbers of discrepancies were calculated. With the same stratified structure of samples used in this study (RLU <0.5, $0.5 \leq$ RLU < 1, $1 \leq$ RLU < 10, and RLU ≥ 10), two different scenarios were tested: 20 samples, 5 from each interval (A); and 40 samples, 10 from each interval (B).

13.4%). Moreover, we have explored the differential probability for a sample to show a discrepancy upon retesting, as a function of the RLU values. The highest probability for discrepancies was concentrated in the RLU interval of 0.5 to 5; the probability for a sample in this interval to show a discrepancy was 10.80% (95% CI, 7.86 to 14.33) compared with 0.85% (95% CI, 0.17 to 1.69) for samples outside this interval. The accumulated probabilities for discrepancies for the different intervals of the RLU variable are given in Table 3. These improved intervals will be implemented in the future for the PT activities when the HC2 assay is used. We will thus ask the participating laboratories to annually provide randomly chosen samples from each of the following RLU intervals with the following structure: RLU <0.5, 5 samples; $0.5 \leq$ RLU $<$ 1, 10 samples; $1 \leq$ RLU < 5, 10 samples; and RLU ≥ 5, 5 samples. In this case, the probability of having three or more discrepancies by chance is 13.5%, whereas the probability of three four or more discrepancies by chance is 4.8%.

DISCUSSION

An interlaboratory PT program was successfully implemented for the detection of high-risk HPV types within the CC screening activities in the public health system in Catalonia. It was considered instrumental to ensure the accuracy of the results obtained by HPV testing. A total of 946 samples were analyzed for PT during the 2008-2011 period, and a total of 44 (4.6%) discrepancies were found. The present PT study was in agreement with the guidelines proposed by Meijer and coworkers for the validation of high-risk HPV tests for primary CC screening (14). These guidelines proposed that interlaboratory agreement should be determined by evaluation of at least 500 samples, with 30% of the samples having tested positive in a reference laboratory using a clinically validated assay and reaching an agreement with a lower confidence bound not less than 87%.

It is very important to note that the distribution of the RLU values used for the PT does not follow the distribution obtained from the general population participating in the screening algorithms in which the HC2 assay was used to assess the presence of DNA of oncogenic HPVs (Table 1). Differences between distributions reflect the analytical nature of our PT program, as we were interested in assessing the overall performance of the laboratories using the HC2 technique. We chose therefore to design a balanced PT sample distribution to explore the full range of the readout variables equally, as otherwise the central values, close to the cutoff and more prone to discrepancies, would have been undersampled.

Paired tests demonstrated an almost excellent interlaboratory agreement for all 12 participating laboratories, for both positive and negative agreement (kappa = 0.91) and for the four RLU categories (kappa = 0.79). In our study, RLU values were slightly but significantly lower in the PT program retesting than those obtained in the original laboratory, with a median decrease of 9.0% from the original RLU value. One possible explanation for this trend is specimen degradation between the two tests, as has been reported to a certain extent for the HC2 assay (20, 21). We tested the hypothesis of the influence of time elapsed between the two consecutive tests to account for this difference, but we did not find a significant association between the decrease in the RLU readout variable and the time between the two HC2 tests. An alternative explanation is the impact of consecutive freeze-thaw cycles, as samples were denatured, analyzed originally, frozen, sent, and thawed for PT analysis, especially because the HC2 test does not include an amplification step. Finally, although in abso-

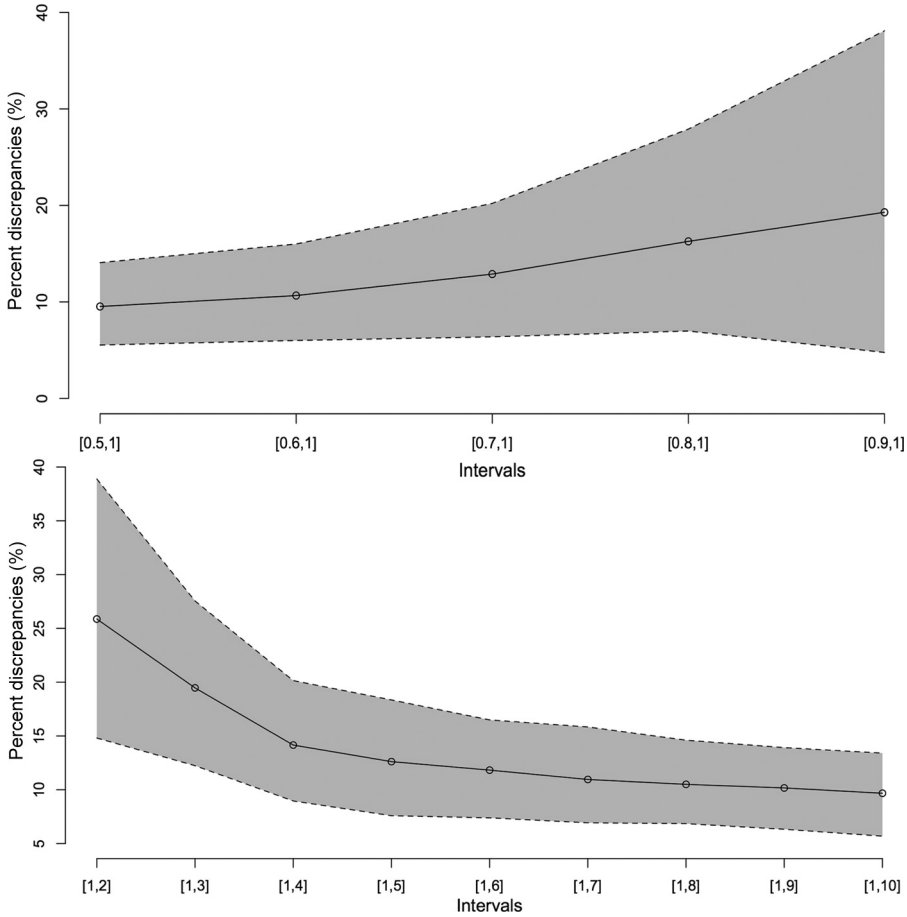

FIG 4 Expected percentages of discrepancies along the relative light unit (RLU) intervals of 0.5 to 0.99 (A) and 1 to 9.99 (B). Mean percentages of discrepancies were obtained by bootstrapping. The gray area encompasses the 95% confidence interval.

lute terms more samples were found to be positive originally and tested negative in the PT than the reverse case, the difference between the two numbers was not statistically significant. Overall, the small decrease in signals during retesting has little or no clinical significance because HPV retesting is not part of the routine clinical practice (21).

TABLE 3 Comparison between the percentage of discrepancies between the original results and the results after proficiency testing, as a function of the RLU value[a]

Categories of samples	Mean % (95% CI) probability of discrepancy
Proficiency testing	
RLU <0.5	0 (0)
0.5 ≤ RLU < 1	9.59 (6.02–13.58)
1 ≤ RLU < 10	9.87 (6.50–13.82)
RLU ≥10	0.41 (0–1.21)
Proposed proficiency testing	
RLU < 0.5	0 (0)
0.5 ≤ RLU < 1	9.59 (6.02–13.58)
1 ≤ RLU < 5	12.56 (7.64–18.47)
RLU ≥5	1.18 (0.30–2.87)

[a] The categories used in the current proficiency testing scheme and those proposed for the improved proficiency testing are shown.

Several reports have addressed the use of the HC2 assay in CC screening and its performance compared to those of other methods for detecting HPV genetic material (13, 22–24). However, few data on the reproducibility of HC2 assays are available. In the LSIL (low-grade squamous intraepithelial lesion) Triage Study (ALTS) study, interlaboratory reproducibility values across four laboratories were found to be similar to those communicated herein (kappa = 0.84) (21). The interlaboratory reproducibility for HC2 in seven laboratories participating in an Italian clinical trial was also high (15). In both cases, the RLU values around the cutoff (RLU = 1) were more prone to show discrepant results, as observed in our study. Thus, the random variation in RLU values had more influence on the classification as positive or negative for samples having values close to the cutoff (15). One fundamental contribution of the present study is the detailed description of the differential reproducibility of the HC2 technique for different intervals of RLU values. We conclude, therefore, that, in our settings, 0.5 ≤ RLU < 5 is the interval in which the PT paired values of the readout RLU values were more likely to show discrepancies.

A central aim of the initial PT assessment after the first 4 years of the CC screening activities was to provide guidelines for a better PT control in the future. As such, we have estimated the number of samples that should test discrepant after retesting to be considered a "significant discordant." In our case and with our data structure,

this critical value is 3 or more individual discrepant samples in the sets of 20 samples to be retested. Two factors in our PT design, the number of retested samples and the stratified structure, could be changed for improvement in the future. Increasing the number of retested samples would allow us to more accurately find a number of discrepancies that were significantly higher than expected by chance. This has been exemplified in the simulation described above, by including 40 instead of 20 samples per laboratory per year. Obviously, this change also doubles the cost of the PT assessment, and a cost-benefit equilibrium must be found. Regarding sampling structure, the aim of the PT program is analytical, not clinical, and we have thus chosen to monitor values throughout the dynamic range of the readout variable equally, instead of randomly sampling from the entire population. This strategy allows us to place special emphasis on the region in which the technique is less reproducible. In order to focus in the sensitive area in which samples were more likely to be discordant (0.5 to 5 RLU), we propose that the following sample structure be applied for future PT assessments: 5 samples with RLU <0.5, 10 samples with $0.5 \leq$ RLU < 1, 10 samples with $1 \leq$ RLU < 5, and 5 samples with RLU ≥ 5.

In conclusion, the results of the PT assessment for the CC screening program in Catalonia have shown that the HC2 assay has a high interlaboratory concordance. The use of a common, standardized protocol with well-defined anticontamination measures for samples and processing fluids in the 12 laboratories involved has been instrumental for achieving these excellent results. We have also detected a significant decrease in the HC2 signals after retesting that cannot be linked to the time elapsed between consecutive tests and that may be attributable to the additional freeze-thaw cycle. We have additionally explored in depth the differential repeatability along the dynamic response of the readout variable and have demonstrated that most discrepant results accumulate in the 0.5 to 5 RLU interval, with samples in this interval being 12.7 times more likely to show discrepant results than samples outside this interval. With this information, we have further defined recommendations and confidence thresholds for interlaboratory PT in the future when the HC2 assay is used as screening test, as analytical quality assessment of HPV DNA detection remains a central component of the screening program for CC prevention.

ACKNOWLEDGMENTS

This study was partially supported by the Pla Director d'Oncologia of the Health Department in Catalonia and grants from the Instituto de Salud Carlos III (Spanish government grants RCESP C03/09, RTICESP C03/10, RTIC RD06/0020/0095, RD12/0036/0056, and CIBERESP) and from the Agència de Gestió d'Ajuts Universitaris i de Recerca (Catalan government grants AGAUR 2005SGR 00695 and 2009SGR126). S. de Sanjosé has received occasional travel grants from Merck and Qiagen and holds a restricted grant from Qiagen and from Merck not related to the study presented. None of the funding bodies had any role in study design; in collection, analysis, and data interpretation; in the writing of the manuscript; or in the decision to submit the manuscript for publication.

We thank Attila Lorincz for his comments and remarks on an initial version of the manuscript. We thank the following persons for their collaboration in this study: Marylène Lejeune, Carlos López, and Cristina Callau from the Molecular Biology and Research Section, IISPV, URV, Hospital de Tortosa Verge de la Cinta (HTVC); Anna Ferran Gibert, Mercé Rey Ruhi, Ruth Orellana Fernandez, and Irmgard Costa Tranchel from the UDIAT-CD, Hospital Universitari Parc Taulí; Roser Esteve from the Barcelona Hospital Clínic; Montserrat Sardà Roca and Àngels Verdaguer Autonell from the Consorci Hospitalari de Vic; Jo Ellen Klaustermeier, Vanesa Camón, Ana Esteban, Yolanda Florencia, Isabel Català, and Núria Baixeras from the Hospital universitari de Bellvitge–Institut Català d'Oncologia; José Maria Navarro Olivella, Carme Perich Alsina, and Belen Viñado from the Laboratori Clínic Bon Pastor; Cristina Antúnes Vitales, Glòria Oliveras Serrat, Lluís Bernadó Turmo, and Pilar Castro Marqueta from the Institut Català d'Oncologia (Girona); Ana Velasco Sánchez, Anna Serrate López, Marta Romero Fernández, Azahar Romero Jiménez, and Nuria Llecha from the Hospital Universitari Arnau de Vilanova; Vanessa Guerrero Hormiga and Ignacia Ramos Mora from the Hospital Universitari Joan XXIII; Rosa Tenllado Gimenez and Adoración Velilla from the Laboratori Barcelonès Nord i Vallès Oriental; and Belén Lloveras, Francesc Alameda, and Merced Muset from the Hospital del Mar.

R. Ibáñez and I. G. Bravo conceived the analyses, R. Ibáñez and S. de Sanjosé coordinated the screening activities, R. Ibáñez, M. Félez-Sánchez, and I. G. Bravo analyzed the data, J. M. Godínez, C. Guardià, E. Caballero, R. Juve, N. Combalia, B. Bellosillo, D. Cuevas, J. Moreno-Crespi, L. Pons, J. Autonell, C. Gutierrez, and J. Ordi performed experiments, and R. Ibáñez, M. Félez-Sánchez, and I. G. Bravo drafted the manuscript. All the authors read and approved the final manuscript.

REFERENCES

1. **de Sanjose S, Quint WG, Alemany L, Geraets DT, Klaustermeier JE, Lloveras B, Tous S, Felix A, Bravo LE, Shin HR, Vallejos CS, de Ruiz PA, Lima MA, Guimera N, Clavero O, Alejo M, Llombart-Bosch A, Cheng-Yang C, Tatti SA, Kasamatsu E, Iljazovic E, Odida M, Prado R, Seoud M, Grce M, Usubutun A, Jain A, Suarez GA, Lombardi LE, Banjo A, Menendez C, Domingo EJ, Velasco J, Nessa A, Chichareon SC, Qiao YL, Lerma E, Garland SM, Sasagawa T, Ferrera A, Hammouda D, Mariani L, Pelayo A, Steiner I, Oliva E, Meijer CJ, Al-Jassar WF, Cruz E, Wright TC, Puras A, Llave CL, et al.** 2010. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. Lancet Oncol. **11:**1048–1056. http://dx.doi.org/10.1016/S1470-2045(10)70230-8.
2. **Muñoz N, Bosch FX, de Sanjosé S, Herrero R, Castellsague X, Shah KV, Snijders PJF, Meijer CJLM.** 2003. Epidemiologic classification of human papillomavirus types associated with cervical cancer. N. Engl. J. Med. **348:** 518–527. http://dx.doi.org/10.1056/NEJMoa021641.
3. **Walboomers JMM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, Snijders PJF, Meifer CJLM, Munoz N.** 1999. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J. Pathol. **189:**12–19. http://dx.doi.org/10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F.
4. **IARC Working Group on the Evaluation of Cancer Prevention Strategies.** 2005. IARC handbooks of cancer prevention, vol 10. Cervix cancer screening. IARC Press, Lyon, France.
5. **Ronco G, Dillner J, Elfstrom KM, Tunesi S, Snijders PJ, Arbyn M, Kitchener H, Segnan N, Gilham C, Giorgi-Rossi P, Berkhof J, Peto J, Meijer CJ.** 2014. Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials. Lancet **383:**524–532. http://dx.doi.org/10.1016/S0140-6736(13)62218-7.
6. **Arbyn M, Ronco G, Anttila A, Meijer CJ, Poljak M, Ogilvie G, Koliopoulos G, Naucler P, Sankaranarayanan R, Peto J.** 2012. Evidence regarding human papillomavirus testing in secondary prevention of cervical cancer. Vaccine **30**(Suppl 5):F88–F99. http://dx.doi.org/10.1016/j.vaccine.2012.06.095.
7. **Anttila A, Kotaniemi-Talonen L, Leinonen M, Hakama M, Laurila P, Tarkkanen J, Malila N, Nieminen P.** 2010. Rate of cervical cancer, severe intraepithelial neoplasia, and adenocarcinoma in situ in primary HPV DNA screening with cytology triage: randomised study within organised screening programme. BMJ **340:**c1804. http://dx.doi.org/10.1136/bmj.c1804.
8. **Naucler P, Ryd W, Tornberg S, Strand A, Wadell G, Elfgren K, Radberg T, Strander B, Johansson B, Forslund O, Hansson BG, Rylander E, Dillner J.** 2007. Human papillomavirus and Papanicolaou tests to screen for cervical cancer. N. Engl. J. Med. **357:**1589–1597. http://dx.doi.org/10.1056/NEJMoa073204.
9. **Rijkaart DC, Berkhof J, Rozendaal L, van Kemenade FJ, Bulkmans NW,**

Heideman DA, Kenter GG, Cuzick J, Snijders PJ, Meijer CJ. 2012. Human papillomavirus testing for the detection of high-grade cervical intraepithelial neoplasia and cancer: final results of the POBASCAM randomised controlled trial. Lancet Oncol. **13:**78–88. http://dx.doi.org/10.1016/S1470-2045(11)70296-0.

10. **Ronco G, Giorgi-Rossi P, Carozzi F, Confortini M, Dalla Palma P, Del Mistro A, Ghiringhello B, Girlando S, Gillio-Tos A, De Marco L, Naldoni C, Pierotti P, Rizzolo R, Schincaglia P, Zorzi M, Zappa M, Segnan N, Cuzick J.** 2010. Efficacy of human papillomavirus testing for the detection of invasive cervical cancers and cervical intraepithelial neoplasia: a randomised controlled trial. Lancet Oncol. **11:**249–257. http://dx.doi.org/10.1016/S1470-2045(09)70360-2.
11. **IARC.** 2012. IARC monographs on the evaluation of carcinogenic risks to humans, vol 100B. A review of human carcinogens: biological agents. IARC Press, Lyon, France.
12. **Cuzick J, Bergeron C, von Knebel Doeberitz M, Gravitt P, Jeronimo J, Lorincz AT, Meijer CJLM, Sankaranarayanan R, Snijders PJF, Szarewski A.** 2012. New technologies and procedures for cervical cancer screening. Vaccine **30**(Suppl 5)**:**F107–F116. http://dx.doi.org/10.1016/j.vaccine.2012.05.088.
13. **Poljak M, Cuzick J, Kocjan BJ, Iftner T, Dillner J, Arbyn M.** 2012. Nucleic acid tests for the detection of alpha human papillomaviruses. Vaccine **30**(Suppl 5):F100–F106. http://dx.doi.org/10.1016/j.vaccine.2012.04.105.
14. **Meijer CJ, Berkhof H, Heideman DA, Hesselink AT, Snijders PJ.** 2009. Validation of high-risk HPV tests for primary cervical screening. J. Clin. Virol. **46**(Suppl 3)**:**S1–S4. http://dx.doi.org/10.1016/S1386-6532(09)00540-X.
15. **Carozzi FM, Del Mistro A, Confortini M, Sani C, Puliti D, Trevisan R, De Marco L, Tos AG, Girlando S, Palma PD, Pellegrini A, Schiboni ML, Crucitti P, Pierotti P, Vignato A, Ronco G.** 2005. Reproducibility of HPV DNA testing by Hybrid Capture 2 in a screening setting. Am. J. Clin. Pathol. **124:**716–721. http://dx.doi.org/10.1309/84E5WHJQHK83BGQD.
16. **Generalitat de Catalunya, Departament de Salut.** 2007. Protocol de les activitats per al cribratge del càncer de coll uterí a l'atenció primària. Institut Català d'Oncologia, Barcelona.
17. **Landis JR, Koch GG.** 1977. The measurement of observer agreement for categorical data. Biometrics **33:**159–174. http://dx.doi.org/10.2307/2529310.
18. **Brennan P, Silman A.** 1992. Statistical methods for assessing observer variability in clinical measures. BMJ **304:**1491–1494. http://dx.doi.org/10.1136/bmj.304.6840.1491.
19. **Tukey JW.** 1977. Exploratory data analysis. Addison-Wesley, Reading, United Kingdom.
20. **Castle PE, Hildesheim A, Schiffman M, Gaydos CA, Cullen A, Herrero R, Bratti MC, Freer E.** 2003. Stability of archived liquid-based cytologic specimens. Cancer **99:**320–322. http://dx.doi.org/10.1002/cncr.11723.
21. **Castle PE, Wheeler CM, Solomon D, Schiffman M, Peyton CL.** 2004. Interlaboratory reliability of Hybrid Capture 2. Am. J. Clin. Pathol. **122:**238–245. http://dx.doi.org/10.1309/BA43HMCAJ26VWQH3.
22. **Gravitt PE, Schiffman M, Solomon D, Wheeler CM, Castle PE.** 2008. A comparison of linear array and hybrid capture 2 for detection of carcinogenic human papillomavirus and cervical precancer in ASCUS-LSIL triage study. Cancer Epidemiol. Biomarkers Prev. **17:**1248–1254. http://dx.doi.org/10.1158/1055-9965.EPI-07-2904.
23. **Hesselink AT, Bulkmans NW, Berkhof J, Lorincz AT, Meijer CJ, Snijders PJ.** 2006. Cross-sectional comparison of an automated hybrid capture 2 assay and the consensus GP5+/6+ PCR method in a population-based cervical screening program. J. Clin. Microbiol. **44:**3680–3685. http://dx.doi.org/10.1128/JCM.02078-05.
24. **Sankaranarayanan R, Gaffikin L, Jacob M, Sellors J, Robles S.** 2005. A critical assessment of screening methods for cervical neoplasia. Int. J. Gynaecol Obstet. 89(Suppl 2)**:**S4–S12. http://dx.doi.org/10.1016/j.ijgo.2005.01.009.