Published in final edited form as:

Cancer Causes Control. 2012 December ; 23(12): 2013–2021. doi:10.1007/s10552-012-0081-0.

A modified Latent Class Model assessment of human papillomavirus-based screening tests for cervical lesions in women with atypical glandular cells: a Gynecologic Oncology Group study

Randy L. Carter^{1,2}, Le Kang², Kathleen M. Darcy^{1,*}, James Kauderer¹, Shu-Yuan Liao³, William H. Rodgers⁴, Joan L. Walker⁵, Heather A. Lankes¹, S. Terence Dunn⁶, and Eric J. Stanbridge⁷

¹GOG Statistical and Data Center; Roswell Park Cancer Institute, Buffalo, NY

²Department of Biostatistics, University at Buffalo, Buffalo, NY

³Department of Epidemiology, University of California at Irvine, School of Medicine, Irvine, CA

⁴Department of Pathology, NSLIJ/Lenox Hill Hospital, New York, NY

⁵Department of Obstetrics and Gynecology, University of Oklahoma Health Sciences Center, Oklahoma City, OK

⁶Department of Pathology, University of Oklahoma Health Sciences Center, Oklahoma City, OK

⁷Department of Microbiology & Molecular Genetics, University of California at Irvine, School of Medicine, Irvine, CA

Abstract

Purpose—Absent gold standard diagnoses, we estimate age-specific false-positive and false-negative prediction rates of HPV-, cytology- and histology-based tests for significant cervical lesions (SCL) in U.S. women with AGC-NOS Pap smear diagnoses.

Methods—Modified Latent Class Model (LCM) analyses, with prevalence of SCL modeled as a function of age, were applied to GOG-0171 study data (n=122). The accuracies of several HPV-based tests, including Hybrid Capture II high-risk HPV (HC2 H-HPV); carbonic anhydrase IX (CA-IX) and invasive histological diagnosis were compared. 1-PPV and 1-NPV were written as functions of sensitivity, specificity and prevalence to obtain age-specific false-positive and false-negative rates.

Results—The histology-based test was nearly perfect (sensitivity=1.00, CI=0.98-1.00; specificity=0.99, CI=0.96-1.00). Otherwise, HC2 H-HPV performed best (sensitivity=1.00, CI=1.00-1.00; specificity=0.87, CI=0.79-0.94). The false-positive detection rates (1-PPV) for HC2 H-HPV were high (>17%) at each age, while those of the histological diagnoses were low (<5% at ages 60 and <17% overall ages). False negative prediction rates (1-NPV) for HC2 H-HPV were < 0.11% at each age and were uniformly lower than those of other tests, including the histology-based test (<0.25%). CA-IX together with HC2 H-HPV did not improve performance.

Corresponding Author: Randy L. Carter, Ph.D., Department of Biostatistics, University at Buffalo, 3435 Main Street, Buffalo, NY 14214, Phone: 716-829-2884, rcarter@buffalo.edu.

*Dr. Darcy is currently employed at Women's Health Integrated Research Center at INOVA Health System

Conflict of Interest Statement

The authors wish to report that there are no conflicts of interest.

Conclusions—Women with negative HC2 H-HPV can safely forego invasive treatment (i.e., cone or LEEP biopsy, hysterectomy) in favor of observational follow-up. Additional biomarkers must be found for use in combination with HC2 H-HPV to reduce false-positive rates. This novel application of a modified Latent Class Model (LCM) exemplifies methods for potential use in future cancer screening studies when gold standard diagnoses are not available.

PURPOSE

Women with a cytological diagnosis of atypical glandular cells of undetermined significance (AGC-NOS) have a high prevalence of intraepithelial neoplasia (CIN2, CIN3), adenocarcinoma in-situ (AIS), or invasive carcinomas, collectively referred to as significant cervical lesions (SCL) [1-4]. Yet, a notably high percentage of these women do not have SCL. No SCL was observed in 72% of AGC-NOS cases in our recent study of U.S. data [5], while 71% of AGUS (*aka*, AGC-NOS) women had no SCL in a meta-analysis of the English literature [6]. Clearly, many women with AGC-NOS do not require the aggressive level of treatment that is often provided to them. It is not clear, however, how to identify them. After a diagnosis of AGC-NOS, additional screening is needed to distinguish between those women who have a SCL from those who do not in order to avoid an unnecessarily high referral rate and over-treatment of healthy women [7]. Our Gynecologic Oncology Group study, GOG-0171, found that additional screening for high-risk HPV (H-HPV) infection by Hybrid Capture II (HC2) among women with AGC-NOS in the U.S. produced an estimated 97% sensitivity, 87% specificity, 99% negative predictive value (NPV), and a positive predictive value (PPV) of only 77%, when treating histological diagnosis as the truth. Only 1% ($=100\text{-NPV}\times 100$) of women with negative HC2 H-HPV status had significant cervical lesions [5].

Many women with AGC-NOS receive invasive treatment that leaves them infertile. This is an especially important concern for young women when choosing between surgery and fertility preserving observational follow up. Since PPV varies with prevalence, it is important to know the false-positive prediction rate (1-PPV) of the screening test as a function of age if the prevalence of SCL varies with age. Negative predictive value (NPV) also varies with prevalence and, therefore, the false-negative rate (1-NPV) of HC2 H-HPV would be a function of age if prevalence varies with age. Our previous global estimate of the false-negative rate for HC2 H-HPV suggests that it is safe for women to forego surgical treatment when they are HC2 H-HPV negative [5]. It is not known, however, whether age-specific false-negative rates would justify such a conclusion for women of childbearing age. It is also not known what the risk a false-positive test result is for women in this age range because age specific false-positive and false-negative rates of HPV-based and other screening tests are currently not available.

Another problem that is often present but seldom addressed in cancer screening studies is the absence of perfectly accurate gold standard diagnoses. In the current context, for example, histological evaluation of the cervix is not perfect and is subject to an unknown level of diagnostic error. “Gold standard” (“G.S.”) diagnoses were based on histological evaluation of the cervix within six months of initial cytological diagnosis of AGC-NOS. Although expected to be quite accurate, the “G.S.” diagnoses were not totally free from misclassifications. For example, CIN2 is known to be an equivocal diagnosis of pre-cancer, and there is less than perfect agreement among pathologist when evaluating specimens. In general, misclassifications by a “G.S.” are known to bias estimates of diagnostic accuracy (i.e., sensitivity and specificity) toward zero, sometimes seriously (See, for example, Valenstein [8]). Latent class model (LCM) analysis was developed by statisticians to address this issue. Gaffikin, *et al.* [9] used LCM analysis to assess the accuracy of several screening tests, including an HPV-based test and a test based on colposcopy/biopsy, and obtained very

different estimates of sensitivity and specificity than from the standard analysis with colposcopy/biopsy treated as the gold standard. Unfortunately, classical LCM analysis does not produce the age-specific estimates of false-positive or false-negative rates that are desired in this ancillary study to GOG-0171.

The purpose of the current study is two-fold: (1) to introduce a novel modification of the LCM that models prevalence as a function of age; and (2) to apply the modified LCM to estimate age-specific false-positive and false-negative prediction rates of HPV-, cytology- and histology-based tests for SCL in U.S. women with AGC-NOS Pap smear diagnoses.

To achieve these goals, we will apply LCM analysis to the GOG-0171 study data [5]. One can infer from the probabilistic modeling of Myers *et al.* [10] that the prevalence of SCL and, hence, 1-PPV and 1-NPV of diagnostic tests depend on age (See their Figure 2). Thus, we modify the usual LCM, which solves the problem of an imperfect reference test, to allow prevalence of SCL to vary with age. We estimate and compare the sensitivities, specificities, and age-specific 1-PPV (*a.k.a.*, false positive test rate) and 1-NPV (*a.k.a.*, false negative test rate) for: (1) the histology-based reference test that was used in our previous study [5] (*i.e.*, imperfect “G.S.”); and for the biomarkers (2) HC2 H-HPV; (3) carbonic anhydrase IX (CA-IX) immunostaining; and (4) several diagnostic tests that are based on HPV genotypes detected by the Roche LINEAR ARRAY (RLA) genotyping test, including two denoted herein by RLA H-HPVG (any one of 13 high-risk HPV genotypes present or not) and RLA HPV (any one of the 37 known HPV genotypes present or not).

MATERIALS AND METHODS

This study was based on a secondary analysis of data collected according to the GOG-0171 study protocol [5]. A detailed presentation of methods is presented in the online supplemental materials accompanying this paper. The essentials follow:

Patients with a cytological diagnosis of AGC-NOS who were 18 years of age or older and who consented to participate were included in the study. Exclusion criteria are those of our previous study [5] and are summarized in the online supplemental materials.

Assessments for detection of SCL included histologic evaluation of the cervix, HPV testing by HC2 and RLA using liquid-based cytology (LBC; ThinPrep, Cytoc/Hologic, Marlborough, MA), and immunocytochemical determination of CA-IX expression in conventional Pap smear specimens. We used the histological-based diagnostic test as the “G.S.” in our previous evaluation of diagnostic accuracy of HPV and CA-IX [5]. Because the histological diagnoses are not perfect, we will re-evaluate the accuracy of this “G.S.” test along with that of CA-IX and three HPV-based diagnoses obtained from LBC Pap specimens (HC2 H-HPV, RLA H-HPVG, RLA HPV). We also evaluate anew several RLA-based tests defined using subsets of the 13 high-risk HPV genotypes. The histology-based test was defined to be positive if CIN2, CIN3, AIS, or an invasive carcinoma were observed in cervical tissue obtained by a cone/LEEP biopsy, or a hysterectomy, or by a regular biopsy of the cervix. A complete evaluation of the cervical transformation zone was required for negative diagnosis. Conventional Pap smear specimens were immunostained for CA-IX as described previously [5,7,11,12]. The presence of any of 13 high-risk HPV (H-HPV) DNA types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, or 68) was detected using HC2 (Digene Corp., Gaithersburg, MD). This test does not distinguish between the different H-HPV types. RLA genotype tests were performed on 112 of the 122 women for whom HC2 H-HPV results were observed to determine the presence or absence of each of 37 anogenital HPV genotypes (high-risk or low-risk). Two separate RLA test results were considered initially for analysis: RLA H-HPVG, was defined as positive if any one of the 13

H-HPV types were present, whereas RLA HPV was positive if any of the 37 types were observed. The RLA tests identify the individual HPV type(s) present in each positive sample [13,14]. The RLA tests were included to: (1) determine whether the test based on all 37 genotypes (RLA HPV) performs better than that based on just the 13 high-risk genotypes (RLA H-HPV), while holding manufacturer, method of DNA detection, *etc.* constant; and (2) confirm our expectation that RLA H-HPV performs similarly to the HC2 H-HPV. Inclusion of the RLA data allowed us to search for tests, based more or fewer genotypes than the 13 high-risk ones, that perform better than HC2 H-HPV.

Statistical Methods

Initially, three Latent Class Model (LCM) analyses [15,16] were performed, each including three diagnostic/predictive variables. Two of the three variables in each analysis were “G.S.” *i.e.*, the histology-based assessment of presence or absence of SCL; and CA-IX. The third test was HC2 H-HPV, RLA HPV, or RLA H-HPV in the first, second, or third analyses, respectively. LCM analysis has been recommended for routine use when assessing diagnostic accuracy in the absence of a perfect gold standard [17]. We extend the classical LCM to allow the prevalence parameter to vary as a function of covariates (*e.g.*, age). The EM-algorithm [18] with Monte Carlo approximation in the E-step [19] was used to obtain the maximum likelihood estimates of model parameters. The prevalence of SCL was modeled as a logistic function of age and age squared. Bayes’ rule was used to obtain 1-PPV and 1-NPV as functions of prevalence [20]. Bootstrap methods [21] were used to estimate standard errors of the estimators and the percentile method was used to calculate 95% confidence intervals for the sensitivities and specificities of each diagnostic test and 95% confidence bands for 1-PPV and 1-NPV as functions of age. A Receiver Operating Characteristics (ROC) analysis [20] was performed to determine how best to combine information in HC2 H-HPV result with that of the CA-IX test to optimize predictive power.

Results of the above analyses raised the question “Can the performance of the RLA H-HPV test be improved by basing it on a subset of the 13 high-risk HPV types?”. To address this question, we performed a logistic regression of histological diagnosis of SCL on indicators of the presence or absence of each of the 13 types. The backward selection strategy was used to obtain a model that included only significantly related types. A test was then defined to be positive if any one of the significant types was positive. A modified LCM analysis was performed to assess the subset test’s accuracy for comparison to the HC2 H-HPV and RLA H-HPV tests. Additionally, six of the most promising subsets of the 13 high-risk genotypes were selected from among the 8192 possible subsets and compared to HC2 H-HPV and RLA H-HPV. Selection of the six subsets was based on the screening test derived from them, as defined above, having a sensitivity or specificity greater than HC2 H-HPV when initially using the histology based diagnoses as the truth. A modified LCM analysis was performed to obtain better estimates of accuracy for each of these six tests. The Youden index ($sensitivity + specificity - 1.0$) was used as the measure of overall accuracy to compare the tests based on subsets to HC2 H-HPV and RLA H-HPV.

RESULTS

The age range in the study sample of 122 patients was 20-71 years. The mean age was 42.5 with an SEM of 11.1. The distribution of patients (frequency-percentage) by age category was: 30 yrs, 19-15.6%; 31-40 yrs, 37-30.3%; 41-50 yrs, 40-32.8%; 51-60 yrs, 16.6%; 61-70 yrs, 6-4.9%; 71, 1-0.8%. A subsample of 74 patients reported their race and was comprised of 83% whites, 16% blacks, and 1% Asian.

Thirty-eight of 122 patients (31%) were diagnosed with a SCL by the “G.S.” method. There were 31 (25%) with a positive test for CA-IX expression, 48 (39%) were positive for H-HPV by HC2, 52 (46%) by RLA H-HPVG, and 65 (58%=65/112) by RLA HPV.

LCM analyses produced estimates of the associations of the “G.S.” and each biomarker with the latent true SCL status. These estimates are presented in Table 1 along with bootstrap standard errors and 95% confidence intervals. The estimates obtained in our previous paper by treating the histological diagnosis as the gold standard [5] also are presented to illustrate the impact of its imperfection on accuracy estimates (Table 1).

The estimated sensitivity and specificity of the imperfect “G.S.” were very good, 100% and 99%, respectively, with 95% confidence intervals 98% to 100% and 96% to 100%, respectively. Sensitivity and specificity estimates obtained from the LCM analyses for each biomarker were larger than those obtained by using the histology-based method as a “gold standard”. The differences generally were not notable, with the exception of the difference in specificity for the RLA HPV test (0.67 vs 0.61). The sensitivity of each of the three HPV-based tests also was excellent. The specificity of each was relatively weak, especially for RLA HPV. For this test we estimated a positive result for 33% of women with no SCL (*i.e.*, 1-specificity=0.33). HC2 H-HPV and RLA H-HPVG had 1-specificity rates of 13% and 20%, respectively. CA-IX had weak sensitivity (67%) and relatively strong specificity (93%). Still, an estimated 7% of women with no SCL were diagnosed with an SCL by CA-IX, while an estimated 32% of those who had a SCL were diagnosed with none. HC2 H-HPV had greater specificity than the other two HPV-based tests and at least equal sensitivity. Its Youden index was greater than that for CA-IX.

The similarity of results in the previous paragraph with those obtained by treating the histology-based test as a true gold standard is not surprising given that the modified LCM results showed that the “G.S.” was nearly perfect. This, of course, is not the case in many studies. More generally, biases in accuracy estimates due to imperfection of the reference test can be severe. The LCM results above are presented because, although similar to those of our previous study [5], they are less biased and to provide readers with a complete introduction to the results produced by modified LCM analysis. Results of the modified LCM analysis presented below are also part of that introduction but, more importantly, are new contributions to the literature; that is, age specific false-positive and false-negative rates of SCL in women with AGC-NOS. The answers to two new *post-hoc* questions: “Can we improve on the accuracy of HC2 H-HPV by combining it with CA-IX?” and “Can we develop a test based on a subset of the 13 high-risk HPV genotypes that improves on HC2 H-HPV or RLA H-HPVG?” also are presented.

The result of modeling prevalence of SCL as a function of age is illustrated in Figure 1. The prevalence peaks at almost 38% at 28-30 years of age and then declines steadily to about 6 to 7% at age 70 (Figure 1).

The 1-PPV and 1-NPV curves are shown for the histological diagnoses (*i.e.*, “G.S.”) in Figures 2 and 3, respectively. The sensitivity and specificity of “G.S.” are given in Table 1 and, together with Figures 2 and 3, show that the “G.S.” method of diagnosis is nearly a true gold standard, at least for women who are less than 50 years old. Estimated false-positive percentages ($100-100\times\text{PPV}$) increase from about 3-4% at age 50 to about 18% at age 70 (Figure 2). It should be noted here that there were only seven patients 61 years old and that results at these older ages are not precise. This sparseness of data at ages 61 is manifest in the very wide confidence bands after age 60 in Figure 2. The reader should keep this in mind throughout the remainder of this section.

1-PPV and 1-NPV plots also were produced for the HC2 H-HPV, RLA H-HPVG, RLA HPV, and CA-IX-based tests. An overall comparison of the three HPV-based tests showed that HC2 H-HPV had better sensitivity, specificity, and PPV than either RLA H-HPVG or RLA HPV. The 1-NPV curve for HC2 H-HPV was, for all intents and purposes, equal to those for RLA H-HPVG and RLA HPV. Thus, for the purposes of the current study, we chose HC2 H-HPV as the HPV-based test for further investigation. The 1-PPV and 1-NPV curves for HC2 H-HPV by age are presented in Figures 4 and 5.

The 1-PPV of HC2 H-HPV was about 0.17 – 0.20 at ages 20-30 years but increased with increasing age up to 70 years (See Figure 4). This means that 17, or more depending on age, out of every 100 HC2 H-HPV positive women will not have a SCL. Furthermore, based on the fitted curve the “false-positive” percentage is expected to be roughly 35% at age 55 and range up to 65% based on the upper 95% confidence band. The expected rate increases with age, and could exceed 65%, among women who are > 55 years of age. On the other hand, 1-NPV was uniformly low, ranging from about 0.0001 to 0.0011 over all ages (See Figure 5). Based on the upper 95% confidence band, at most 0.425% of AGC-NOS women who forego invasive excision procedures (*i.e.*, cone/LEEP biopsy, including the transformation zone, or hysterectomy) based on a negative HC2 H-HPV test have a SCL, and the expected percentage is <0.11% (*i.e.*, 1 in every 909 HC2 H-HPV negative women are expected to be true positive, and we can be 95% confident that this error rate is less than 1 in every 235 HC2 H-HPV negatives).

The root cause of the unacceptably high 1-PPV of the HC2 H-HPV test was its lack of specificity. While CA-IX showed a lack of sensitivity (68%), it had a high specificity (93%). Thus, the question arises “Can the information in HC2 H-HPV and CA-IX results be combined to develop a test that is both sensitive and specific?” A ROC analysis showed that the best combination of HC2 H-HPV and CA-IX results classified a woman as positive if HC2 H-HPV was positive and negative if HC2 H-HPV was negative, regardless of the CA-IX result. Thus, CA-IX added no detection accuracy beyond that attainable from HC2 H-HPV alone.

The answer to the second *post-hoc* question, “Can we develop a test based on a subset of the 13 high-risk HPV genotypes that improves on HC2 H-HPV or RLA H-HPVG?”, appears to be no. The logistic regression of histological diagnoses yes/no codes (1/0) for SCL on yes/no codes (1/0) for each of the 13 high-risk types resulted in a model that included nine significant types ($p < 0.05$): 16, 18, 33, 35, 39, 45, 52, 58, and 59. The test that was defined to be positive if any of these nine types was observed was found to have sensitivity 0.9586, specificity 0.8426, and Youden index 0.8012. The Youden index is less than that for HC2 H-HPV (0.8681) and only slightly higher than that for RLA H-HPVG (0.7946). All of the other six selected subsets had Youden indices within round-off error of that for HC2 H-HPV. Since these six were selected out of 8192 possible subsets based on relative performance, the results for them are likely to be sample specific overestimates of accuracy. Thus, we expect that no subset of the 13 high-risk HPV genotypes is better than HC2 H-HPV.

CONCLUSIONS

The current study was conducted as a GOG ancillary data study (GOG ADS-0925) in follow-up to GOG-0171. In this study, we address potential biases in diagnostic accuracy estimates and estimate age-specific false positive and false-negative rates of competing screening tests by applying newly developed modifications to latent class model analysis methods for assessing diagnostic test accuracy in the absence of true gold standard diagnoses. Differences between LCM and estimates of these parameters that were obtained by using the histology-based test as if it was a gold standard were small for HC2 H-HPV,

CA-IX, and RLA H-HPVG, validating the use of histologic evaluation as a near perfect reference when assessing the accuracy of these tests. The difference was also small for RLA HPV sensitivity but was notable for specificity, for which the usual estimate (0.61) varied from the LCM estimate (0.67) by 8.2%. Overall, our results show that HC2 H-HPV is a highly sensitive and moderately specific predictor of SCL.

It is not generally true, however, that estimates will be accurate when an imperfect “G.S.” is used. Valenstein [8] showed that there can be a substantial bias in sensitivity and specificity estimates. In the context of cervical cancer screening, Pretorius, *et al.* [22] argued: “Studies of VIA that used colposcopic directed biopsy as the gold standard require reevaluation”, because of imperfections in this “gold standard”. Gaffikin, *et al.* [9] observed notable differences between classical LCM estimates of sensitivity and specificity (0.973 and 0.807, respectively) and naïve estimates (0.643 and 0.638, respectively) obtained from a standard analysis with an imperfect reference test (colposcopy/biopsy). While this problem was not expected to be very consequential, because our histological evaluation was quite rigorous, our previous conclusions [5] needed confirmation by an analysis that addresses the issue; *i.e.*, LCM analysis. More importantly, age-specific false-positive and false-negative rates are needed and were produced by our modified LCM. The resulting estimates of age-specific prevalence, false-positive, and false-negative rates are the primary substantive contribution of the current study. In addition, our demonstration that classical LCM analysis can be modified to allow prevalence to be a function of covariates, such as age, is a novel contribution to the cancer screening test literature.

Before proceeding with a discussion of substantive contributions, we present conditions under which the modified LCM methods should be applied in order to avoid bias due to misclassification error in an imperfect reference test. They are: (1) No true gold standard available; (2) At least three screening test results are available; (3) The screening tests are independent given true disease status; and (4) There is a need/desire to model prevalence, 1-PPV and 1-NPV as functions of covariates. It should be noted that the conditional independence requirement in (3) cannot be tested empirically and must be assumed based on expert knowledge of the tests under consideration and the etiology of the disease. In the current application, we deemed it reasonable to assume that the histology-based test, CA-IX, and each HPV-based test are mutually independent given true disease status. This is not a reasonable assumption for HC2 H-HPV, RLA H-HPVG, and RLA HPV, however; which are clearly not independent given disease status. In recognition of this conditional dependence among the HPV-based tests, we performed three separate modified LCM analyses, each with three tests: the “G.S.”, CA-IX, and one of the HPV tests; for each of the three HPV-based tests. It is important that future investigators give careful thought to whether this assumption is satisfied in their application of LCM or modified LCM analyses.

Our presentation of SCL prevalence, false positive and false negative rates as functions of age in the population of AGC-NOS women are of clinical and scientific importance. The false negative rate curve is particularly important to younger patients in this high-risk population who may wish to preserve fertility. The fact that the upper 95% confidence bound on 1-NPV for the HC2 H-HPV test (*i.e.*, the false-negative rate for HC2 H-HPV) is less than 0.0043 at all ages (see Figure 5) means that an HPV-negative woman can be 95% confident that her chances of having a SCL is less than 0.43% and that the risk is likely to be less than 0.11%, even if she is at the age of peak risk (28-30 years). She can therefore feel safe in foregoing invasive treatment in favor of follow up testing. On the scientific side, the false-positive curve in Figure 4 suggests that future research is needed to search for additional biomarkers that will distinguish true positives from false positives to avoid over-treatment of test-positive women. A follow-up study to GOG-0171 (GOG-0237) is in progress to determine whether biomarkers such as p16, Ki-67, or MCM2 expression can be

used to develop a test with a reduced number of false positive test results. The general objective of the follow-up study is to develop tests that improve on the ability of HC2 H-HPV to distinguish true-positive test results from false-positives among women who are test-positive, thus lowering 1-PPV and increasing specificity, without decreasing sensitivity.

The relatively low specificity and high 1-PPV of HC2 H-HPV may be due to cross-reactivity with at least 15 untargeted low-risk genotypes [23-26]. Thus, we compared the specificity and 1-PPV of HC2 H-HPV to those of RLA H-HPVG to determine whether the latter might produce superior results. It did not, suggesting that RLA H-HPVG detection of at least some of the 13 specific high-risk genotypes also may be affected by cross-reactivity with low-risk types. The specificity of RLA H-HPVG was lower and the 1-PPV was generally higher across all ages than for HC2 H-HPV, while the sensitivities and 1-NPVs were essentially the same for these two tests. The same comparisons of HC2 H-HPV to RLA HPVG results showed similar but more pronounced differences due to the fact that the latter targets 24 low-risk HPV genotypes in addition to the 13 high-risk types. Clearly, cross-reactivity with low-risk types or inclusion of low-risk types as targets of the HPV test decreases the accuracy of the test. An interesting question then, is “Can a diagnostic test with better specificity and 1-PPV be derived from RLA H-HPV results by reducing the number of high-risk genotypes used to define a positive result to some subset of the most high-risk types?”. The results of our analyses of tests based on subsets of HPV high-risk genotypes show that the answer to this question is “no”. Reducing the number of high-risk subtypes on which an HPV test is based appears not to improve performance of the test.

The results for HC2 H-HPV more closely approximated those for the “G.S.” than did those of any other biomarker considered. CA-IX was strong where HC2 H-HPV was weak (i.e., specificity of 0.93 compared with 0.87 in Table 1) and weak where HC2 H-HPV was strong (i.e., sensitivities of 0.68 and 1.00, respectively). The combination of CA-IX with HC2 H-HPV testing, however, did not improve the accuracy of detecting cervical neoplasia in women with an AGC-NOS diagnosis over that of HC2 H-HPV testing alone.

In summary, the histology-based “G.S.” was validated as a near perfect reference test. Among the four biomarker tests for SCL considered, HC2 H-HPV alone is recommended as the best non-invasive alternative to histological diagnosis, which involves either a highly invasive cone or LEEP biopsy of the cervix or a hysterectomy. The accuracy of HC2 H-HPV was not enhanced by the optimal test obtained from CA-IX results in combination with HC2 H-HPV, and using subsets of the 13 high risk HPV genotypes to develop tests did not improve on HC2 H-HPV. Most importantly, in terms of novel contributions, the study presents age-specific SCL prevalence rates, false-positive rates (1-PPV) and false-negative rates (1-NPV) These age-specific risks have both clinical value to women with AGC-NOS and their physicians, who must decide between invasive and non-invasive treatment alternatives, and scientific value to scientists who wish to develop screening tests that have low false-negative and false-positive rates at all ages in this or in more general populations. The modified LCM analyses employed in the current study are novel to the screening test literature and offer improvements over methods that ignore potential misclassifications by a reference test, provided that the conditional independence assumption discussed above is valid. When possible, future studies should be designed to allow the use of these methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The following member institutions participated in this study: Abington Memorial Hospital, Walter Reed Army Medical Center, University of Mississippi Medical Center, University of California at Los Angeles, University of Pennsylvania Cancer Center, University of Cincinnati, University of Texas Southwestern Medical Center at Dallas, Wake Forest University School of Medicine, University of California Medical Center at Irvine, Tufts-New England Medical Center, The Cleveland Clinic Foundation, SUNY at Stony Brook, Washington University School of Medicine, Cooper Hospital/University Medical Center, Columbus Cancer Council, Fox Chase Cancer Center, Women's Cancer Center – University of Nevada, University of Oklahoma, Tacoma General Hospital, Tampa Bay Cancer Consortium, Gynecologic Oncology Network/Brody School of Medicine, Ellis Fischel Cancer Center, Fletcher Allen Health Care, University of Wisconsin Hospital, University of Texas-Galveston.

This study was supported by National Cancer Institute grants to the Gynecologic Oncology Group Administrative Office (CA 27469) and the Gynecologic Oncology Group Statistical and Data Center (CA 37517).

REFERENCES

1. Kennedy AW, Salmieri SS, Wirth SL, Biscotti CV, Tuason LJ, Travarca MJ. Results of the clinical evaluation of atypical glandular cells of undetermined significance (AGCUS) detected on cervical cytology screening. *Gynecol Oncol.* 1996; 1996; 3:14–18. [PubMed: 8898161]
2. Wilbur DC. Endocervical glandular atypia: A “new” problem for the cytologist. *Diagn Cytopathol.* 1995; 13:463–469. [PubMed: 8834321]
3. Lee KR, Manna EA, St John T. Atypical endocervical glandular cells: Accuracy of cytologic diagnosis. *Diagn Cytopathol.* 1995; 13:202–208. [PubMed: 8575278]
4. Nasu I, Meurer W, Fu YS. Endocervical glandular atypia and adenocarcinoma: A correlation of cytology and histology. *Int J Gynecol Pathol.* 1993; 12:208–211. [PubMed: 8344758]
5. Liao SY, Rodgers WH, Kauderer, et al. Carbonic anhydrase IX and human papillomavirus as diagnostic biomarkers of cervical dysplasia/neoplasia in women with a cytologic diagnosis of atypical glandular cells: A Gynecologic Oncology Group study in United. *Int J Cancer.* 2009; 125:2434–2440. [PubMed: 19670419]
6. Schnatz PF, Guile M, O'Sullivan DM, Sorosky JL. Clinical significance of atypical glandular cells on cervical cytology. *Obstet Gynecol.* 2006; 107:701–708. [PubMed: 16507944]
7. Sieber AG, Massuger LFAG, Bulten J. Referral compliance, outcome and predictors of CIN after repeated borderline cervical smears in the Netherlands. *Cytopathology.* 2007; 18:96–104. [PubMed: 17397494]
8. Valenstein P. Evaluating diagnostic tests with imperfect standards. *Am J Clin Pathol.* 1990; 93:252–258. [PubMed: 2405632]
9. Gaffikin L, McGrath JA, Arbyn M, Blumenthal PD. Visual inspection with acetic acid as a cervical cancer test: accuracy validated using latent class analysis. *BMC Medical Research Methodology.* 2007; 7:36. [PubMed: 17663796]
10. Myers ER, McCrory DC, Nanda K, Bastian L, Matchar DB. Mathematical model for the natural history of Human Papillomavirus infection and cervical carcinogenesis. *Am J Epidemiol.* 2000; 156:1158–1171. [PubMed: 10905528]
11. Liao SY, Brewer C, Závada J, et al. Identification of the MN antigen as a diagnostic biomarker of cervical intraepithelial squamous and glandular neoplasia and cervical carcinoma. *Am J Pathol.* 1994; 145:598–609. [PubMed: 8080042]
12. Liao SY, Stanbridge EJ. Expression of the MN antigen in cervical Papanicolaou smears is an early diagnostic biomarker of cervical dysplasia. *Cancer Epidemiol Biomarkers Prev.* 1994; 5:549–557. [PubMed: 8827360]
13. Gravitt PE, Schiffman M, Solomon D, Wheeler CM, Castle PE. 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.* 2008; 17:1248–1254. [PubMed: 18483347]
14. Castle PE, Sadorra M, Garcia F, Holladay EB, Kornegay J. Pilot study of a commercialized human papillomavirus (HPV) genotyping assay: comparison of HPV risk group to cytology and histology. *J Clin Microbiol.* 2008; 44:3915–3917. [PubMed: 16971652]

15. Lazarsfeld, P.; Henry, N. Latent structure analysis. Houghton Mifflin Co; Boston: 1968.
16. Walter SD, Irwig LM. Estimation of error rates, disease prevalence and relative risk from misclassified data: a review. *J Clin Epidemiol.* 1988; 41:923–938. [PubMed: 3054000]
17. Uebersax JS, Grove WM. Latent class analysis of diagnostic agreement. *Stat Med.* 1990; 9:559–572. [PubMed: 2190288]
18. Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm. *J Royal Stat Soc: Series B.* 1977; 39:1–38.
19. Wei GCG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J Am Stat Assoc.* 1990; 85:699–704.
20. Bland, M. An introduction to medical statistics. 3rd ed. Oxford University Press; Oxford, UK: 2000.
21. Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Chapman and Hall/CRC; Boca Raton, FL: 1993.
22. Pretorius RG, Bao YP, Belinson JL, Burchette RJ, Smith JS, Qiao YL. Inappropriate gold standard bias in cervical cancer screening studies. *Int J Cancer.* 2007; 121(10):2218–2224. [PubMed: 17657715]
23. Poljak M, Marin IJ, Seme K, Vince A. Hybrid Capture II HPV test detects at least 15 human papillomavirus genotypes not included in its current high risk cocktail. *J Clin Virol.* 2002; 25(Suppl. 3):S89–S97. [PubMed: 12467782]
24. Castle PE, Gravitt PE, Solomon D, Wheeler CM, Schiffman M. Comparison of linear array and line blot assay for detection of human papillomavirus and diagnosis of cervical precancer and cancer in the atypical squamous cell of undetermined significance and low-grade squamous intraepithelial lesion triage study. *J Clin Microbiol.* 2008; 46:109–117. [PubMed: 17989194]
25. Castle PE, Schiffman M, Burk RD, et al. Restricted cross-reactivity of hybrid capture 2 with non-oncogenic human papillomavirus types. *Cancer Epidemiol Biomarkers Prev.* 2002; 11:1394–1399. [PubMed: 12433717]
26. Castle PE, Solomon D, Wheeler CM, Gravitt PE, Wacholder S, Schiffman M. Human papillomavirus genotype specificity of Hybrid Capture 2. *J Clin Microbiol.* 2008; 46:2595–2604. [PubMed: 18579716]
27. Liao SY, Rodgers WH, Kauderer J, et al. Carbonic anhydrase IX (CA-IX) and high-risk human papillomavirus (H-HPV) as diagnostic biomarkers of cervical dysplasia/neoplasia in Japanese women with a cytologic diagnosis of atypical glandular cells (AGC): a Gynecologic Oncology Group (GOG) Study. *Br J Cancer.* 2010; 104:353–360. [PubMed: 21157448]
28. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papilloma virus and cervical cancer. *Lancet.* 2007; 370:890–907. [PubMed: 17826171]
29. Kinney W, Stoler MH, Castle PE. Patient safety and the next generation of HPV DNA tests. *Am J Clin Pathol.* 2010; 134:193–199. [PubMed: 20660320]
30. Castle PE, Katki HA. Benefits and risks of HPV testing in cervical cancer screening. *Lancet Oncol.* 2010; 11:214–215. [PubMed: 20089448]
31. Arbyn M, Kyrgiou M, Simoons C, et al. Perinatal mortality and other severe adverse pregnancy outcomes associated with treatment of cervical intraepithelial neoplasia: meta-analysis. *BMJ.* 2008; 337:a1284. [PubMed: 18801868]

Prevalence of SCL among women with AGC-NOS

Figure 1.

Prevalence of SCL among women with AGC-NOS by age with 95% confidence bands and observed proportions in 5 age categories.

1-PPV of histological ("G.S.") diagnoses

Figure 2.

Proportion of false-positive test results by age with 95% confidence bands for the diagnostic test based on histological evaluation of tissue from a cone or LEEP biopsy or hysterectomy.

1-NPV of histological ("G.S.") diagnoses

Figure 3.

Proportion of false-negative test results by age with 95% confidence bands for the diagnostic test based on histological evaluation of tissue from a cone or LEEP biopsy or hysterectomy.

1-PPV of HC2 H-HPV diagnoses

Figure 4.

Proportion of false-positive test results by age with 95% confidence bands for the diagnostic test based on Hybrid Capture II assessment of HPV infection with one or more of 13 HPV types.

1-NPV of HC2 H-HPV diagnoses

Figure 5.

Proportion of false-negative test results by age with 95% confidence bands for the diagnostic test based on Hybrid Capture II assessment of HPV infection with one or more of 13 HPV types.

Table 1

Sensitivity and specificity of competing diagnostic tests for SCL

	"G.S."	CA-IX	HC2 H-HPV	RLA H-HPVG	RLA HPVVG
Sensitivity					
Liao, <i>et al.</i> est.	--	0.66	0.97	0.97	0.97
LCM estimate	1.00	0.68	1.00	1.00	1.00
Bootstrap SE	0.02	0.08	0.00	0.00	0.00
95% CI	(0.98, 1.00)	(0.51, 0.83)	(1.00, 1.00)	(1.00, 1.00)	(1.00, 1.00)
Specificity					
Liao, <i>et al.</i> est.	--	0.93	0.87	0.79	0.61
LCM estimate	0.99	0.93	0.87	0.80	0.67
Bootstrap SE	0.01	0.03	0.04	0.05	0.06
95% CI	(0.96, 1.00)	(0.86, 0.98)	(0.79, 0.94)	(0.70, 0.92)	(0.54, 0.86)