Published in final edited form as:

J Clin Epidemiol. 2014 March ; 67(3): 343–353. doi:10.1016/j.jclinepi.2013.09.013.

A New Method to Address Verification Bias in Studies of Clinical Screening Tests: Cervical Cancer Screening Assays as an Example

Xiaonan Xue, **Mimi Y Kim**, **Philip E Castle**, and **Howard D Strickler**

Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY

Abstract

Objective—Studies to evaluate clinical screening tests often face the problem that the "gold standard" diagnostic approach is costly and/or invasive. It is therefore common to verify only a subset of negative screening tests using the gold standard method. However, under-sampling the screen-negatives can lead to substantial overestimation of the sensitivity and underestimation of the specificity of the diagnostic test. Our objective was to develop a simple and accurate statistical method to address this "verification bias".

Study Design and Setting—We developed a weighted generalized estimating equation approach to estimate, in a single model, the accuracy (e.g., sensitivity/specificity) of multiple assays as well as simultaneously compare results between assays while addressing verification bias. This approach can be implemented using standard statistical software. Simulations were conducted to assess the proposed method. An example is provided using a cervical cancer screening trial that compared the accuracy of human papillomavirus and Pap tests, with histological data as the gold standard.

Results—The proposed approach performed well in estimating and comparing the accuracy of multiple assays in the presence of verification bias.

Conclusion—The proposed approach is an easy to apply and accurate method for addressing verification bias in studies of multiple screening methods.

Keywords

clinical screening tests; sensitivity; specificity; weighted generalized estimating equations; verification bias

1. Introduction

The most accurate clinical methods to diagnose a disease or its precursor state are often not practical for routine screening purposes in the general population. Biopsy and histology, for example, are the gold standard for the diagnosis of many types of cancer, liver fibrosis, and

Correspondence to: Xiaonan Xue, Department of Epidemiology & Population Health, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Belfer 1303C, Bronx, NY 10461, Tel. (718)430-2431; Fax. (718)430-8780, Xiaonan.xue@einstein.yu.edu.

renal disease, but can cause pain as well as infection and bleeding, which on rare occasions may result in significant physical harm. The high cost of certain types of testing may also make them impractical for routine use. Therefore, screening tests that are less invasive and/or less expensive than the gold standard methods are commonly employed in disease surveillance and to triage patients for more definitive testing. However, the sensitivity, specificity, positive (PPV) and negative (NPV) predictive value of these screening tests must first be established by comparing them to a gold standard.

If the disease is uncommon, only a small subset of those negative using a screening test will have the disease. It may therefore be unethical and/or inefficient to request that every subject undergo a more expensive and invasive gold standard procedure for verification of their disease status. Instead, a representative sample of screen-negatives might be selected to be verified with the gold standard. However, this results in over-representation of subjects with positive screening tests in the sample and may bias estimates of sensitivity and specificity (i.e., verification bias) (1–2).

Several statistical methods to correct for verification bias have been proposed. Begg and Greenes (3), for example, developed a bias-correction procedure that assumes that the disease prevalence estimated in the subset of screen-negative individuals who undergo verification applies to all screen-negatives. Empirical methods such as bootstrapping (4) are commonly employed to estimate the confidence intervals (CIs) for these bias-corrected estimates of sensitivity and specificity. Zhou (5) allowed for a more general verification selection procedure and developed a maximum likelihood (ML) approach to estimate sensitivity and specificity as well as corresponding CIs. Neither of these approaches, however, can be used to directly compare the performance of different assays.

To date, a few statistical methods able to directly compare two or more diagnostic tests while addressing verification bias have been reported. The McNemar test can be used to compare two screening tests when only a subset of subjects who are negative on both tests undergo disease verification (6). Zhou developed a non-parametric ML method for comparing sensitivity and specificity (7–8), as well as an ML approach to compare PPVs and NPVs (9). Alonzo (10) proposed a ML estimator for relative sensitivity and relative specificity that addresses verification bias. However, the complexity of these ML methods and the computational difficulty in implementing them with standard software has restricted their use.

The limited use of these methods is reflected in recent studies of cervical cancer screening assays. Our review of 15 recent comparative cervical screening trials which performed colposcopy (gold standard) in a subset of screen-negatives (11–25), only one study used Zhou's likelihood method (7) to compare the sensitivity and specificity of the assays (Table 1). A simple and easily implemented statistical method is therefore needed to estimate and compare the performance of screening assays while accounting for verification bias.

2. Methods

2.1. Overview

Subjects who did not undergo the gold standard procedure to verify their disease status can be viewed as having missing data (7, 26). Therefore, methods for incomplete data can be adapted to correct for verification bias. Weighted generalized estimating equation models, a direct extension of the generalized estimating equation approach, were proposed for the analysis of correlated categorical data with missing data and censoring (27–29). Herein we propose a similar method to compare screening assays when the majority of the screen negatives are not verified and have missing gold standard case status. Specifically, we propose incorporating a weight determined by the inverse of the probability of a subject undergoing the gold standard procedure. This approach allows for simultaneous estimation of (i) sensitivity and specificity, (ii) PPV and NPV, (iii) differences in these parameters between multiple screening tests, and iv) 95% CIs for these parameters. Moreover, the proposed statistical approach can be readily implemented with commonly available software.

2.2. Weighted Generalized Estimating Equation Models to Address Verification Bias

Consider a study to assess and compare two screening tests (e.g., Pap and HPV DNA tests). Let T_{ij} denote the test result for the ith person (i= 1,..., N) with the jth test (j=1,2) where T_{ij} = 1 for a positive test result and 0 otherwise. We model P_{ij} = P(T_{ij} =1) by

$$\log it P_{ij}=\beta_0+\beta_1 test_j+\beta_2 Y_i+\beta_3 Y_i test_j \quad (1)$$

where $test_j$ is an indicator variable equal to 0 for j=1 and 1 for j=2 and Y_i = 1 if the ith person has disease and 0 otherwise. Since each subject contributes two test results, the results can be correlated. A generalized estimating equation analysis can be used to account for the correlation (30) with either an independent or exchangeable working correlation structure. If all subjects undergo the gold standard test to verify their diagnoses, then no weighting is necessary. However, if only a random sample of subjects with negative results in both screening tests undergo the gold standard test, then a weighted generalized estimating equation analysis should be used in which the subjects in the random sample are given a weight equal to the inverse of the sampling fraction and the others given a weight of 1 in the estimation procedure. For example, if 10% of screen-negative subjects undergo the gold standard testing to verify their diagnoses, these subjects receive a weight of 10 (i.e., each subject in this subset is representative of 10 screen-negative subjects). The sensitivity, specificity and the diagnostic likelihood ratios (DLR) for each test along with their 95% CIs can then be directly estimated from model (1). In addition, model (1) allows for direct comparison of the sensitivity and the specificity between any two tests as described further in Appendix A.

To estimate predictive values, we model μ_i = P(Y_i = 1) by log it μ_i = θ_0 + $\theta_1 test_j$ + $\theta_2 T_{ij}$ + $\theta_3 T_{ij} test_j$ (2). This model also allows direct comparison of PPVs and NPVs for different methods. When all subjects with at least one test positive are referred for diagnostic verification by the gold standard, the PPV can be estimated from the verified subjects without any adjustment. The NPV, however, needs to be estimated with the proper weight incorporated.

Odds ratios of sensitivity, specificity, PPV or NPV between the two tests provided in models (1–2) are less intuitive measures than the relative sensitivity, relative false positive fraction (FPF) or specificity and relative predictive values. We therefore also consider using a log link for P_{ij} and μ_{ij} (models (3–5)). Model convergence may be difficult to achieve with the log link because log(P) and log(μ) are less than 0. We therefore adopted a modified Poisson regression model, the validity of which has recently been demonstrated for correlated binary data (31).

Appendix A describes details of each model. These models can also be generalized to incorporate more than two test results.

2.3. Addressing Testing Non-Compliance and Inadequate Test

Results

The above models can also be used to account for potential subject non-compliance in undergoing tests and inadequate test results, whether among screen-positives or screen negatives. For example, in practice, not every subject who is referred for diagnostic verification using the gold standard assay will be compliant with that referral and some subjects who are tested may not have an adequate test result. Therefore, verification bias can also occur due to non-compliance and missing data even if all subjects are referred for gold standard testing. For example, in Mayrand (24), the sampling fraction for a study subject depended not only on her Pap and HPV test results, but also on the screening arm to which she was randomly assigned as missing due to non-compliance and inadequate results differed by screening arms.

2.4. Simulation Studies

We used simulations to evaluate the performance of our proposed method. We generated a sample of 5,000 subjects with a disease prevalence of 5%, two correlated binary screening tests (32) with a range of possible values for sensitivity and specificity, and a gold standard diagnostic test with 100% accuracy (see details in Appendix B). In the first scenario, all individuals positive in either or both of the screening tests and a randomly selected 10% of individuals who tested negative in both screening tests were assumed to have their disease status verified by the gold standard. Models (1)–(5) were applied to each simulated data set using an independent working correlation.

We next considered the real-world situation in which not only do we lack complete gold standard test results for screen-negative patients but also from the screen-positive patients who were referred for the gold standard test but either failed to comply or had inadequate test results. Specifically, in each simulated data set, we assumed 10% of subjects with negative results in both screening tests, 90% of those who had positive results in only the first screening test, 80% of those who had positive results in only the second screening test and all the subjects with positive results on both screening tests, had undergone diagnostic verification and had adequate gold standard test results. Log link models (3)–(5) were applied to each simulated dataset.

These models were fit using both a weighted and un-weighted generalized estimating equation approach. The latter approach was evaluated as the basis for comparison since it provides results that are unadjusted for verification bias, and when used for comparing two screening tests, is equivalent to McNemar’s test (33). Thus, our results help to highlight the possible limitations of the McNemar test when comparing screening tests in the presence of verification bias (20, 21).

The simulation was repeated 500 times. The performance of each method was summarized according to % bias and % coverage for sensitivity, specificity, PPV, NPV and DLR and empirical power or empirical type I error when comparing two tests.

2.5. Cervical Cancer Screening Example

The Canadian Cervical Cancer Screening Trial (CCCaST) was designed to compare HPV DNA testing and Pap tests as stand-alone screening assays for the detection of cervical pre-cancer and cancer among women ages 30–69 years who presented for routine screening (24). A total of 10,154 women were randomly assigned 1:1 to a “focus on Pap” or a “focus on HPV” screening arm: the women received a Pap test first and HPV test next in the former group whereas the women received a HPV test first and Pap next in the latter group (34).

Coloposcopy was recommended for a random sample of 10% of participants who had a normal Pap and negative HPV test from each screening arm and all women who were positive in either or both assays. A histological diagnosis of pre-cancer/cancer in these specimens was considered the gold standard diagnosis. For the women with a positive Pap and/or HPV test result who failed to comply with their referral to colposcopy, we assumed this was unrelated to their unobserved disease status.

Within each arm, the investigators calculated corrected estimates of sensitivity and specificity using Begg and Green's (3) method and 95% CIs with Zhou's method (5). Since test performance was similar in both arms of the study, the data were combined to obtain overall estimates of sensitivity, specificity, PPV and NPV. However, 95% CIs were not provided since Zhou's (5) method cannot be readily applied with variable compliance rates across screening arms. A z-test was used to test the difference in sensitivity and specificity between the HPV and Pap tests with the Pap test performance estimated using only data from the "focus on Pap" arm and the HPV test performance estimated using only data from the "focus on HPV" arm. Although this statistical approach was inefficient given that women in each arm underwent both HPV and Pap tests, existing methods could not be used to readily combine data across the arms and directly assess the significance of any differences between the HPV versus Pap tests.

We re-analyzed the CCCaST data using the proposed method to demonstrate how the new method can: (i) be used to efficiently summarize data across different study arms, and (ii) address non-compliance with referral to the gold-standard test and other sources of missing data, which are additional forms of verification bias.

3. Results

3.1. Simulation Study

Results using a logit link function (Tables 2) show that the proposed method provides valid point and interval estimates for sensitivity, specificity, PPV, NPV, and DLR as well as for ORs comparing the performance between two screening tests as measured by % bias and % coverage; the proposed method also provides valid statistical inferences for comparing the two tests since the empirical significance level does not exceed 5%. Similar results were observed using a log-link function (Table 3) where RRs instead of ORs were used for comparing the performance of the two screening tests. As mentioned previously, the major advantage to using the log link rather than logit function is that relative sensitivity and specificity (or FPF) are easier to interpret than ORs. We thus recommend using the logit link if the statistical comparison between tests is of primary interest but the log link if the magnitude of the difference between tests is of primary interest.

The proposed method also yielded unbiased estimates of sensitivity, specificity, PPV, NPV, DLR and the differences in these parameters between the two screening tests in the real-world situation in which some subjects are non-compliant with their referral to undergo gold standard testing (Table 4).

Estimates obtained using the standard generalized estimating equation approach were in general biased (results reported in Appendix C), since the method does not account for verification bias. However, when all positive tests were verified, absolute and relative PPV estimates were unbiased; relative sensitivity and relative FPF estimates obtained using a log-link function were also valid since each parameter estimate was inflated by the same factor (35). Furthermore, the standard generalized estimating equation approach achieves similar levels of empirical power/type I error as the proposed method when comparing sensitivity, specificity (or FPF) and PPV. This result agrees with Schatzkin et al.'s finding that the

McNemar test is valid for comparing sensitivity, specificity and PPV when all positive screening tests are verified (6). However, the McNemar test is no longer valid in the presence of non-compliance and missing data.

3.2. Cervical Cancer Screening Example

Using an independent working correlation and a log link function (model (4)), our estimates of sensitivity, specificity, PPV and NPV for HPV DNA and Pap tests match those of Mayrand et al (24). We also determined 95% CI for each of these measures, and estimated DLRs and their 95% CIs. We then compared each performance measure between the screening tests using all test data from both study arms, whereas the statistical methods in the original study used only partial data. As shown in Table 5, HPV had significantly better sensitivity and negative DLR but worse specificity, PPV and positive DLR. We also provided the estimates of their difference and their CIs. For example, the relative sensitivity between HPV and PAP is estimated to be 1.72 (CI: 1.30, 2.29) and relative specificity is estimated to be 0.97 (CI: 0.96, 0.98). CIs could not be determined in the original paper using existing statistical methods. Thus, our method demonstrates that the HPV test is as specific as the Pap test and can be up to two-fold more sensitive.

4. Discussion

Although methods to directly compare the performance of different screening tests that are subject to verification bias have been described in the statistical literature, they have not been widely adopted in cervical cancer screening trials since existing methods, including the ML methods, are complicated and difficult to implement using standard statistical software.

We have proposed a conceptually simple method based on a weighted generalized estimating equation approach; an approach originally proposed to address missing data. Specifically, we generalized this method to estimate and compare the performance of binary diagnostic tests when only a random subset of screen-negatives undergo diagnostic verification with a gold standard. The proposed method can also address verification bias due to the real-world problem that not every subject referred to gold standard testing actually complies and/or obtains an adequate test result.

Simulation studies showed that the new method can accurately estimate the sensitivity and specificity, PPV and NPV, DLR of multiple individual assays, calculate the corresponding 95% CIs, and simultaneously determine the differences in performance measures between any two screening tests examined in the model. Using an example dataset, we also showed that the proposed method made it possible to combine data from two or more study arms, which could not be accomplished using prior statistical methods. Lastly, the new approach can be readily implemented with commonly used statistical software such as SAS, R and STATA (see Appendix D).

An important underlying assumption of the new model is that diagnostic verification depends only on screening test results and not on any other variables, i.e., the missing at random assumption. However, verification might sometimes depend on either recorded or unrecorded variables which are related to disease state (26, 36–37) so that a separate regression equation is needed to model the weight(s) (38). While extension of the proposed approach to address these more sophisticated situations of disease verification is of interest, the development of a simple and straightforward computational method for this purpose will be a challenge. Another assumption of our model involves the large sample normal approximation and, thus, the proposed method may not be appropriate for small studies (39, 40).

In conclusion, we developed a straightforward approach to estimate and compare the accuracy of two or more screening tests in the presence of verification bias, which can be readily implemented with standard software. The development of this method is timely since new screening tests for cancer continue to be developed based on DNA methylation, tissue microRNA expression levels, proteomics, metabolomics, and other not-yet-imagined molecular assays. The performances of these screening tests need to be examined and compared to existing methods to determine if the new tests significantly improve diagnostic accuracy. We hope that the proposed statistical method will be adopted to facilitate these comparisons and to improve the efficiency and validity of clinical screening studies.

Acknowledgments

This work was supported by NCI R01 CA174634 and NCI R01 CA85178.

References

1. Mower WR. Evaluating Bias and Variability in Diagnostic Test Reports. Annals of Emergency Medicine. 1999; 33:85–91. [PubMed: 9867892]
2. Zhou, XH.; Obuchowski, NA.; McClish, DK. Statistical Methods in Diagnostic Medicine. Vol. Chp 10. Wiley; New Jersey: 2011. p. P329
3. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subjects to selection bias. Biometrics. 1983; 39:207–215. [PubMed: 6871349]
4. Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall; London: 1993.
5. Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Communication in Statistics-Theory and Methods. 1993; 22:3177–98.
6. Schatzkin A, Connor RJ, Taylor PR, Bunnag B. Comparing new and old screening tests when a reference procedure cannot be performed on all screenees. Example of automated cytometry for early detection of cervical cancer. AJE. 1987; 25:672–678.
7. Zhou XH. Comparing accuracies of two screening tests in a two-phase study for dementia. Appl Statist. 1998; 46:135–147.
8. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. Statistical Methods in Medical Research. 1998; 7:337–353. [PubMed: 9871951]
9. Zhou XH. Effect of verification bias on positive and negative predictive values. Statistics in Medicine. 1994; 13:1737–1745. [PubMed: 7997707]
10. Alonzo TA. Verification bias-corrected estimators of the relative true and false positive rates of two binary screening tests. Statistics in Medicine. 2005; 24(3):403–417. [PubMed: 15543634]
11. Kulasingam KL, Hughes JP, Kiviat NB, Mao C, Weiss NS, Kuypers JM, Koutsky LA. Evaluation of Human Papillomavirus Testing in primary screening for cervical abnormalities: Comparison of sensitivity, specificity and frequency of referral. JAMA. 2002; 288:1749–1757. [PubMed: 12365959]
12. Almonte M, Ferreccio C, Winkler JL, Cuzick J, Tsu V, Robles S, Takahashi R, Sasieni P. Cervical screening by visual inspection, HPV testing, liquid-based and conventional cytology in Amazonian Peru. Int J Cancer. 2007; 121:796–802. [PubMed: 17437272]
13. Dalstein V, Riethmuller D, Sautiere JL, Pretet JL, Kantelip B, Schaal JP, Mougin C. Detection of cervical precancer and cancer in a hospital population benefits of testing for human papillomavirus. European Journal of Cancer. 2004; 40:1225–1232. [PubMed: 15110887]
14. Ferrecio C, Barriga MI, Lagos M, Ibanez C, Poggi H, Gonzalez F, et al. Screening trial of human papillomavirus for early detection of cervical cancer in Santiago, Chile. IJC. 2012 in press.
15. Mahmud SM, Sangwa-Lugoma G, Nasr SH, Kayembe PK, Tozin RR, Drouin P, Lorincz A, Ferenczy A, Franco EL. Comparison of human papillomavirus testing and cytology for cervical cancer screening in a primary health care setting in the Democratic Republic of the Congo. Gynecologic Oncology. 2012; 124:286–291. [PubMed: 22062546]
16. Gravitt PE, Paul P, Katki HA, Vendantham H, Ramakrishna G, Sudula M, Kalpana B, Ronnett BM, Vijayaraghavan K, Shah KV. Effectiveness of VIA, Pap, and HPV DNA Testing in a

Cervical Cancer Screening Program in a Peri-Urban Community in Andhra Pradesh, India. PLoS ONE. 2010; 5(10):e13711.10.1371/journal.pone.0013711 [PubMed: 21060889]

17. Li N, Shi J-F, Franceschi S, Zhang W-H, Dai M, Liu B, Zhang Y-Z, Li L-K, Wu R-F, Vuyst HD, Plummer M, Qiao Y-L, Clifford G. Different cervical cancer screening approaches in a Chinese Multicentre study. Br J Cancer. 2009; 100:532–537. [PubMed: 19127262]
18. Sarian LO, Derchain S, Shabalova I, Tatti S, Naud P, Longatto-Filho A, Syrjanen S, Syrjanen K. Optional screening strategies for cervical cancer using standalone tests and their combinations among low- and medium-income populations in Latin America and Eastern Europe. J Med Screen. 2010; 17:195–203. [PubMed: 21258130]
19. Petry KU, Menton S, Menton M, van Loenen-Frosch F, de Carvalho Gomes H, Holz B, Schopp B, Garbrecht-Buettner S, Davies P, Boehmer G, van den Akker E, Iftner T. Inclusion of HPV testing inroutine cervical cancer screening for women above 29 years in Germany: results for 8466 patients. Br J Cancer. 2003; 88:1570–7. [PubMed: 12771924]
20. Castle PE, Stoler MH, Wright TC Jr, Sharma A, Wright T, Behrens CM. Performance of caricinogenic human papillomavirus (HPV) testing and HPV16 or HPV 18 genotyping for cervical cancer screening of women aged 25 years and older: a subanalysis of the ATHENA study. Lancet Oncol. 2011; 12:880–890. [PubMed: 21865084]
21. Ratnam S, Franco EL, Ferenczy A. Human Papillomavirus testing for primary screening of cervical cancer precursors. CEBP. 2000; 9:945–951.
22. Clavel C, Masure M, Bory JP, Putaud I, Mangeonjean C, Lorenzato M, Nazeyrollas P, Gabriel R, Quereux C, Birembaut P. Human papillomavirus testing in primary screening for the detection of high-grade cervical lesions: a study of 7932 women. Br J Cancer. 2001; 89:1616–23. [PubMed: 11401314]
23. Cuzick J, Szarewski A, Cubie H, Hulman G, Kitchener H, Luesley D, McGoogan E, Menon U, Terry G, Edwards R, Brooks C, Desai M, et al. Management of women who test positive for high-risk types of human papillomavirus: the HART study. Lancet. 2003; 362:1871–6. [PubMed: 14667741]
24. Mayrand MH, Duarte-Franco E, Rodrigues I, Walter SD, Hanley J, Ferenczy A, Ratnam S, Coutlee F, Franco EL. Human Papillomavirus DNA versus papanicolaou screening tests for cervical cancer. The New England Journal of Medicine. 2007; 357:1579–1588. [PubMed: 17942871]
25. Monsonego J, Hudgens MG, Zerat L, Zerat J-C, Syrjanen K, Halfon P, Ruiz F, Smith JS. Evaluation of oncogenic human papillomavirus RNA and DNA test with liquid-based cytology in primary cervical cancer screening: the FASE study. Int J of Cancer. 2011; 129:691–701. [PubMed: 20941740]
26. Toledano AY, Gatsonis C. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. Biometrics. 1999; 55:488–496. [PubMed: 11318205]
27. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. JASA. 1995; 90:106–121.
28. Preisser JS, Lohman KK, Rathouz PJ. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. Statistics in Medicine. 2002; 21:3035–3054. [PubMed: 12369080]
29. Zeger SC, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 1988; 44:1049–1060. [PubMed: 3233245]
30. Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press; New York: 2003. p. P59
31. Zou GY, Donner A. Extension of the modified Poisson regression model to prospective studies with correlated binary data. Statistical Methods in Medical Research. 2012 in press.
32. Park CG, Park T, Shin DW. A simple method for generating correlated binary variables. The American Statistician. 1996; 50:306–310.
33. Leisenring W, Alonzo T, Pepe MS. A marginal regression modeling framework for evaluating medical diagnostic tests. Statistics in Medicine. 1997; 16:1263–81. [PubMed: 9194271]
34. Mayrand MH, Duarte-Franco E, Coutlee F, et al. Randomized controlled trial of human papillomavirus testing versus Pap cytology in the primary screening for cervical cancer precursors:

design, methods and preliminary accrual results of the Canadian Cervical Cancer Screening Trial (CCCaST). Int J Cancer. 2006; 119:615–23. [PubMed: 16572425]

35. Cuzick J, Clavel C, Petry KU, et al. Overview of the European and North American studies on HPV testing in primary cervical cancer screening. Int J Cancer. 2006; 119:1095–1101. [PubMed: 16586444]

36. Albert PS, Dodd LE. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. JASA. 2008; 103(481):61–73. [PubMed: 19802353]

37. Baker SG. Evaluating multiple diagnostic tests with partial verification. Biometrics. 1995; 51:330–337. [PubMed: 7539300]

38. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. JASA. 1994; 89:846–66.

39. Cheng H, Macaluso M, Hardin JM. Validity and coverage of estimates of relative accuracy. Annals of Epidemiology. 2000; 10:251–60. [PubMed: 10854959]

40. Obuchowski NA, Zhou ZH. Prospective studies of diagnostic test accuracy when disease prevalence is low. Biostatistics. 2002; 3(4):477–92. [PubMed: 12933593]

Appendix A: Details for Models (1)–(5)

We model *Pij* = *P*(*Tij* =1) by the following regression model:

$$\log itP_{ij}=\beta_0+\beta_1 test_j+\beta_2 Y_i+\beta_3 Y_i test_j \quad (1)$$

where $test_j$ = 0 for j = 1 and 1 for j = 2; Y_i = 1 if the ith person has disease and 0 otherwise and T_{i1} & T_{i2} are correlated. The weighted generalized estimating equation method defined below is used to estimate model (1):

$$U(\beta)=\sum_{i=1}^{N} f_i^{-1}\left(\frac{\partial P_{i1}}{\partial\beta},\frac{\partial P_{i2}}{\partial\beta}\right)V_i^{-1}(T_i-P_i)$$

where $V_i=A_i^{1/2}R_i(\alpha)A_i^{1/2}$ and $Cov(T_i) = A_i\phi$ and $R_i(\alpha)$ is the working correlation within T_i = (T_i, T_{i2}), i.e., the two test results from the same subject and φ is the dispersion parameter which is treated as a nuisance parameter here and f_i is the sample fraction of disease verification for the subject in his/her subset, $0 < 1\ f_i \leq 1$. Either independent or exchangeable working correlations can be assumed between test results.

Based on Model (1), the sensitivity for the jth test is

$$\frac{e^{\beta_0+(\beta_1+\beta_3)test_j+\beta_2}}{1+e^{\beta_0+(\beta_1+\beta_3)test_j+\beta_2}}$$

and the corresponding confidence interval can be obtained by first estimating the confidence interval for $\beta_0 + (\beta_1 + \beta_3)test_j + \beta_2$ based on the robust variance estimate for the βs obtained using the WGEE method and then taking the anti-logit transformation. The specificity for the jth test is

$$\frac{1}{1+e^{\beta_0+\beta_1 test_j}}$$

and its CI can be obtained similarly as described above.

Direct comparison of the sensitivity and the specificity between any two tests reduces to a Wald test of the coefficients or a linear function of coefficients in the model: comparison of the sensitivity of two different screening tests is equivalent to testing the significance of $\beta_1 + \beta_3$. The odds ratio of sensitivities of the two tests is $e^{\beta_1 + \beta_3}$. Comparison of specificities between the two tests is equivalent to testing for significance of β_1, where the odds ratio of their specificities is $e^{-\beta_1}$.

The diagnostic likelihood ratio (DLR), the ratio of the likelihood of the observed test result in the diseased versus non-diseased population, is

$$\frac{e^{\beta_2+\beta_3 test_j}}{1+e^{\beta_0+(\beta_1+\beta_3)test_j+\beta_2}}(1+e^{\beta_0+\beta_1 test_j})$$

for a positive test j and is

$$\frac{1+e^{\beta_0+\beta_1 test_j}}{1+e^{\beta_0+(\beta_1+\beta_3)test_j+\beta_2}}$$

for a negative test j. Their variance estimates can be obtained using the delta method.

We model $\mu_i = P(Y_i = 1)$ using the following logistic regression model:

$$\log it\mu_i=\theta_0+\theta_1 test_j+\theta_2 T_{ij}+\theta_3 T_{ij} test_j \quad (2)$$

Based on model (2), the PPV for the jth test is $\frac{e^{\theta_0+(\theta_1+\theta_3)test_j+\theta_2}}{1+e^{\theta_0+(\theta_1+\theta_3)test_j+\theta_2}}$, which is the same as the crude estimate of PPV. The NPV for the jth test is $\frac{1}{1+e^{\theta_0+\theta_1 test_j}}$. In addition to estimating the PPV and NPV, the model allows the comparison of PPVs between tests through testing the significance of $\theta_1 + \theta_3$ and the comparison between NPVs through testing the significance of θ_1.

To obtain relative sensitivity, relative specificity, the following model:

$$\log P_{ij}=\gamma_0+\gamma_1 test_j+\gamma_2 Y_i+\gamma_3 Y_i test_{ij} \quad (3)$$

is considered in which a log link function is used. The sensitivity for the jth test is then $e^{\gamma_0 + (\gamma_1 + \gamma_3)test_j + \gamma_2}$ and the FPF (false positive fraction=1-specificity) for the jth test is given by $e^{\gamma_0 + (\gamma_1 test_j}$. The relative sensitivity comparing the two screening tests is $e^{\gamma_1 + \gamma_3}$ and the relative false positive fraction is e^{γ_1}. If direct calculation of specificity is preferred, a model on the agreement between the disease status and the corresponding test result, i.e., a model on $\pi_{ij} = P(Y_i = T_{ij})$ can be used instead (30):

$$\log \pi_{ij}=\gamma_0+\gamma_1 test_j+\gamma_2 Y_i+\gamma_3 Y_i test_{ij} \quad (4)$$

so that relative specificity is e^{γ_1}.

Similarly, a log link can be used for the model of PPV and NPV. We model the test and disease agreement $\pi_{ij} = P(Y_i = T_{ij})$ as the following

$$\log\pi_{ij}=\delta_0+\delta_1 test_j+\delta_2 T_{ij}+\delta_3 T_{ij}\, test_j. \quad (5)$$

The PPV for the jth test is therefore $e^{\delta_0+ (\delta_1 + \delta_3)test_j + \delta_2}$ and the NPV for the jth test is $e^{\delta_0 + \delta_1 test_j}$. The comparison between PPVs is equivalent to testing the significance of $\delta_1 + \delta_3$ and the comparison between NPVs is then equivalent to testing the significance of δ_1.

Appendix B: Simulation of the data

In each simulation, a sample of 5,000 subjects with a disease prevalence of 5% were generated with two binary screening tests under a range of possible values for sensitivity and specificity for both tests (sen_1, sen_2, spe_1, spe_2), and a gold standard diagnostic test that was 100% accurate. Specifically, we assume

$$\begin{gathered}(T_1,T_2)|case{\sim}Bi\,\mathrm{var}\,iateBernoulli(1,sen_1,1,sen_2,\rho_1)\\ \text{and}\\ (T_1,T_2)|control{\sim}Bi\,\mathrm{var}\,iateBernoulli(1,1-spe_1,1,1-spe_2,\rho_2)\end{gathered}$$

where ρ_1 and ρ_2 are the correlations between the two screening test results for a case and a control, respectively. Here we set $\rho_1 = \rho_2 = 0.3$. Correlated binary variables were generated from correlated Poisson variables, where the correlated Poisson variables were expressed as a convolution of independent Poisson random variables (32).

Appendix C: Simulation Results of the Generalized Estimating Equation Approach without adjustment for verification bias

Table 1a

The performance of the weighted generalized estimating equation approach using a logit link (models (1)–(2))

[1] Test 1						Comparison					
sen_1		spe_1		dlr+$_1$ Bias	dlr−$_1$ Bias	or_{sen}			or_{spe}		
[2] Bias	Cov	Bias	Cov			Bias	Cov	Power	Bias	Cov	Power
sen1=sen2=0.50, spe1=0.75, spe$_2$=0.80											
41.7	0.0	−47.3	0.0	−41.5	11.3	−2.6	94.6	5.4	22.8	18.4	100
sen$_1$=0.50, sen$_2$=0.60, spe$_1$=0.75, spe$_2$=0.80											
32.3	0.4	−47.4	0.0	−45.2	28.0	33.4	79.2	76.6	23.0	18.0	100
sen$_1$=0.50, sen$_2$=0.60, spe$_1$=spe$_2$=0.75											
33.0	0.2	−41.9	0.0	−41.1	15.4	30.7	84.0	74.2	0.3	95.6	4.4

Test 1				Comparison					
ppv_1		npv_1		or_{ppv}			or_{npv}		
Bias	Cov	Bias	Cov	Bias	Cov	Power	Bias	Cov	Power
ppv$_1$=ppv$_2$=0.10, npv$_1$=0.96, npv$_2$=0.97									
0.5	97.6	−3.4	0.0	−0.5	95.8	4.2	107	48.2	99.8

Test 1				Comparison					
ppv_1		npv_1		or_{ppv}			or_{npv}		
Bias	**Cov**	**Bias**	**Cov**	**Bias**	**Cov**	**Power**	**Bias**	**Cov**	**Power**
ppv_1=0.06, ppv_2=0.07, npv_1=npv_2=0.96									
0.2	99.4	−1.3	54.8	−0.7	95.4	33.2	8.7	91.8	8.2

Note: Results based on 500 simulated data sets of size=5,000 among whom 10% of negative on both tests and all positive on any test obtained disease verification and model (1);

[1] As several performance measures are calculated for each test and the findings from the two tests are similar, for simplicity, only the results for test 1 and the comparison between test 1 and 2 are reported (test 2 data not shown).

[2] Bias(%) = (estimate-true value)/true value *100; Cov(%) =the percent that the estimated CI contained the true parameter; Power(%)=the percent that the estimated CI does not contain the null value: when the true parameter equals to the null value, this percent is the empirical type I error rate; when the true parameter does not equal to the null value, this percent is the empirical power.

Abbreviation: sen_1 and sen_2, sensitivity of test 1 and 2 respectively; spe_1 and spe_2, specificity of test 1 and 2 respectively; or_{sen}, odds of sensitivity of test 1 vs odds of sensitivity of test 2; orspe, odds of specificity of test 1 vs odds of specificity of test 2; $dlr+_1$ and $dlr-_1$, positive and negative diagnostic likelihood ratio for test 1 respectively; ppv_1 and ppv_2, positive predicative value of test 1 and 2 respectively; npv_1 and npv_2, negative predicative value of test 1 and 2 respectively; or_{ppv}, odds of PPV of test 1 vs odds of PPV of test 2; or_{npv}, odds of NPV of test 1 vs odds of NPV of test 2.

Table 2a

The performance of the generalized estimating equation approach using a log link (models (3) & (5))

Test 1						Comparison					
sen_1		fpf_1		$dlr+_1$ Bias	$dlr-_1$ Bias	rr_{sen}			rr_{fpf}		
Bias	**Cov**	**Bias**	**Cov**			**Bias**	**Cov**	**Power**	**Bias**	**Cov**	**Power**
sen_1=sen_2=0.50, fpf_1=0.25, fpf_2=0.20											
41.3	0.0	−47.3	0.0	−41.5	11.3	−0.8	94.0	6.0	−0.2	100	100
sen_1=0.50, sen_2=0.60, fpf_1=0.25, fpf_2=0.20											
32.3	0.2	142	0.0	−45.0	28.0	0.3	95.4	77.0	0.0	95.0	100
sen_1=0.50, sen_2=0.60, fpf_1=fpf_2=0.25											
33.0	0.4	126	0.0	−41.1	15.4	−0.4	93.4	76.2	0.0	97.0	3.0

Test 1				Comparison					
ppv_1		npv_1		rr_{ppv}			rr_{npv}		
Bias	**Cov**	**Bias**	**Cov**	**Bias**	**Cov**	**Power**	**Bias**	**Cov**	**Power**
ppv_1=ppv_2=0.10, npv_1=0.96, npv_2=0.97									
−0.1	98.0	−3.5	0.0	−0.3	97.0	3.0	−6.4	8.0	99.8
ppv_1=0.06, ppv_2=0.07, npv_1=npv_2=0.96									
0.2	99.6	−1.4	66.2	0.1	96.0	35.6	−0.5	92.8	7.2

Abbreviation: fpf_1 and fpf_2, false positive rate positive rate of test 1 and test 2 respectively; rr_{sen}, rr_{fpf}, rr_{ppv}, rr_{ppv}, relative sensitivity, relative FPF, relative PPV and relative NPV between test 2 and test 1 respectively.

Table 3a

The performance of the generalized estimating equation approach using a log link in a general setting (Models (3) & (5))

Test 1						Comparison					
sen_1		fpf_1		$dlr+_1$ Bias	$dlr-_1$ Bias	rr_{sen}			rr_{fpf}		
Bias	Cov	Bias	Cov			Bias	Cov	Power	Bias	Cov	Power
sen_1=sen_2=0.50, fpf_1=0.25, fpf_2=0.20											
42.7	0.0	136	0.0	−39.5	−4.9	3.6	95.4	4.6	7.9	34.8	100
sen_1=0.50, sen_2=0.60, fpf_1=0.25, fpf_2=0.20											
34.9	0.2	136	0.0	−42.9	19.2	1.2	95.4	85.0	8.0	33.6	100
sen_1=0.50, sen_2=0.60, fpf_1=fpf_2=0.25											
34.9	0.4	121	0.0	−39.3	9.5	0.9	94.0	83.0	5.7	54.4	45.6

Test 1				Comparison					
ppv_1		npv_1		rr_{ppv}			rr_{npv}		
Bias	Cov	Bias	Cov	Bias	Cov	Power	Bias	Cov	Power
ppv_1=ppv_2=0.10, npv_1=0.96, npv_2=0.97									
2.0	96.4	−3.3	0.0	−2.8	95.6	4.4	−6.3	10.8	99.8
ppv_1=0.06, ppv_2=0.07, npv_1=npv_2=0.96									
2.5	99.2	−1.4	70.8	−1.7	94.6	29.4	−0.6	92.0	8.0

Appendix D: R and SAS program

Define sweight=1/disease verification fraction, program to use for estimating and comparing test performances are given as follow:

```
In R:
For sensitivity and FPF and log link function:
geese(Test˜Testtype+case
+Testtype:case,id=subid,weights=sweight,data=geedata,corstr=“indep”,family
=poisson)
For PPV and NPV
agree<-1-abs(case-Test)
geese(agree˜Test+Testtype
+Testtype:Test,id=subid,weights=sweight,data=geedata,corstr=“indep”,famil
y =poisson)
In SAS: For sensitivity and FPF and log link function (similar code for PPV
and NPV)
proc genmod data= descending;
class subid;
model Test=Testtype case Testtype*case/dist=poisson;
repeated subject=subid/corr=indep;
weight sweight;
run;
```

Table 1

Summary of a review of recent large cervical screening studies (over 1,000 women) which conducted colposcopy in only a subset of screening negatives

Number of Studies	Methods used to compare performance between assays
Nine studies (11–19)	No formal statistical comparison is provided
Two studies (20–21)	McNemar Test
One study (22)	Fisher's Exact Test
One study (23)	Bootstrap method
One study (24)	Two-sampled t-test
One study (25)	Zhou's likelihood method (7)

Table 2

The performance of the proposed weighted generalized estimating equation model using a logit link (models (1)–(2))

[1]Test 1						Comparison					
sen_1		spe_1		dlr+$_1$ Bias	dlr−$_1$ Bias	or_{sen}			or_{spe}		
[2]Bias	Cov	Bias	Cov			Bias	Cov	Power	Bias	Cov	Power
sen_1=sen_2=0.50, spe_1=0.75, spe_2=0.80											
0.6	94.6	−0.1	99.6	0.9	−0.8	−1.5	95.0	5.0	0.2	95.6	100
sen_1=0.50, sen_2=0.60, spe_1=0.75, spe_2=0.80											
0.8	93.0	0.0	99.8	0.5	− 0.8	1.7	95.8	74.8	0.2	95.2	100
sen_1=0.50, sen_2=0.60, spe_1=spe_2=0.75											
1.6	93.0	0.0	100	1.2	−1.1	0.1	95.2	72.0	0.2	95.6	4.4

Test 1				Comparison					
ppv_1		npv_1		or_{ppv}			or_{npv}		
Bias	Cov	Bias	Cov	Bias	Cov	Power	Bias	Cov	Power
ppv_1=ppv_2=0.10, npv_1=0.96, npv_2=0.97									
0.5	97.6	0.0	96.4	−0.5	95.8	4.2	0.7	94.8	99.2
ppv_1=0.06, ppv_2=0.07, npv_1=npv_2=0.96									
0.2	99.4	0.0	96.6	−0.7	95.4	33.2	−0.6	95.4	4.6

Note: Results based on 500 simulated data sets of size=5,000 among whom 10% of negative on both tests and all positive on any test obtained disease verification and model (1);

[1] As several performance measures are calculated for each test and the findings from the two tests are similar, for simplicity, only the results for test 1 and the comparison between test 1 and 2 are reported (test 2 data not shown).

[2] Bias(%)= (estimate-true value)/true value *100; Cov(%) =the percent that the estimated CI contained the true parameter; Power(%)=the percent that the estimated CI does not contain the null value: when the true parameter equals to the null value, this percent is the empirical type I error rate; when the true parameter does not equal to the null value, this percent is the empirical power.

Abbreviation: sen_1 and sen_2, sensitivity of test 1 and 2 respectively; spe_1 and spe_2, specificity of test 1 and 2 respectively; or_{sen}, odds of sensitivity of test 1 vs odds of sensitivity of test 2; or_{spe}, odds of specificity of test 1 vs odds of specificity of test 2; dlr+$_1$ and dlr−$_1$, positive and negative diagnostic likelihood ratio for test 1 respectively; ppv_1 and ppv_2, positive predicative value of test 1 and 2 respectively; npv_1 and npv_2, negative predicative value of test 1 and 2 respectively; or_{ppv}, odds of PPV of test 1 vs odds of PPV of test 2; or_{npv}, odds of NPV of test 1 vs odds of NPV of test 2.

Table 3

The performance of the proposed weighted generalized estimating equation approach using a log link (models (3), (5))

Test 1						**Comparison**					
sen_1		**fpf_1**		**dlr^{+}_{1} Bias**	**dlr^{-}_{1} Bias**	**rr_{sen}**			**rr_{fpf}**		
Bias	**Cov**	**Bias**	**Cov**			**Bias**	**Cov**	**Power**	**Bias**	**Cov**	**Power**
sen_1=sen_2=0.50, fpf_1=0.25, fpf_2=0.20											
0.9	94.6	0.2	99.6	0.9	−0.8	−0.8	94.0	6.0	−0.2	95.6	100
sen_1=0.50, sen_2=0.60, fpf_1=0.25, fpf_2=0.20											
0.6	93.6	0.11	99.8	0.5	−0.8	0.3	95.4	77.0	0.0	95.0	100
sen_1=0.50, sen_2=0.60, fpf_1=fpf_2=0.25											
1.2	92.4	−0.0	100	1.2	−1.1	−0.4	93.4	76.2	0.0	97.0	3.0

Test 1				**Comparison**					
ppv_1		**npv_1**		**rr_{ppv}**			**rr_{npv}**		
Bias	**Cov**	**Bias**	**Cov**	**Bias**	**Cov**	**Power**	**Bias**	**Cov**	**Power**
ppv_1=ppv_2=0.10, npv_1=0.96, npv_2=0.97									
−0.1	98.0	0.1	94.6	−0.3	97.0	3.0	0.0	97.4	99.8
ppv_1=0.06, ppv_2=0.07, npv_1=npv_2=0.96									
0.2	99.6	0.0	93.6	0.1	96.0	35.6	0.0	95.2	4.8

Abbreviation: fpf_1 and fpf_2, false positive rate positive rate of test 1 and test 2 respectively; rr_{sen}, rr_{fpf}, rr_{ppv}, rr_{ppv}, relative sensitivity, relative FPF, relative PPV and relative NPV between test 2 and test 1 respectively.

Table 4

The performance of the proposed weighted generalized estimating equation approach using a log link in a general setting (models (3), (5))

Test 1						Comparison					
sen_1		**fpf_1**		**$dlr+_1$ Bias**	**$dlr-_1$ Bias**	**rr_{sen}**			**rr_{fpf}**		
^{2}Bias	**Cov**	**Bias**	**Cov**			**Bias**	**Cov**	**Power**	**Bias**	**Cov**	**Power**
sen_1=sen_2=0.50, fpf_1=0.25, fpf_2=0.20											
0.9	95.6	0.1	99.8	0.8	−1.2	−0.1	96.0	4.0	0.0	96.4	100
sen_1=0.50, sen_2=0.60, fpf_1=0.25, fpf_2=0.20											
0.0	91.8	−0.1	99.8	0.4	−0.3	−0.1	95.0	68.2	0.1	96.0	100
sen_1=0.50, sen_2=0.60, fpf_1=fpf_2=0.25											
0.6	92.6	0.2	100	0.5	−0.8	−0.5	94.6	67.8	−0.1	96.4	3.6

Test 1				Comparison					
ppv_1		**npv_1**		**rr_{ppv}**			**rr_{npv}**		
Bias	**Cov**	**Bias**	**Cov**	**Bias**	**Cov**	**Power**	**Bias**	**Cov**	**Power**
ppv_1=ppv_2=0.10, npv_1=0.96, npv_2=0.97									
−0.6	96.4	0.1	94.8	−0.2	96.0	4.0	0.0	97.4	100.0
ppv_1=0.06, ppv_2=0.07, npv_1=npv_2=0.96									
0.2	99.2	0.0	92.6	0.2	95.4	31.2	0.0	96.2	3.8

Table 5

Application to the Canadian Cervical Cancer Screening Trial in comparing the performance between Pap and HPV in detecting CIN 2 & plus based on the conservative definition using combined study groups

	Pap$_1$	HPV$_2$	p-value	RR (hpv vs pap)
sensitivity	**56.4** (43.1, 74.0)	**97.4** (92.4, 1.00)	0.0002	1.72 (1.30,2.29)
specificity	**97.3** (94.9,97.7)	**94.3** (93.7,96.9)	0.0000	0.97 (0.96,0.98)
PPV	**8.4** (5.7,12.4)	**7.0** (5.2, 9.4)	0.1933	0.83 (0.62,1.10)
NPV	**99.8** (99.7,99.9)	**99.9** (99.8,1.00)	0.0002	1.002 (1.001,1.003)
DLR Positive	21.2 (15.61, 28.77)	17.2 (14.47, 19.41)	0.1960	0.81 (0.59, 1.11)
DLR Negative	0.45 (0.31, 0.64)	0.03 (0.004, 0.192)	0.0070	0.06 (0.01,0.47)

[1] ASCUS or worse threshold is used;

[2] ≥ 1pg HPV DNA/ml is used;

Numbers in bold were also provided in the results of Mayrand, et al. 2007 (24).