Published in final edited form as:
Stat Med. 2012 February 28; 31(5): . doi:10.1002/sim.4422.

Estimating the agreement and diagnostic accuracy of two diagnostic tests when one test is conducted on only a subsample of specimens

Hormuzd A. Katki[a,*], **Yan Li**[b], **David W. Edelstein**[c], and **Philip E. Castle**[d]

[a]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, U.S.A.

[b]Department of Mathematics, University of Texas at Arlington, Arlington, Texas, U.S.A.

[c]Carnegie Mellon University, Pittsburgh, PA, U.S.A.

[d]American Society for Clinical Pathology, Washington, DC, U.S.A.

Abstract

We focus on the efficient usage of specimen repositories for the evaluation of new diagnostic tests and for comparing new tests with existing tests. Typically, all pre-existing diagnostic tests will already have been conducted on all specimens. However, we propose retesting only a judicious subsample of the specimens by the new diagnostic test. Subsampling minimizes study costs and specimen consumption yet estimates of agreement or diagnostic accuracy potentially retain adequate statistical efficiency. We introduce methods to estimate agreement statistics and conduct symmetry tests when the second test is conducted on only a subsample and no gold standard exists. The methods treat the subsample as a stratified two-phase sample and use inverse-probability weighting (IPW). Strata can be any information available on all specimens and can be used to oversample the most informative specimens. The verification bias framework applies if the test conducted on only the subsample is a gold standard. We also present IPW-based estimators of diagnostic accuracy that take advantage of stratification. We present three examples demonstrating that adequate statistical efficiency can be achieved under subsampling while greatly reducing the number of specimens requiring re-testing. Naively using standard estimators that ignore subsampling can lead to drastically misleading estimates. Through simulation, we assess the finite-sample properties of our estimators and consider other possible sampling designs for our examples that could have further improved statistical efficiency. To help promote subsampling designs, our R package `CompareTests` computes all of our agreement and diagnostic accuracy statistics.

Keywords

Verification Bias; Symmetry Test; Kappa; Two-Phase Design; HPV; sensitivity; specificity; Gold Standard

[*]Correspondence to: Hormuzd Katki, 6120 Executive Blvd Room 8014, Rockville, Maryland, U.S.A. katkih@mail.nih.gov.

Conflict of Interest: None declared.

1. Introduction

Discovering new biomarkers, developing new diagnostic tests for those biomarkers, and validating new diagnostic tests are fundamental goals in medical research. Improving this "Biomarker Pipeline" [1, 2, 3] to speedily deliver validated and useful diagnostic tests is a major research priority. A key to improving the pipeline is the establishment and efficient usage of large specimen repositories to allow for the evaluation of new diagnostic tests and for comparing them with existing diagnostic tests [2]. Typically, all pre-existing diagnostic tests will already have been conducted on all specimens. When a gold standard exists, the standard diagnostic accuracy statistics are sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) [4].

Often, no gold standard test exists, as is often the case for the presence of a pathogen or for history of occupational exposure to a chemical. When no gold standard exists, agreement statistics (% agreement, % agreement by category, and Kappa) and odds ratios for estimating and testing symmetry amongst specimens with discordant test results are computed [5, Ch.8 & 11]). Although diagnostic accuracy statistics receive much attention, in our experience, agreement statistics play a prominent role in the development of diagnostic tests. For example, poor overall agreement can weed out obviously poor diagnostic tests [6]. Disagreements for particular test categories have been repeatedly and successfully used to find differences between tests that have led to successive improvements in test performance [7]. In the Discussion, we will discuss the possibility of using latent-class models that try to estimate diagnostic accuracy without a gold standard [8].

Standard methods apply if all specimens are retested by a second diagnostic test. However, precious specimen repositories can often be more efficiently used by selecting only a subsample of the specimens for retesting by the new diagnostic test. For example, we will present data estimating agreement and symmetry for two diagnostic tests that detect the presence of *Chlamydia Trachomatis*, a sexually-transmitted pathogen for which there is no gold standard diagnostic test [9]. The standard test was already conducted in the specimen repository, yielding 827 positive for Chlamydia and 4998 negative for Chlamydia. Retesting all 4998 negative specimens by the new test would have been prohibitively expensive and unnecessary for adequately estimating agreement statistics. Thus only an 8% subsample of the 4998 negative specimens was retested, preserving 4596 specimens and saving $230,000 in testing costs. Yet, as we show, the subsampling retained enough statistical information to achieve all study goals.

The subsampling can be stratified by any prior information available on all specimens, such as other previously-conducted diagnostic test results or demographic variables. In a second example, we will present data for comparing two human papillomavirus (HPV) DNA tests that we categorized into five ordinal categories. We sampled 25% of the repository within 48 sampling strata. Stratification ensured that rare subgroups were subsampled, that informative specimens were oversampled, and that uninformative specimens were undersampled.

Subsampling is also analogous to the situation of verification bias [10, 11] when conducting a gold standard diagnostic test. Many likelihood-based methods have already been developed to estimate diagnostic accuracy for binary or ordinal tests (e.g. [12, 13, 14, 15, 16, 17, 18, 19, 20, 21]) under verification bias. Under subsampling, the investigator chooses specimens for re-testing with a gold standard in a beneficial way to reduce costs and specimen usage while achieving enough statistical efficiency [27]. Thus we can rely on the missing at random (MAR) assumption: the reasons why the second test (which may be a

gold standard) was conducted depend only on the diagnostic test result and known specimen characteristics but not on the unknown test results.

The subsampling problem can also be formulated as a stratified two-phase sampling design (also known as double sampling [22, Ch. 12]) and use inverse-probability weighting (IPW) methods. Pickles et al [23] reviews applications of two-phase sampling in epidemiologic research. Much work has been done exploting two-phase sampling methods for determining diagnostic test accuracy when a gold standard is available, both for analysis (e.g. [24, 25, 26] and for optimal design (e.g. [28, 29, 30, 31, 32]). In particular, [30] gives expressions for optimal sampling fractions for estimating sensitivity, specificity, and positive predictive value under fixed study costs without auxiliary sampling strata. Although IPW estimators for diagnostic accuracy have been previously presented, we present simple closed-form expressions for all estimators and variances for polychotomous test results and polychotomous gold standard results under stratified two-phase sampling. We present simulations that show the value of stratification in the two-phase study design. However, we are unaware of work on agreement statistics under subsampling. Therefore, this paper focuses on introducing IPW-based estimators of agreement statistics under stratified two-phase sampling.

We show three examples that demonstrate the insights we gained, the resources we saved, the issues we encountered when selecting a subsampling scheme, and the ease of practical application of our methods. Our first example compares binary tests with the simplest possible subsampling to distill the simple essence of subsampling without needless detail. Our second example compares two ordinal tests using 48 sampling strata to demonstrate the full power of our proposal to oversample the most informative specimens to gain more information while consuming fewer specimens. Our third example pertains to estimating diagnostic accuracy under subsampling. We show that naive usage of standard estimators that ignore subsampling would have produced badly misleading estimates had they been used. Through simulation, we assess the finite-sample properties of our estimators and consider other possible sampling designs for two of our data examples that could have further improved statistical efficiency. Our R package `CompareTests` [33] computes our agreement and diagnostic accuracy estimators under stratified two-phase sampling.

2. Methods

Denote the diagnostic test conducted on all specimens as test A. Test B denotes the diagnostic test conducted on only a subsample. Each test has I ordinal categories of test results. We divide the population into S sampling strata. If each test were conducted on every specimen in the repository, the number falling into each possible combination of categories is N_{ijs} for $i, j = 1 \ldots I$ and $s = 1 \ldots S$, where i represents test A and j represents test B. The table of inferential interest marginalizes the $IxIxS$ table over strata into an IxI table. We denote the entries and margins of the marginalized table by dropping the sum over the third subscript, so for instance $N_{ij} = N_{ij+}$.

If every specimen in the repository is retested by Test B, then the agreement statistics are (neither diagnostic test is a gold standard):

$$\%agreement: P_0 = \frac{1}{N}\sum_{i=1}^{I} N_{ii}$$

$$\%agreement\ by\ category i: P_i = \frac{N_{ii}}{N_{+i} + \sum_{j \neq i}^{I} N_{ij}}$$

$$Kappa: \kappa = \frac{P_0 - P_e}{1 - P_e},\ P_e = \sum_{i=1}^{I} N_{i+} N_{+i} / N^2$$

$$Symmetry\ Odds\ Ratio (OR): OR = \frac{N_{ij}}{N_{ji}},\ i \neq j$$

$$Omnibus\ Symmetry\ Test: T = \sum_{i>j} \frac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}} \sim \chi^2_{I(I-1)/2}$$

For a $2x2$ table, the % agreement by categories are the % positive agreement and the % negative agreement, and the Symmetry Test reduces to McNemar's Test [34]. The Kappa is unweighted and thus treats all disagreeing categories the same. For a full discussion of agreement statistics and symmetry tests, see [5, Ch.8 & 11]. If Test B is a gold-standard, then diagnostic test accuracy is measured by (simplified to $2x2$ tables): $PPV = N_{22}/N_{+2}$, $NPV = N_{11}/N_{+1}$, $sensitivity = N_{22}/N_{2+}$, and $specificity = N_{11}/N_{1+}$.

To denote subsampling, each cell count is denoted with lower-case n_{ijs}. Note that under stratified two-phase sampling, we observe the total number of samples by results of Test A which are the margins N_{+js}. The margins N_{i+s} are not observable because only a subsample is retested by Test B.

For each stratum s, each column j has sampling fraction π_{js}, denoting the fraction of the N_{js} selected for retesting by Test B. We estimate each N_{ijs} as $\hat{N}_{ijs} = n_{ijs} w_{js}$, weighting the observed n_{ijs} by the inverse of the sampling probability $w_{js} = 1/\pi_{js}$. We marginalize over strata to estimate each cell N_{ij} as $\hat{N}_{ij} = \sum_{s=1}^{S} \hat{N}_{ijs}$. The Test B margin is estimated as $\hat{N}_{i+} = \sum_{j=1}^{I} \hat{N}_{ij}$. Note that the weighted estimates are consistent estimates of each total count [35, Ch.7]. Thus each agreement statistic and diagnostic accuracy statistic is consistently estimated by plugging in the weighted estimate of the required cell or margin (the symmetry test will be dealt with later).

We are interested in making inference with respect to a superpopulation where the each test specimen is fixed to its sampling stratum and its Test A results are also fixed, and only Test B results are intrinsically random. This superpopulation ensures that our inference is maximally relevant to the specimen repository. Thus the margins N_{+js} are fixed. The variance of each cell count, treating sample weights w_{js} as constant, is

$$Var(\hat{N}_{ij}) \sum_{s=1}^{S} Var(\hat{N}_{ijs}) = \sum_{s=1}^{S} Var(n_{ijs}) w_{js}^2.$$

Since $n_{ijs} \sim multinomial(n_{+js}, p_{1s}, \ldots, p_{Is})$ where $\hat{p_{is}} = n_{ijs}/n_{+js}$ estimates the probability of a specimen in stratum s and Test A category j testing into category i on Test B, then $Var(n_{ijs}) = n_{+js} p_{is}(1 - p_{is})$. Note that $Cov(\hat{N_{ij}}, \hat{N_{ij'}}) = 0$ (for $j \neq j'$ since columns are independent, and

$$Cov(\widehat{N}_{ij},\widehat{N}_{i'j})=Cov(\sum_{s=1}^{S}\widehat{N}_{ijs},\sum_{s=1}^{S}\widehat{N}_{i'js})=\sum_{s=1}^{S}Cov(\widehat{N}_{ijs},\widehat{N}_{i'js})=\sum_{s=1}^{S}-n_{+js}p_{is}p_{i's}w_{js}^{2},$$

since only terms in the same column and stratum covary. Finally

$$Cov(\widehat{N}_{i+},\widehat{N}_{i'+})=Cov(\sum_{j=1}^{I}\widehat{N}_{ij},\sum_{j'=1}^{I}\widehat{N}_{i'j'})=\sum_{j=1}^{I}Cov(\widehat{N}_{ij},\widehat{N}_{i'j}).$$

In particular, $Var(\widehat{N}_{i+})=\sum_{j=1}^{I}Var(\widehat{N}_{ij})$.

2.1. % Agreement and % Agreement by Category

The variance of % agreement is

$$Var(\widehat{P}_{0})=\frac{1}{N^{2}}\sum_{i=1}^{I}Var(\widehat{N}_{ii}).$$

The delta-method [36, pg.26] obtains the variance of the percent agreement by category

$$Var(\widehat{P}_{i})=\frac{1}{(\widehat{N}_{i+}+N_{+i}-\widehat{N}_{ii})^{2}}\left(Var(\widehat{N}_{ii})+P_{i}^{2}\sum_{j\neq i}^{I}Var(\widehat{N}_{ij})\right),$$

since N_{+i} is fixed and $Cov(\widehat{N}_{ii},\sum_{j\neq i}^{I}\widehat{N}_{ij})=0$ since the columns are independent.

2.2. Kappa

The previous section computes $\hat{P_0}$ and its variance. We must compute P_e, its variance, and its covariance with P_0. We have

$$\begin{aligned}Var(\widehat{P}_{e})&=\frac{1}{N^{4}}Var\left(\sum_{i=1}^{I}N_{+i}\widehat{N}_{i+}\right)\\&=\frac{1}{N^{4}}\sum_{i,j=1}^{I}Cov(N_{+i}\widehat{N}_{i+},N_{+j}\widehat{N}_{j+})=\frac{1}{N^{4}}\left(\sum_{i=1}^{I}N_{+i}^{2}Var(\widehat{N}_{i+})\right.\\&\left.+\sum_{i\neq j}N_{+i}N_{+j}Cov(\widehat{N}_{i+},\widehat{N}_{j+})\right),\end{aligned}$$

Finally

$$
\begin{aligned}
&Cov(\widehat{P}_0, \widehat{P}_e) \\
&= Cov\left(\frac{1}{N}\sum_{i=1}^{I}\widehat{N}_{ii}, \frac{1}{N^2}\sum_{i=1}^{I}N_{+i}\widehat{N}_{i+}\right) \\
&= \frac{1}{N^3}\sum_{i=1}^{I}\sum_{j=1}^{I}N_{+j}\,Cov(\widehat{N}_{ii}, \widehat{N}_{j+}) = \frac{1}{N^3}\left(\sum_{i=1}^{I}N_{+i}\,Var(\widehat{N}_{ii})\right. \\
&\left. + \sum_{i=1, j\neq i}^{I}N_{+j}\,Cov(\widehat{N}_{ii}, \widehat{N}_{ji})\right)
\end{aligned}
$$

since $Cov(\hat{N}_{ii}, \hat{N}_{j+}) = Cov(\hat{N}_{ii}, \hat{N}_{ji})$.

The delta-method obtains the variance of κ

$$Var(\kappa) = \frac{1}{(1-P_e)^2} \times \left(Var(P_0) + 2Cov(P_0, P_e)(\kappa - 1) + Var(P_e)(\kappa - 1)^2\right)$$

2.3. Symmetry Log Odds-Ratio and Test of Symmetry

The symmetry log odds-ratio $log(\hat{N}_{ij})/\hat{N}_{ji}$, using the delta-method, has variance

$$Var(\widehat{N}_{ij})/\widehat{N}_{ij}^2 + Var(\widehat{N}_{ji})/\widehat{N}_{ji}^2$$

A test of overall Symmetry in the *IxI* table that does not condition on $N_{ij} + N_{ji}$ being fixed is based on the estimated $\hat{N}_{ij} - \hat{N}_{ji}$ and its variance $Var(\hat{N}_{ij}) + Var(\hat{N}_{ji})$. Then the test

$$\sum_{i>j}\frac{(\widehat{N}_{ij} - \widehat{N}_{ji})^2}{Var(\widehat{N}_{ij}) + Var(\widehat{N}_{ji})}$$

should be distributed in large samples as $\chi^2_{I(I-1)/2}$ [37]. Since this test does not condition on the sum of the discordant pairs, it will not reduce to McNemar's test if all specimens are retested.

2.4. Sensitivity, Specificity, Positive and Negative Predictive Values

The variances of PPV and NPV are $Var(PPV) = Var(\widehat{N}_{22})/N_{+2}^2$ and $Var(NPV) = Var(\widehat{N}_{11})/N_{+1}^2$. For sensitivity and specificity, the delta-method obtains the variances

$$Var(spec) = \frac{1}{\widehat{N}_{1+}^2} \times \left(Var(\widehat{N}_{11}) + (spec)^2(Var(\widehat{N}_{1+})) - 2(spec)\,Var(\widehat{N}_{11})\right) = \frac{1}{\widehat{N}_{1+}^2} \times \left((1-spec)^2\,Var(\widehat{N}_{11}) + (spec)^2\,Var(\widehat{N}_{12})\right)$$

$$Var(sens) = \frac{1}{\widehat{N}_{2+}^2} \times \left((1-sens)^2\,Var(\widehat{N}_{22}) + (sens)^2\,Var(\widehat{N}_{21})\right)$$

To extend the above to *IxI* tables, the variance of the i^{th} predictive value N_{ii}/N_{+i} is $Var(\widehat{N}_{ii})/N_{+i}^2$ and the variance of the i^{th} category-specific classification probability (extension of sensitivity and specificity) N_{ii}/N_{i+} is

$$\frac{1}{\widehat{N}_{i+}^2} \times \left(\left(1 - \frac{N_{ii}}{N_{i+}}\right)^2 Var(\widehat{N}_{ii}) + \left(\frac{N_{ii}}{N_{i+}}\right)^2 (Var(\widehat{N}_{i+}) - Var(\widehat{N}_{ii})) \right).$$

We note that if there are no sampling strata, and then estimates of PPV and NPV that ignore the sampling are identical to estimates that account for sampling. This is because when sampling is ignored, the denominators of PPV and NPV are not N_{i+}, but instead the margin for those with both tests observed n_{i+}. Since the sampling weight is N_{i+}/n_{i+}, the sampling weight in the numerator and denominator cancels out of the weighted estimate. Note that this argument does not hold for sensitivity or specificity.

3. Examples and simulation

3.1. Example: Comparing the HC2 and Ct-DT assays for Chlamydia

We assessed the agreement of a Hybrid Capture 2 test (HC2; Qiagen Corporation, Gaithersburg, MD, USA) and the *C. trachomatis* Detection and genoTyping assay (Ct-DT; DDL Diagnostic Laboratory, Voorburg, The Netherlands) for detecting a Chlamydia infection [9]. HC2 had been previously conducted at enrollment, as part of a complete pelvic examination, on all women in the Cervarix Vaccine Trial (CVT) in Costa Rica, yielding 827 women testing positive by HC2 for Chlamydia and 4998 women testing negative by HC2 for Chlamydia. All 827 HC2-positive were retested by Ct-DT, yielding 27 negative and 800 positive. To minimize study costs and specimen consumption, 402 (8%) of the HC2-negative specimens were randomly chosen for retesting by Ct-DT. Of those 402, 6 tested Ct-DT positive and 396 tested Ct-DT negative. Weighting the 402 to the 4998 total HC2-negative, 4923 were estimated Ct-DT negative and 75 were estimated Ct-DT positive.

The agreement statistics for HC2 and Ct-DT are vastly different between ignoring and accounting for the sampling design (Table 1). The symmetry OR changes from strongly positive to strongly negative and goes from strongly significant to borderline significant. The p-value is reduced to borderline significance because it accounts for the fact that the 75 is an estimate from weighting up just 6 observations. In addition, the category with the stronger % agreement changes and the confidence interval for Kappa lengthens after accounting for the sampling design.

3.2. Example: Agreement of two HPV tests using complex subsampling

We compared two tests for HPV DNA: SPF_{10}-$LiPA_{25}$ test with HPV16 and HPV18 genotype-specific detection (SPF_{10}; DDL Diagnostic Laboratory, Voorburg, The Netherlands) and the Linear Array (LA; Roche Molecular Systems) test. The SPF_{10} test detects 25 different HPV types and the LA test detects 37 different HPV types. Both tests detect the same 15 HPV carcinogenic types that are known to cause cervical cancer (HPV16, −18, −31, −33, −35, −39, −45, −51, −52, −56, −58, −59, −66, −68, and −73); other HPV types cause cervical lesions but are not considered to be carcinogenic. For comparison, we categorized both tests into the same five ordinal categories graded from least to most carcinogenic: HPV-negative, HPV+ but non-carcinogenic (or unknown) HPV type, HPV+ and carcinogenic type but not HPV16 or HPV18, HPV18+, and HPV16+.

As part of the baseline pelvic examination in the CVT, testing for HPV DNA was already conducted using the SPF_{10} test on all 5659 available specimens. The cost of retesting all

5659 specimens with LA would have exceeded $250,000. Instead, we conducted LA on a complex subsample of 1427 specimens [7].We used 48 sampling strata based on a Hybrid Capture 2 test (HC2; Qiagen Corporation, Gaithersburg, MD, USA) for HPV DNA (2 levels), Pap smear grade (6 levels), and SPF_{10} categorized into 4 levels 2. This sampling plan has several advantages over a 25% simple random sample. First, the sampling plan ensured that all strata with few specimens would be sampled. Strata with few specimens are rare but interesting situations and sampling them is necessary to ensure weighting up to the full cohort. Second, interesting strata could be oversampled. For example, 13 specimens were negative by HC2, positive by SPF_{10} for a carcinogenic HPV type, and the Pap smear was ASC-US, an equivocal result. We wanted to see how LA adjudicated this difficult situation, so we sampled 12 out of the 13 (one specimen was not available for testing). Third, nearly half of the specimens (2650) were negative by all three tests, and thus extensive testing of this category by LA is unlikely to yield insight; instead, we took a 4% sample of this category.

Using the weighted table (Table 3), the % agreement is 83 (80–86). The % agreements by category are: PCR-: 80 (75–85), NC/HPV+: 45 (39–51), C/HPV+: 71 (66–77), HPV18+: 64 (53–74), and HPV16+: 82 (76–88). The non-carcinogenic HPV+ category has low agreement because it tends to be confused with no HPV being found (PCR-). HPV18 also has relatively low agreement because SPF_{10} has a harder time distinguishing HPV18 from other carcinogenic types (note the 37 HPV18+ by LA but only other carcinogenic HPV+ by SPF_{10}). The kappa was 0.75 (0.71–0.79). The omnibus symmetry test has p=0.003. The major reason for lack of symmetry is that LA tends to be able find more HPV16 and HPV18 versus only finding other carcinogenic types (the bottom right 3*x*3 subtable). Thus we conclude that when SPF_{10} and LA disagree, LA tends to detect a more carcinogenic HPV type than SPF_{10}.

3.3. Example: Diagnostic Accuracy of HPV tests to detect cervical precancer

The goal of the Community Access to Cervical Health (CATCH) study was to assess the diagnostic accuracy of three different cervical cancer screening modalities (Pap smears, HPV DNA testing, and Visual Inspection with Acetic Acid (VIA)) for detecting cervical intraepithelial neoplasia grade 2 or worse (CIN2+ or ”cervical precancer”) [38]. In this study, 2331 women living in Ranga Reddy District, Andra Pradesh State, India underwent all three cervical screening tests. Women who tested positive on any of the three tests, plus a random sample of 1/5 of all women, were asked to undergo the gold-standard of colposcopically-directed biopsy to definitively determine the presence of CIN2+. In addition, of the 1052 women asked to undergo colposcopically-directed biopsy, 670 have the presence of CIN2+ determined. Although women were allowed to refuse biopsy, the MAR assumption within strata of the screening tests is reasonable since there are no outward signs of cervical precancer that these woman could use to divine their gold standard disease status, and cervical cancer (which could cause noticeable symptoms) was rare (only 4 cases were found).

We estimate diagnostic accuracy of the HPV test for detecting CIN2+, accounting for the verification bias within the eight strata of possible test results (Table 4). Each stratum in Table 4 forms a 2*x*2 table of HPV test result by presence of CIN2+, with each 2*x*2 table having a zero row, depending on whether that stratum is for HPV+ or HPV-. Marginalizing over strata, the observed 2*x*2 table of interest of CIN2+ by HPV (denoting cells as (CIN2+,HPV)) is: (−,−):529, (+,−): 3, (−,+):122, and (+,+)=16. The weighted table is: (−,−): 2081.1, (+,−): 9.9, (−,+): 211.4, and (+,+)=28.6. Note that each column of the weighted table sums to the total number of HPV test results in the full study (2091 HPV-negative and 240 HPV-positive).

Ignoring sampling, the sensitivity is 84% (60%–97%) but drops to 74% (43%–92%) by accounting for sampling. Ignoring sampling, the specificity is 81% (78%–84%) but jumps to 90.8% (90.3%–91.3%) by accounting for sampling. Both point estimates and confidence intervals change drastically. In contrast, ignoring sampling, the PPV is 12% (6.7%–18%) and by accounting for sampling is 12% (7.6%–18%). Ignoring sampling, the NPV is 99.4% (98.4%–99.9%) and by accounting for sampling is 99.5% (98.3%–99.9%). As mentioned in section 2.4, accounting for sampling does not alter the PPV or NPV estimate when there are no strata, and here the difference in sampling fractions between strata is not enough to cause important changes.

3.4. Performance of Estimators

We assess the bias, variance, and coverage properties of the agreement and symmetry statistics in Table 5 by simulating from the Chlaymdia example, and for diagnostic accuracy by simulating from the HPV example in Table 4. We simulate by first choosing a superpopulation mean table, then doing binomial sampling on each column (within strata) to generate 4000 superpopulation tables within each stratum. Within each superpopulation table, we conduct binomial sampling on the columns with specified sampling probabilities to get the observed tables under sampling. Finally, we marginalize over strata for the final table to estimate all agreement and diagnostic accuracy statistics.

For the agreement statistics, we use as the superpopulation mean table the weighted Chlamydia table, except that we change the 27 to a 75 to ensure the null hypothesis holds for the superpopulation mean. In the first null scenario, we sample as in the observed data: all the HC2 positives and 402/4998 of the HC2 negatives. Although the estimators and their variance estimators are unbiased, the coverage of the 95% CIs is generally too small. For a stated $\alpha = 0.05$, the unconditional symmetry test had an observed $\alpha = 0.080$. Thus asymptotics do not quite apply, likely because the n_{21} cell averaged only six counts. To improve coverage, in the second scenario, we doubled the sampling fraction of HC2-negatives. Now the n_{21} cell averaged twelve counts and all estimators achieved adequate coverages and for a stated $\alpha = 0.05$, the unconditional symmetry test had an observed $\alpha = 0.065$.

For diagnostic accuracy, we use as the superpopulation mean table the weighted Table 4 (rounding off to the nearest one) and test four sampling schemes. In all sampling schemes, the estimates were unbiased. For the first set of rows, the sampling fractions are set to the observed fractions for each stratum. The variance estimators for sensitivity and NPV are unbiased or a little large, but for specificity and PPV are biased too small, leading to undercoverage of their 95% confidence intervals. To rectify the coverage, we tried three other sampling schemes. First, we doubled the sampling fraction for those negative on all three screening tests, but this had no impact on any estimator except NPV. Second, we sampled all 16 individuals positive on all three screening tests, an increase of only 7 individuals, but in the group at highest risk for developing CIN2+. This improved the underestimation of the specificity and PPV variance estimators and improved their coverages. Third, we sampled all HPV+ individuals since HPV is a necessary cause of cervical cancer. This dramatically improved the underestimation of the specificity and PPV variance estimators and improved their coverages.

In the agreement simulations, the variances for all quantities are small and statistical significance for the symmetry test was still achieved under sampling, suggesting that no qualitative difference in conclusions would have been achieved had a greater sampling fraction been used (although the symmetry OR has high variance because it is the ratio of potentially small cell counts). Thus the (875+402)/5873=21.7% subsample achieved the scientific conclusions that would have been drawn had the entire repository been retested.

In the diagnostic accuracy simulations, the sampling design provided efficiency gains for estimating specificity and PPV (Table 6). Under the observed sampling fractions, 670/2331=28.7% of the women had the gold standard ascertained, yet the specificity and PPV have variances inflated by a factor of only roughly 2 versus having all women undergo the gold standard. This occurs because the sampling design oversamples the few women who developed CIN2+. For the final design that also samples all HPV+ women (670+102)/2331=33.1%), the variances of the specificity and PPV are within 10% of the variance under all women undergoing the gold standard. Again, this is because this design would capture, on average, 34.2 out of the estimated 38.5 CIN2+ in the full study. However, enriching the sample for CIN2+ has little effect on the variances of sensitivity and NPV. Instead, the sampling design that doubles the sampling fraction for those negative on all three screening tests provides efficiency gains for estimating sensitivity and NPV (but not for specificity and PPV). These findings exemplify the general result derived in [30] that cost-efficient estimation of specificity (or PPV) using two-phase designs requires oversampling the test-positives (which tend to capture most of the true gold-standard positives) but for sensitivity (or NPV) requires oversampling the test-negatives.

4. Discussion

To reduce study costs and specimen consumption in specimen repositories where the standard test has already been conducted on all specimens, we proposed conducting the new test on a judicious subsample of specimens. To estimate agreement statistics, we introduced IPW estimators that account for subsampling as stratified two-phase sampling from within a specimen repository. Stratification allows the researcher to use all relevant information available on all specimens to oversample the most informative specimens. Although IPW estimators of diagnostic accuracy statistics are not novel, we provided simple closed-form estimates for variances under general stratified two-phase sampling and present an example with simulations that demonstrate the wealth of sampling options available from this design. To accelerate the use of subsampling designs into practice, our R package `CompareTests` [33] computes our agreement and diagnostic accuracy estimators under stratified two-phase sampling.

In both the Chlamydia and HPV testing examples, sampling 25% of the specimens for retesting saved over $200,000 in testing costs, yet resulted in small variances and we still achieved statistical significance for the symmetry tests. The HPV testing example used multiple prior test results to stratify a complex subsample, ensuring that all strata would be sampled, informative strata oversampled, and uninformative strata undersampled. In the cervical cancer screening example, our chosen sampling plan oversampled women likely to have cervical precancer, allowing for improved efficiency versus simple random sampling for estimating specificity and PPV. Our methods allow the researcher to wisely choose only the most informative specimens for retesting, thus reducing costs and specimen consumption without an important loss of statistical efficiency. Furthermore, in both examples, ignoring the sampling in the analysis yielded badly biased estimators and poor variance estimates. When subsampling, one must use methods that account for subsampling.

At 8% sampling, the Chlamydia simulation showed undercoverage of true parameters due to lack of adequacy of normal approximation. However, at 16% sampling, the estimators had adequate coverage. Similarly, the HPV simulation showed undercoverage for estimating specificity and PPV for our chosen sampling scheme. Instead, sampling schemes that oversample even more CIN2+, a rare event, reduced the variances and had improved coverage. Waller et al. [39] show that confidence interval undercoverage for weighted proportions can occur when some observed stratum proportions are zero. Resampling-based

variance calculations may be useful because agreement statistics can be complicated functions of weighted proportions.

Survey sampling provides a general framework for the analysis of complex subsampling or verification bias. For example, a simplifying assumption made in the HPV example was that 111 women who appeared for colposcopy, were asked to undergo biopsy, but refused biopsy or had inadequate biopsies, were at the same CIN2+ risk as any other woman in their sampling stratum who did not appear for colposcopy. This is not true because being asked to undergo a biopsy indicates a greater risk for CIN2+ than for a woman not asked to undergo biopsy. To account for this, a three-phase sampling design would be required: the 670 women with CIN2+ status would be weighted up to represent the 781 women who appeared for colposcopy, then the 781 women would be weighted up to represent the full cohort of 2331 women. Similarly, if it was deemed desirable to account for the fact that women came from 42 villages in Ranga Reddy District, then standard cluster sampling methods could be readily employed. We note that the R package survey [40] handles general sampling designs and has a function `svykappa()` to compute Kappa.

In the introduction, we noted the availability of likelihood-based methods for estimating diagnostic accuracy under verification bias. IPW-based methods provide an alternative. Likelihood-based methods are well-known for their efficiency and weighted methods can be inefficient if the weights vary substantially [36]. However, as mentioned above, standard IPW methods from survey sampling can be readily employed under very complex subsampling. In such situations, likelihood-based methods may require the simultaneous estimation of many parameters. Comparing likelihood-based and IPW-based estimators is an important future area of research.

When no gold standard exists, an alternative to using agreement statistics is to estimate diagnostic accuracy by positing latent-class models that make assumptions about the dependence between the tests, given the unknown gold standard [8]. Although latent-class models are an important methodologic research area, crucial methodologic hurdles remain to be overcome before their routine use can be recommended. Latent-class models require a large number of tests to ensure identifiability [41], and having more degrees of freedom than parameters does not ensure identifiability [42]. Equally seriously, latent-class models are sensitive to the assumed dependence structure and it is difficult to distinguish between different dependence structures using model selection criteria [43]. Most seriously, it is unclear whether the model-based consensus of tests that latent-class models provide is useful to scientists, who may be understandably skeptical about estimating diagnostic accuracy without gold standards. Although partial gold standard evaluation can alleviate those problems [21], under subsampling of tests and no gold standard, those problems may worsen, but the research remains to be done. Agreement statistics do not solve those problems. However, agreement statistics have the crucial advantage of presenting solely what the data directly supports, i.e., whether two tests agree and patterns in how they disagree.

As noted in the Introduction, much work has been on optimal design for estimating diagnostic accuracy statistics, most notably by McNamee [30]. Our findings exemplify her general result that cost-efficient estimation of specificity (or PPV) requires oversampling test-positives for gold-standard evaluation, but for sensitivity (or NPV) requires oversampling test-negatives. She describes her general result as ”counter-intuitive”. We would like to offer an intuitive explanation. Oversampling negative test results is clearly optimal for estimating NPV. Then because NPV is affected much more by sensitivity than specificity, we might expect that oversampling test-negatives ought to be optimal for estimating sensitivity as well. The same intuition would apply to PPV and specificity.

McNamee also notes that, since most specimens will test negative, oversampling negative test results provides little cost savings, and thus the two-phase design may provide little cost-efficiency for estimating sensitivity. However, McNamee's general result did not consider extra sampling strata beyond stratifying on first-phase test results. Our simulations showed that using extra sampling strata with the two-phase design helped to estimate sensitivity. In particular, the design that oversamples the triple-negatives reduced by about 30–40% the variance for sensitivity and NPV. Optimal use of auxiliary information for stratification in two-phase designs is an important future area of research.

Proposing efficient sampling designs for comparing the agreement of diagnostic tests is another important future area of research. Choice of sample size and allocation of sample size to strata require consideration, especially for strata with rare events requiring large numbers of specimens to be retested to find a single positive specimen. Another important research avenue is efficiently using available auxiliary information on each specimen for stratification. The HPV testing example showed the value of having multiple prior test results to oversample the most informative specimens: those that are likely to have disagreeing test results. Such efficient sampling designs could achieve most of the statistical efficiency of retesting all specimens with major reductions in study costs and consumption of precious specimens. We believe that efficient sampling designs should be more widely used to improve the efficiency of the "Biomarker Pipeline".

Acknowledgments

The authors thank Barry Graubard for his support and comments on earlier versions of this manuscript. We thank two anonymous reviewers for their helpful comments and suggestions. We thank Patti Gravitt for sharing the data from the CATCH study with us. This research was supported by the Intramural Research Program of the NIH/ National Cancer Institute.

References

1. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst. 2001 Jul; 93(14):1054–1061. [PubMed: 11459866]
2. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst. 2008 Oct; 100(20):1432–1438. URL http://dx.doi.org/10.1093/jnci/djn326. [PubMed: 18840817]
3. Baker SG. Improving the biomarker pipeline to develop and evaluate cancer screening tests. J Natl Cancer Inst. 2009 Aug; 101(16):1116–1119. URL http://dx.doi.org/10.1093/jnci/djp186. [PubMed: 19574417]
4. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press; 2003.
5. Bishop, YMM.; Fienberg, SE.; Holland, PW. Discrete multivariate analysis: theory and practice. Cambridge, Mass.-London: The MIT Press; 1975. With the collaboration of Richard J. Light and Frederick Mosteller
6. Kinney W, Stoler MH, Castle PE. Special commentary: patient safety and the next generation of HPV DNA tests. Am J Clin Pathol. 2010 Aug; 134(2):193–199. URL http://dx.doi.org/10.1309/AJCPRI8XPQUEAA3K. [PubMed: 20660320]
7. Castle PE, Porras C, Quint WG, Rodriguez AC, Schiffman M, Gravitt PE, Gonzlez P, Katki HA, Silva S, Freer E, et al. Comparison of two PCR-based human papillomavirus genotyping methods. J Clin Microbiol. 2008 Oct; 46(10):3437–3445. URL http://jcm.asm.org/cgi/content/full/46/10/3437?view=long&pmid=18716224. [PubMed: 18716224]
8. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Stat Methods Med Res. 1998 Dec; 7(4):354–370. [PubMed: 9871952]

9. Quint K, Porras C, Safaeian M, Gonzalez P, Hildesheim A, Quint W, van Doorn LJ, Silva S, Melchers W, Schiffman M, et al. Evaluation of a novel PCR-based assay for detection and identification of chlamydia trachomatis serovars in cervical specimens. J Clin Microbiol. 2007 Dec; 45(12):3986–3991. [PubMed: 17959760]
10. Begg CB. Biases in the assessment of diagnostic tests. Stat Med. 1987 Jun; 6(4):411–423. [PubMed: 3114858]
11. Zhou XH. Effect of verification bias on positive and negative predictive values. Stat Med. 1994 Sep; 13(17):1737–1745. [PubMed: 7997707]
12. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983 Mar; 39(1):207–215. [PubMed: 6871349]
13. Tosteson TD, Titus-Ernstoff L, Baron JA, Karagas MR. A two-stage validation study for determining sensitivity and specificity. Environ Health Perspect. 1994 Nov; 102(Suppl 8):11–14. [PubMed: 7851324]
14. Baker SG. Evaluating multiple diagnostic tests with partial verification. Biometrics. 1995 Mar; 51(1):330–337. [PubMed: 7539300]
15. Zhou XH. A nonparametric maximum likelihood estimator for the receiver operating characteristic curve area in the presence of verification bias. Biometrics. 1996 Mar; 52(1):299–305. [PubMed: 8934599]
16. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. Stat Med. 1996 Aug; 15(16):1807–1826. URL http://dx.doi.org/3.0.CO;2-U. [PubMed: 8870162]
17. Rodenberg C, Zhou XH. ROC curve estimation when covariates affect the verification process. Biometrics. 2000 Dec; 56(4):1256–1262. URL http://www3.interscience.wiley.com/journal/119032449/abstract?CRETRY=1&SRETRY=0. [PubMed: 11129488]
18. Schneider DL, Burke L, Wright TC, Spitzer M, Chatterjee N, Wacholder S, Herrero R, Bratti MC, Greenberg MD, Hildesheim A, et al. Can cervicography be improved? an evaluation with arbitrated cervicography interpretations. Am J Obstet Gynecol. 2002 Jul; 187(1):15–23. [PubMed: 12114883]
19. 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]
20. Alonzo TA, Kittelson JM. A novel design for estimating relative accuracy of screening tests when complete disease verification is not feasible. Biometrics. 2006 Jun; 62(2):605–612. URL http://dx.doi.org/10.1111/j.1541-0420.2005.00445.x. [PubMed: 16918926]
21. Albert PS, Dodd LE. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. J Am Stat Assoc. 2008 Mar; 103(481):61–73. URL http://dx.doi.org/10.1198/016214507000000329. [PubMed: 19802353]
22. Cochran, WG. Sampling techniques. 1st ed. John Wiley & Sons; 1953.
23. Pickles A, Dunn G, Vzquez-Barquero JL. Screening for stratification in two-phase ('two-stage') epidemiological surveys. Stat Methods Med Res. 1995 Mar; 4(1):73–89. [PubMed: 7613639]
24. Tenenbein A. A double sampling scheme for estimating from binomial data with misclassifications. Journal of the American Statistical Association. 1970; 65(331):1350–1361. URL http://www.jstor.org/stable/2284301.
25. Hochberg Y. On the use of double sampling schemes in analyzing categorical data with misclassification errors. Journal of the American Statistical Association. 1977; 72:914–921.
26. Haitovsky Y, Rapp J. Conditional resampling for misclassified multinomial data with applications to sampling inspection (Corr: 94V36 p334). Technometrics. 1992; 34:473–483.
27. Kraemer, HC. Evaluating Medical Tests: Objective and Quantitative Guidelines. Sage Publications Inc; 1992.
28. Tenenbein A. A double sampling scheme for estimating from binomial data with misclassifications: Sample size determination. Biometrics. 1971; 27:935–944.
29. Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient study designs to assess the accuracy of screening tests. Am J Epidemiol. 1994 Oct; 140(8):759–769. [PubMed: 7942777]

30. McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. Stat Med. 2002 Dec; 21(23):3609–3625. URL http://dx.doi.org/10.1002/sim.1318. [PubMed: 12436459]
31. Wruck LM, Yiannoutsos CT, Hughes MD. A sequential design to estimate sensitivity and specificity of a diagnostic or screening test. Stat Med. 2006 Oct; 25(20):3458–3473. URL http://dx.doi.org/10.1002/sim.2451. [PubMed: 16374904]
32. Kosinski AS, Chen Y, Lyles RH. Sample size calculations for evaluating a diagnostic test when the gold standard is missing at random. Stat Med. 2010 Jul; 29(15):1572–1579. URL http://dx.doi.org/10.1002/sim.3899. [PubMed: 20552570]
33. Katki, HA.; Edelstein, DW. CompareTests: Estimating agreement and diagnostic accuracy when one test is not conducted on all specimens. 2011. URL http://dceg.cancer.gov/bb/tools/CompareTests, r package version 1.0
34. McNemar Q. Note on the sampling error of the differences between correlated proportions of percentages. Psychometrika. 1947; 12:153–157. [PubMed: 20254758]
35. Raj, D. Sampling theory. New York: McGraw-Hill Book Co; 1968.
36. Korn, EL.; Graubard, BI. Analysis of Health Surveys. John Wiley & Sons; 1999.
37. Koch GG, Freeman DH Jr, Freeman JL Jr. Strategies in the multivariate analysis of data from complex surveys. International Statistical Review. 1975; 43:59–78.
38. Gravitt PE, Paul P, Katki HA, Vendantham H, Ramakrishna G, Sudula M, Kalpana B, Ronnett BM, Vijayaraghavan K, Shah KV, et al. 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. URL http://dx.doi.org/10.1371/journal.pone.0013711. [PubMed: 21060889]
39. Waller JL, Addy CL, Jackson KL, Garrison CZ. Confidence intervals for weighted proportions. Stat Med. 1994 May; 13(10):1071–1082. [PubMed: 8073202]
40. Lumley T. survey: analysis of complex survey samples. 2010 URL http://cran.r-project.org/web/packages/survey/index.html, r package version 3.23-3.
41. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med. 1999 Nov; 18(22):2987–3003. [PubMed: 10544302]
42. Jones G, Johnson WO, Hanson TE, Christensen R. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics. 2009 Sep. URL http://dx.doi.org/10.1111/j.1541-0420.2009.01330.x.
43. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics. 2004 Jun; 60(2):427–435. URL http://dx.doi.org/10.1111/j.0006-341X.2004.00187.x. [PubMed: 15180668]

Table 1

Agreement statistics for HC2 and Ct-DT for detecting Chlamydia, computed separately for estimators that ignore and estimators that account for the sampling design

	% agreement	% negative agreement	% positive agreement	Kappa	Symmetry OR
Ignoring sampling	97 (96–98)	92 (90–95)	96 (95–97)	0.94 (0.92–0.96)	4.5, p=0.0003
Accounting for sampling	98 (97–99)	98 (97–99)	89 (83–95)	0.93 (0.89–0.97)	0.36, p=0.02

Table 2

Sampling plan to select specimens for testing by the LA test to compare with the SPF_{10} test [7, Table 1]. ”No. tested” is the actual number of specimens tested. −, negative result; +, positive result; C, carcinogenic; NC, noncarcinogenic; SPF+ LIPA−, SPF_{10} found HPV DNA but could not determine the type; HSIL, high-grade squamous intraepithelial lesion; AGUS, atypical glandular cells of undetermined significance; ASC-H, atypical squamous cells, HSIL cannot be ruled out; LSIL, low-grade squamous epithelial lesion.; ASC-US, Atypical Squamous Cells of Undetermined Significance

	HC2 Negative				HC2 Positive				
Pap Smear	**SPF−**	**SPF+ LIPA−**	**SPF+ NC**	**SPF+ C**	**SPF−**	**SPF+ LIPA−**	**SPF+ NC**	**SPF+ C**	**Total**
HSIL									
N (Cohort)	1	0	1	2	0	3	1	68	76
n (Tested)	1	0	1	2	0	2	1	67	74
Sampling Fraction	100%		100%	100%		67%	100%	99%	97%
LSIL									
N (Cohort)	10	16	25	16	3	21	25	406	522
n (Tested)	9	13	24	14	2	20	25	184	291
Sampling Fraction	90%	81%	96%	88%	67%	95%	100%	45%	56%
AGUS									
N (Cohort)	4	1	0	1	0	0	0	7	13
n (Tested)	3	1	0	1	0	0	0	5	10
Sampling Fraction	75%	100%		100%				71%	77%
ASC-H									
N (Cohort)	0	0	0	0	0	0	0	16	16
n (Tested)	0	0	0	0	0	0	0	15	15
Sampling Fraction								94%	94%
ASC-US									
N (Cohort)	43	14	13	13	0	13	12	174	282
n (Tested)	39	14	12	12	0	13	12	78	180
Sampling Fraction	91%	100%	92%	92%		100%	100%	45%	64%
Negative									
N (Cohort)	2650	325	252	316	101	63	66	977	4750
n (Tested)	115	221	67	87	91	50	51	175	857
Sampling Fraction	4%	68%	27%	28%	90%	79%	77%	18%	18%

Stat Med. Author manuscript; available in PMC 2013 October 28.

	HC2 Negative				HC2 Positive				
Pap Smear	**SPF−**	**SPF+ LIPA−**	**SPF+ NC**	**SPF+ C**	**SPF−**	**SPF+ LIPA−**	**SPF+ NC**	**SPF+ C**	**Total**
Total									
N (Cohort)	2708	356	291	348	104	100	104	1648	5659
n (Tested)	167	249	104	116	93	85	89	524	1427
Sampling Fraction	6%	70%	36%	33%	89%	85%	86%	32%	25%

Stat Med. Author manuscript; available in PMC 2013 October 28.

Table 3

Comparison of SPF_{10} and LA test results, categorized hierarchically according to HPV cancer risk. Highlighted in bold are the results for concordant cells i.e., cells with the same HPV cancer risk group detected by both methods. NC, noncarcinogenic; C, carcinogenic HPV genotypes other than HPV16 and HPV18. −, negative result; +, positive result.

	Number of specimens with indicated LA result					
SPF_{10} result	**PCR−**	**NC/HPV+**	**C/HPV+**	**HPV18+**	**HPV16+**	**Total**
PCR−	2,498	199	109	1	6	2,812
NC/HPV+	196	520	113	9	13	851
C/HPV+	77	103	1,200	37	33	1,451
HPV18+	8	5	2	120	6	141
HPV16+	13	7	5	0	379	404
Total	2,791	835	1,429	167	437	5,659

Table 4

CATCH study. Observed gold standard CIN2+ results within eight strata defined by possible screening test results.

HPV	VIA	PAP	observed CIN2+	observed not CIN2+	stratum size	sampling weight	weighted CIN2+	weighted not CIN2+
−	−	−	1	281	1598	5.7	5.7	1592.3
+	−	−	5	87	153	1.7	8.3	144.7
−	+	−	0	119	213	1.8	0.0	213.0
+	+	−	1	10	23	2.1	2.1	20.9
−	−	+	2	108	235	2.1	4.3	230.7
+	−	+	6	20	48	1.8	11.1	36.9
−	+	+	0	21	45	2.1	0.0	45.0
+	+	+	4	5	16	1.8	7.1	8.9
			19	651	2331		38.5	2292.5

Table 5

Simulation of performance of estimators of agreement under the null hypothesis using (1) sampling fractions used in the study, and (2) doubling the sampling fraction amongst the HC2-.

		% agreement	% negative agreement	% positive agreement	Kappa	Symmetry log(OR)
	true	97.5	97.0	84.2	0.899	0
	full cohort variance	4.1e-06	5.4e-06	1.3e-04	6.4e-05	0.03
scenario 1	mean estimate	97.5	97.0	84.3	0.900	−0.093
$\pi_1=\frac{402}{4998}$	true variance	2.8e-05	3.7e-05	7.6e-04	3.8e-04	0.23
	mean estimated variance	2.8e-05	3.8e-05	7.7e-05	3.8e-04	0.22
	estimated 95%CI coverage	92.4	92.4	92.7	92.6	97.0
scenario 2	mean estimate	97.5	97.0	84.3	0.899	−0.033
$\pi_1=\frac{2*402}{4998}$	true variance	1.5e-05	2.1e-05	4.4e-04	2.1e-04	0.11
	mean estimated variance	1.5e-05	2.0e-05	4.3e-05	2.1e-04	0.10
	estimated 95%CI coverage	94.0	94.0	94.0	94.1	95.1

Stat Med. Author manuscript; available in PMC 2013 October 28.

Table 6

Simulation of performance of estimators of diagnostic accuracy using the HPV data from Table 4

		sensitivity	specificity	PPV	NPV
	true	0.737	0.908	0.117	0.995
	full cohort variance	4.8e-03	3.4e-06	3.8e-04	2.3e-06
observed	mean estimate	0.734	0.908	0.117	0.995
sampling	true variance	1.5e-02	7.0e-06	7.6e-04	8.8e-06
fractions	mean estimated variance	1.5e-02	5.8e-06	6.4e-04	1.0e-05
	estimated 95%CI coverage	96.4	92.4	93.5	97.5
Double	mean estimate	0.738	0.908	0.117	0.995
triple-neg	true variance	1.1e-02	7.0e-06	7.7e-04	5.4e-06
sampling	mean estimated variance	1.1e-02	5.8e-06	6.4e-04	5.9e-06
	estimated 95%CI coverage	97.1	91.5	93.7	96.9
	mean estimate	0.736	0.908	0.117	0.995
Sample all	true variance	1.5e-02	6.1e-06	6.7e-04	9.2e-06
triple-pos	mean estimated variance	1.5e-02	5.5e-06	6.0e-04	1.0e-05
	estimated 95%CI coverage	96.6	92.9	94.2	97.4
	mean estimate	0.740	0.908	0.116	0.995
Sample all	true variance	1.4e-02	3.6e-06	3.9e-04	9.0e-06
HPV-pos	mean estimated variance	1.4e-02	3.4e-06	3.7e-04	1.0e-05
	estimated 95%CI coverage	96.6	94.1	95.2	97.3

Stat Med. Author manuscript; available in PMC 2013 October 28.