STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:2170–2183
Published online 23 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2677

Confidence intervals for predictive values with an emphasis to case–control studies

Nathaniel D. Mercaldo[1], Kit F. Lau[2] and Xiao H. Zhou[3,1,*,†]

[1] *Department of Biostatistics, University of Washington, Seattle, WA 98027, U.S.A.*
[2] *Celera Diagnostics, 1401 Harbor Bay Parkway, Alameda, CA, U.S.A.*
[3] *HSR&D Center of Excellence, VA Puget Sound Health Care System, Seattle, WA, U.S.A.*

SUMMARY

The accuracy of a binary-scale diagnostic test can be represented by sensitivity (Se), specificity (Sp) and positive and negative predictive values (PPV and NPV). Although Se and Sp measure the intrinsic accuracy of a diagnostic test that does not depend on the prevalence rate, they do not provide information on the diagnostic accuracy of a particular patient. To obtain this information we need to use PPV and NPV. Since PPV and NPV are functions of both the accuracy of the test and the prevalence of the disease, constructing their confidence intervals for a particular patient is not straightforward. In this paper, a novel method for the estimation of PPV and NPV, as well as their confidence intervals, is developed. For both predictive values, standard, adjusted and their logit transformed-based confidence intervals are compared using coverage probabilities and interval lengths in a simulation study. These methods are then applied to two case–control studies: a diagnostic test assessing the ability of the e4 allele of the apolipoprotein E gene (ApoE.e4) on distinguishing patients with late-onset Alzheimer's disease (AD) and a prognostic test assessing the predictive ability of a 70-gene signature on breast cancer metastasis. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: predictive values; confidence intervals; diagnostic tests; Alzheimer's disease; breast cancer

1. INTRODUCTION

The case–control study design is often used in various types of epidemiological studies, including genetic epidemiology. A case–control study is advantageous over a cohort study in that the collection of samples can be performed in a short period of time. Moreover, the difficulty of controlling for confounding factors and inferring causal relations is its primary disadvantage. In addition, prevalence of the medical endpoints cannot be estimated. Despite these drawbacks and in light of

*Correspondence to: Xiao H. Zhou, Department of Biostatistics, University of Washington, Seattle, WA 98027, U.S.A.

†E-mail: azhou@u.washington.edu

Received 5 August 2005
Accepted 27 June 2006

the efficiency, many genetic and genomic association studies employ the case–control study design in the discovery stage to find genetic or genomic markers associated with the medical endpoints of interest [1, 2]. The molecular markers can be single nucleotide polymorphisms (SNP), mRNA expression profiles, or protein expression profiles. If the goal of the study is to form a molecular diagnostic/prognostic, the associated markers will often be combined to form a classifier/predictor.

While the diagnostic accuracy of a molecular test will ultimately be evaluated in a prospective clinical study with a representative sample from the target population, it is very helpful to evaluate the diagnostic values of genetic markers in discovery research phase using case–control samples [3]. This can aid in the decision of which markers to follow in replication studies. Upon replication, estimation of diagnostic values can help one decide whether to develop a molecular diagnostic test and to go forward with an expensive clinical trial. The accuracy of a binary-scale diagnostic test can be represented by sensitivity (Se), specificity (Sp) and positive and negative predictive values (PPV and NPV). Although Se and Sp measure the intrinsic accuracy of a diagnostic test that does not depend on the prevalence rate, they do not provide information on the diagnostic accuracy on a particular patient. To get this information we need to employ PPV or NPV. Since PPV and NPV are functions of both the inherent accuracy of the test and the prevalence of the disease, constructing confidence intervals for PPV and NPV for a particular patient is not straightforward, regardless of the study design. In this paper, we will derive a new formula for constructing such confidence intervals.

It has been known, due to the discreteness and skewness of the binomial distribution, that the standard estimation of binomial proportions, Se and Sp in our case, and the subsequent confidence interval generation of these proportions is less than adequate, especially when the proportion being estimated is near 0 or 1. This estimation process leads to an oscillatory pattern of coverage probabilities when certain parameters are slightly perturbed. Certain combinations of sample size and prevalence can cause this oscillation to extend well below the nominal coverage probability [4]. Because of this, the standard estimates of PPV and NPV may fail to meet expectations, as well as their associated confidence intervals.

To correct for these problems, a continuity correction is incorporated into the estimation of Se and Sp. This has been shown to greatly improve coverage probability of binomial proportions and we will show that with a better estimate of Se and Sp, more accurate confidence intervals can be developed [4]. Specifically, we will develop two types of confidence intervals. The first of these methods includes the standard confidence interval while the other will utilize the logit transformation. This transformation is used to help achieve normality and alleviate problems associated with estimating proportions near the boundary values of 0 and 1 [5].

Standard estimates of Se, Sp, PPV and NPV as well as adjusted estimates of each are derived in Section 2. Section 3 details a simulation to assess each of these in terms of coverage probability and confidence interval length. These methods are applied to two case–control studies: a diagnostic test for late-onset Alzheimer's disease (AD) and a prognostic test for metastasis in breast cancer patients in Section 4.

2. METHODS

2.1. Estimation of PPV and NPV

Regardless of the study design that was used to evaluate a diagnostic test, the estimation sensitivity and specificity is straightforward. If a case–control study design is used to evaluate the test, then

DOI: 10.1002/sim

Table I. Standard contingency table.

	Disease+	Disease−
Test+	x_{11}	x_{10}
Test−	x_{01}	x_{00}
	n_1	n_0

additional information, namely the prevalence of the disease in the population, is needed to calculate the predictive values of the test. Typically, this is required due to the oversampling of one group over the other as denoted by n_1 and n_0 which represent the numbers of cases and controls, respectively. Therefore, throughout this section the prevalence is either assumed to be estimated from the observed data, as with a cohort study, or obtained from an additional source, as with a case–control study. The information from either of these observational studies is summarized by the contingency table below (Table I).

Sensitivity and specificity of the test in this population will be denoted by Se and Sp and are estimated from Table I by (1).

$$\widehat{\mathrm{Se}}=\frac{x_{11}}{n_1},\quad \widehat{\mathrm{Sp}}=\frac{x_{00}}{n_0} \tag{1}$$

It is of interest to estimate the positive and negative predictive values of a test, PPV and NPV, in a prospective population with a given disease prevalence rate of p. Since Se and Sp of a test are intrinsic and do not depend on the disease prevalence, we can assume that Se and Sp of the test in the case–control population are the same as those in the prospective population, which is already assumed when performing a cohort study. The formulas for PPV and NPV of the test in the prospective population are

$$\mathrm{PPV}=\frac{\mathrm{Se}\cdot p}{\mathrm{Se}\cdot p+(1-\mathrm{Sp})\cdot(1-p)},\quad \mathrm{NPV}=\frac{\mathrm{Sp}\cdot(1-p)}{(1-\mathrm{Se})\cdot p+\mathrm{Sp}\cdot(1-p)} \tag{2}$$

Their logit transformations are

$$\mathrm{logit(PPV)}=\log\left[\frac{\mathrm{Se}\cdot p}{(1-\mathrm{Sp})(1-p)}\right],\quad \mathrm{logit(NPV)}=\log\left[\frac{\mathrm{Sp}\cdot(1-p)}{(1-\mathrm{Se})\cdot p}\right] \tag{3}$$

After obtaining the standard estimates of Se and Sp (1), the standard estimates for PPV and NPV are found by replacing Se and Sp in (2) with $\widehat{\mathrm{Se}}$ and $\widehat{\mathrm{Sp}}$.

To adjust for the inherent drawbacks of the estimation of a binomial proportion, a continuity correction is added to each cell of Table I, resulting in Table II. The addition of a constant, $k^2/2$ where $k=z_{1-\alpha/2}=\Phi^{-1}(1-\alpha/2)$ and $\Phi^{-1}(x)$ denotes the xth quantile of the inverse normal distribution, is related to the Wilson estimation method for finding the midpoint of a skewed distribution [4].

Using Table II, adjusted estimates of Se and Sp are

$$\widetilde{\mathrm{Se}}=\frac{n_1\cdot\widehat{\mathrm{Se}}+\frac{k^2}{2}}{\widetilde{n_1}},\quad \widetilde{\mathrm{Sp}}=\frac{n_0\cdot\widehat{\mathrm{Sp}}+\frac{k^2}{2}}{\widetilde{n_0}} \tag{4}$$

DOI: 10.1002/sim

Table II. Adjusted contingency table using a continuity correction.

	Disease+	Disease−
Test+	$x_{11}+\frac{k^2}{2}$	$x_{10}+\frac{k^2}{2}$
Test−	$x_{01}+\frac{k^2}{2}$	$x_{00}+\frac{k^2}{2}$
	$\widetilde{n_1}=n_1+k^2$	$\widetilde{n_0}=n_0+k^2$

With the adjusted estimates of Se and Sp, using (2) we can obtain adjusted estimates for PPV and NPV. Using these estimates, four different confidence intervals will be constructed.

2.2. *Confidence intervals of PPV and NPV*

The $100(1-\alpha)$ per cent confidence intervals that are generated follow the Wald-type formulation: $A\pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(A)}$, where A is the function being estimated. In this paper, A will include the standard and adjusted estimates of PPV and NPV, as well as their logit transformations. All confidence intervals will be represented by the estimation method of PPV/NPV. For example, if the standard estimate of PPV is calculated and the logit transformation is then performed, then the associated interval is deemed the standard logit interval. Other confidence intervals exist for binomial proportions, such as Wilson's interval and Jefferys interval, but these were not pursued due to the possibility of having more than two roots and not being able to use beta distribution as a prior, respectively.

The variance of each estimate is derived to complete the above formulation of the confidence intervals. Via the binomial theorem, the variances of Se and Sp are

$$\mathrm{Var(Se)}=\frac{\mathrm{Se}\cdot(1-\mathrm{Se})}{n_1},\quad \mathrm{Var(Sp)}=\frac{\mathrm{Sp}\cdot(1-\mathrm{Sp})}{n_0} \tag{5}$$

The delta method was applied to determine the variances of PPV and NPV:

$$\mathrm{Var(PPV)}=\frac{[p\cdot(1-\mathrm{Sp})\cdot(1-p)]^2\cdot\dfrac{\mathrm{Se}\cdot(1-\mathrm{Se})}{n_1}+[p\cdot\mathrm{Se}\cdot(1-p)]^2\cdot\dfrac{\mathrm{Sp}\cdot(1-\mathrm{Sp})}{n_0}}{[\mathrm{Se}\cdot p+(1-\mathrm{Sp})\cdot(1-p)]^4} \tag{6}$$

$$\mathrm{Var(NPV)}=\frac{[\mathrm{Sp}\cdot(1-p)\cdot p]^2\cdot\dfrac{\mathrm{Se}\cdot(1-\mathrm{Se})}{n_1}+[(1-\mathrm{Se})\cdot(1-p)\cdot p]^2\cdot\dfrac{\mathrm{Sp}\cdot(1-\mathrm{Sp})}{n_0}}{[(1-\mathrm{Se})\cdot p+\mathrm{Sp}\cdot(1-p)]^4} \tag{7}$$

The variance of the logit(PPV) and logit(NPV) can be found in a similar manner.

$$\mathrm{Var(logit(PPV))}=\left[\frac{1-\mathrm{Se}}{\mathrm{Se}}\right]\cdot\frac{1}{n_1}+\left[\frac{\mathrm{Sp}}{1-\mathrm{Sp}}\right]\cdot\frac{1}{n_0} \tag{8}$$

$$\mathrm{Var(logit(NPV))}=\left[\frac{\mathrm{Se}}{1-\mathrm{Se}}\right]\cdot\frac{1}{n_1}+\left[\frac{1-\mathrm{Sp}}{\mathrm{Sp}}\right]\cdot\frac{1}{n_0} \tag{9}$$

The variances of the standard and adjusted methods can be fashioned by replacing PPV and NPV with their respected estimates.

As previously mentioned, the $100(1-\alpha)$ per cent confidence intervals for PPV and logit(PPV) are:

$$\text{PPV} \pm z_{1-\alpha/2}\sqrt{\text{Var(PPV)}}, \quad \text{logit(PPV)} \pm z_{1-\alpha/2}\sqrt{\text{Var(logit(PPV))}} \tag{10}$$

For interpretability purposes, as well as for the comparison of estimation methods, the intervals for the logit transformed PPV estimates can be retransformed to their original scale:

$$\text{PPV} : \left[\frac{e^{\text{logit(PPV)}-z_{1-\alpha/2}\sqrt{\text{Var(logit(PPV))}}}}{1+e^{\text{logit(PPV)}-z_{1-\alpha/2}\sqrt{\text{Var(logit(PPV))}}}}, \frac{e^{\text{logit(PPV)}+z_{1-\alpha/2}\sqrt{\text{Var(logit(PPV))}}}}{1+e^{\text{logit(PPV)}+z_{1-\alpha/2}\sqrt{\text{Var(logit(PPV))}}}} \right] \tag{11}$$

Similar intervals for NPV and logit(NPV) are obtained by replacing PPV with NPV in (10) and (11). Again, the evaluation of the following intervals using standard or adjusted estimates will create the standard logit or adjusted logit intervals.

Using the developed results, a simulation study was performed to determine which method was superior based on the criteria of confidence interval length and coverage probability.

3. SIMULATION

3.1. *Description*

In this section we discuss the results from a simulation that assesses the coverage probability and confidence interval length of three proposed methods in comparison to the standard method. There were 36 different combinations for the prevalence, p, (0.05, 0.30, 0.50), Se (0.55, 0.75, 0.85, 0.90), and Sp (0.70, 0.85, 0.90), but only 12 will be discussed in this paper, as seen in Table III. Values and combinations of p, Se, and Sp were determined to roughly mimic the examples of Section 4 as well as providing a more general parameter set in which other studies may find appropriate. Other results not explicitly detailed can be obtained from the author.

Four different values for n_0 and n_1, (25, 50, 100, 400), were used with each design and were chosen to reflect the examples of Section 4, and also to symbolize small, medium and large study sizes. For each of the designs and n_0 and n_1 combination, 10,000 binomial samples were generated using `rbinom` in **R** [6]. With these data, standard contingency tables were generated. From the standard contingency tables, the continuity correction was added to each cell to create the adjusted contingency tables (Table II).

True PPV and NPV values were calculated corresponding to a given set of p, Se, and Sp. This was followed by estimating PPV and NPV via the four previously derived methods. Confidence

Table III. Design configurations for the simulation.

Design	(p, Se, Sp)	Design	(p, Se, Sp)	Design	(p, Se, Sp)
1	(0.05, 0.55, 0.70)	5	(0.30, 0.55, 0.70)	9	(0.50, 0.55, 0.70)
2	(0.05, 0.75, 0.85)	6	(0.30, 0.75, 0.85)	10	(0.50, 0.75, 0.85)
3	(0.05, 0.85, 0.90)	7	(0.30, 0.85, 0.90)	11	(0.50, 0.85, 0.90)
4	(0.05, 0.90, 0.70)	8	(0.30, 0.90, 0.70)	12	(0.50, 0.90, 0.70)

DOI: 10.1002/sim

Table IV. Summary of PPV coverage probabilities where the cell values denote the coverage probability for a fixed design parameter and averaging over the remaining parameters.

Design parameter	Standard	Standard logit	Adjusted	Adjusted logit
$p = 0.05$	0.9351	0.9540	0.8900	0.9203
$p = 0.30$	0.9328	0.9536	0.9300	0.9200
$p = 0.50$	0.9321	0.9536	0.9425	0.9198
$\mathrm{Se} = 0.55$	0.9422	0.9515	0.9406	0.9517
$\mathrm{Se} = 0.75$	0.9286	0.9558	0.9156	0.9137
$\mathrm{Se} = 0.85$	0.9213	0.9577	0.9000	0.8818
$\mathrm{Se} = 0.90$	0.9412	0.9499	0.9271	0.9329
$\mathrm{Sp} = 0.70$	0.9417	0.9507	0.9338	0.9423
$\mathrm{Sp} = 0.85$	0.9286	0.9558	0.9156	0.9137
$\mathrm{Sp} = 0.90$	0.9213	0.9577	0.9000	0.8818
$n_0 = 25$	0.9104	0.9566	0.8920	0.8862
$n_0 = 50$	0.9325	0.9552	0.9163	0.9171
$n_0 = 100$	0.9426	0.9521	0.9291	0.9304
$n_0 = 400$	0.9478	0.9509	0.9459	0.9465
$n_1 = 25$	0.9343	0.9539	0.9184	0.9183
$n_1 = 50$	0.9337	0.9548	0.9206	0.9210
$n_1 = 100$	0.9332	0.9536	0.9208	0.9195
$n_1 = 400$	0.9320	0.9527	0.9235	0.9214
Overall	0.9333	0.9537	0.9208	0.9200

interval lengths and coverage probabilities were then calculated. The possibility of obtaining a standard estimate of PPV or NPV of 1 led to the potential problem of division by zero when applying the logit transformation. Throughout these simulations when this occurred the point estimate and its confidence interval were coded as 'NA', and thus the confidence interval length and coverage probability were not applicable. The number of times this occurred was recorded for further analysis when the methods were compared. It was also a possibility that the calculated confidence intervals extended outside their allowed bounds, [0, 1]. Throughout these simulations these occurrences were noted, but not modified so that the [0, 1] condition was not forced to be satisfied.

Simulation results for PPV and NPV can be found in Tables IV and V and Tables VI and VII, respectively, where each table represents either confidence interval length or coverage probability summaries for each predictive value. Due to the number of input parameters of this simulation and the number of possible values of each parameter, these tables were constructed by holding one parameter constant and averaging over the remaining parameters. For example, if p is fixed at 0.05, then the interval lengths and coverage probabilities associated with this prevalence value are the average of all possible design and (n_0, n_1) configuration's interval lengths and coverage probabilities with $p = 0.05$, i.e. $(p = 0.05,\ \mathrm{Se} = 0.55,\ \mathrm{Sp} = 0.70,\ n_0 = 100,\ n_1 = 100)$, $(p = 0.05,\ \mathrm{Se}, = 0.75,\ \mathrm{Sp} = 0.85,\ n_0 = 25,\ n_1 = 400)$, etc. In Tables IV–VII, the rows associated with $\mathrm{Se} = 0.70$ and $\mathrm{Sp} = 0.85$, as well as $\mathrm{Se} = 0.85$ and $\mathrm{Sp} = 0.90$ are identical due to the design configuration of Table III. The values were not omitted due to the increased interpretability when included.

From Tables IV and V, the estimation of PPV via the standard logit method is 'superior' to the other tested methods in terms of coverage probabilities with an overall value of 0.9537 and

Table V. Summary of PPV confidence interval lengths where the cell values denote the confidence interval length for a fixed design parameter and averaging over the remaining parameters.

Design parameter	Standard	Standard logit	Adjusted	Adjusted logit
$p=0.05$	0.1819	0.1771	0.1409	0.1398
$p=0.30$	0.2234	0.2237	0.2138	0.2102
$p=0.50$	0.1587	0.1630	0.1579	0.1583
$\text{Se}=0.55$	0.1677	0.1656	0.1545	0.1528
$\text{Se}=0.75$	0.2120	0.2110	0.1891	0.1876
$\text{Se}=0.85$	0.2300	0.2338	0.2067	0.2051
$\text{Se}=0.90$	0.1423	0.1414	0.1330	0.1322
$\text{Sp}=0.70$	0.1550	0.1535	0.1438	0.1425
$\text{Sp}=0.85$	0.2120	0.2110	0.1891	0.1876
$\text{Sp}=0.90$	0.2300	0.2338	0.2067	0.2051
$n_0=25$	0.2916	0.2938	0.2498	0.2465
$n_0=50$	0.2143	0.2127	0.1948	0.1934
$n_0=100$	0.1557	0.1550	0.1488	0.1482
$n_0=400$	0.0904	0.0902	0.0900	0.0898
$n_1=25$	0.2006	0.2002	0.1846	0.1828
$n_1=50$	0.1896	0.1896	0.1727	0.1713
$n_1=100$	0.1833	0.1834	0.1658	0.1645
$n_1=400$	0.1784	0.1785	0.1604	0.1592
Overall	0.1880	0.1879	0.1708	0.1695

the adjusted logit method generated the shortest interval length of 0.1695. Overall, the standard estimation methods yielded higher coverage probabilities along with having slightly longer interval lengths than the adjusted methods. Among the methods investigated, the interval lengths were fairly uniform only differing by at most 0.02, but only the standard logit method attained the desired 95 per cent nominal level. In terms of the parameters, typically only the numbers of cases and controls (n_1, n_0) had a clear effect on the coverage probability and interval length, while p, Se, and Sp fluctuated.

Regardless of n_0 or n_1, as either increased the interval length decreased among all estimation methods. The increase in n_1 did not drastically affect the coverage probabilities for the standard methods, but a slight increase was noted for the adjusted methods. The increase in p only had a positive effect on the adjusted logit coverage probability, while the other methods remained relatively unchanged. There was not a clear effect of the increase in prevalence on interval length. As Se increased, coverage probabilities tended to decrease except for the standard logit method where a minuscule increase was noted and their associated interval lengths tended to increase regardless of the estimation method. These previous observations were deviated from when $\text{Se}=0.90$ where the opposite effect was observed. As with Se, similar patterns were observed with an increase in Sp.

From Tables VI and VII, overall coverage probabilities and interval lengths of the standard logit, $(0.9540, 0.0863)$, and adjusted, $(0.9527, 0.0838)$, estimation methods were virtually identical and only differed after the hundredth digit, and thus were considered the ‘superior’ estimation methods of NPV’s. Both of these methods attained the desired 95 per cent threshold while maintaining interval lengths comparable to the other methods.

DOI: 10.1002/sim

Table VI. Summary of NPV coverage probabilities where the cell values denote the coverage probability for a fixed design parameter and averaging over the remaining parameters.

Design parameter	Standard	Standard logit	Adjusted	Adjusted logit
$p = 0.05$	0.9333	0.9543	0.9608	0.9212
$p = 0.30$	0.9334	0.9538	0.9533	0.9202
$p = 0.50$	0.9342	0.9540	0.9440	0.9207
Se = 0.55	0.9442	0.9497	0.9576	0.9559
Se = 0.75	0.9387	0.9528	0.9499	0.9315
Se = 0.85	0.9278	0.9558	0.9488	0.9025
Se = 0.90	0.9238	0.9578	0.9545	0.8929
Sp = 0.70	0.9340	0.9537	0.9560	0.9244
Sp = 0.85	0.9387	0.9528	0.9499	0.9315
Sp = 0.90	0.9278	0.9558	0.9488	0.9025
$n_0 = 25$	0.9333	0.9548	0.9587	0.9186
$n_0 = 50$	0.9341	0.9544	0.9525	0.9218
$n_0 = 100$	0.9342	0.9534	0.9503	0.9203
$n_0 = 400$	0.9329	0.9536	0.9492	0.9221
$n_1 = 25$	0.9172	0.9548	0.9483	0.8888
$n_1 = 50$	0.9297	0.9561	0.9508	0.9148
$n_1 = 100$	0.9410	0.9541	0.9541	0.9324
$n_1 = 400$	0.9466	0.9512	0.9576	0.9468
Overall	0.9336	0.9540	0.9527	0.9207

As with the PPV estimates, the increase in n_0 and n_1 clearly decreased all NPV interval lengths. Contrary to the PPV estimates, the increase in n_0 did not drastically alter the NPV's coverage probabilities among the estimation methods, but the increase in n_1 increased coverage probabilities for all estimation methods except for the standard logit method. The increase in p only appeared to have a noticeable negative effect on the coverage probability of the adjusted method, whereas other methods remained unchanged. The interval lengths greatly increased with the increase in p, but all methods increased at the same rate. As Se increased, the coverage probabilities decreased except for the standard logit method and for all methods the interval length decreased. Similar results for the increase in Sp are observed for interval length, while coverage probabilities remain relatively unchanged except for a slight decrease with the adjusted method.

Superior is conveniently placed in quotes due to previously mentioned caveat that if estimates of PPV and NPV are 1, then the standard logit method is not applicable. Throughout this simulation study whichever iteration generated this estimate of PPV or NPV the standard logit method would be skipped with the total number of skipped iterations being tallied. Many factors contribute to the occurrence of a predictive value estimate of 1, such as n_0, n_1, Se, and Sp. It was noted that many of the parameter configurations did not generate these estimated values, but of those that did the percentage of 'problematic' estimates was as high as ~ 7.5 per cent. For PPV estimates, these typically occurred when n_0 was small, 25, and Sp>Se, with both Sp and Se being large, 0.90 and 0.85. Similar occurrences happened with NPV estimates, but n_1 was small, 25, and Se>Sp, 0.90 and 0.70. Even though various proportions of iterations were not executed, the sample size of each simulation was large enough that the results of the standard logit method remained valid.

DOI: 10.1002/sim

Table VII. Summary of NPV confidence interval lengths where the cell values denote the confidence interval length for a fixed design parameter and averaging over the remaining parameters.

Design parameter	Standard	Standard logit	Adjusted	Adjusted logit
$p=0.05$	0.0135	0.0145	0.0140	0.0147
$p=0.30$	0.0855	0.0903	0.0877	0.0899
$p=0.50$	0.1490	0.1540	0.1498	0.1506
$\mathrm{Se}=0.55$	0.0944	0.0940	0.0917	0.0913
$\mathrm{Se}=0.75$	0.0828	0.0839	0.0818	0.0823
$\mathrm{Se}=0.85$	0.0725	0.0767	0.0747	0.0765
$\mathrm{Se}=0.90$	0.0810	0.0906	0.0871	0.0901
$\mathrm{Sp}=0.70$	0.0877	0.0923	0.0894	0.0907
$\mathrm{Sp}=0.85$	0.0828	0.0839	0.0818	0.0823
$\mathrm{Sp}=0.90$	0.0725	0.0767	0.0747	0.0765
$n_0=25$	0.0897	0.0934	0.0930	0.0942
$n_0=50$	0.0834	0.0871	0.0848	0.0860
$n_0=100$	0.0801	0.0837	0.0805	0.0817
$n_0=400$	0.0773	0.0809	0.0770	0.0782
$n_1=25$	0.1252	0.1359	0.1267	0.1294
$n_1=50$	0.0936	0.0963	0.0946	0.0961
$n_1=100$	0.0699	0.0707	0.0710	0.0716
$n_1=400$	0.0421	0.0422	0.0430	0.0431
Overall	0.0827	0.0863	0.0838	0.0850

In addition to having the standard logit method not being able to be computed, the overflowing of intervals lower than 0 and greater than 1 was a significant problem with the untransformed standard and adjusted methods. For the standard, untransformed interval estimate of PPV, the upper interval estimate exceeded the 1 up to 47 per cent of the iterations for a single parameter configuration. This remarkably high percentage of non-sensical confidence intervals typically stemmed from a small n_0 value, 25, and large values of Se and Sp, 0.85 and 0.90. Under similar parameter configurations, the adjusted, untransformed estimates of PPV produced these types of intervals, but not of the magnitude of the standard estimates, ~ 7.5 per cent.

This phenomenon was observed with the estimation of NPV as well. Regardless of the n_0 value, the design configuration when Se, Sp and n_1 were 0.85, 0.90 and 25, respectively, resulted in 46 per cent of the simulated standard, untransformed intervals to exceed the upper bound of 1. When the combination of Se and Sp was changed to 0.90 and 0.70, this percentage increased to 69 per cent when n_1 was 25. The adjusted, untransformed upper limit estimate of NPV resulted in 1 being exceeded 9 and 27 per cent using the previously mentioned design configurations. Based on these observations, the results of the standard logit method, as well as the untransformed standard and adjusted methods need to be approached cautiously.

Each evaluated method produced comparable interval lengths, and thus the preferred method would be able to be constantly applied, regardless of estimate value, as well as producing meaningful interval ranges and have a coverage probability that asymptotically attains the 95 per cent desired threshold. None of the evaluated methods meet all of these ideal criteria, but if the estimate of the predictive values is not 1, then the standard logit method is recommended. If the estimate of the predictive value is 1, then the adjusted method is preferred. The standard and adjusted methods yield almost identical results in terms of PPV, but the coverage probability of only the adjusted method reached the nominal 95 per cent level and it is because of this that the adjusted method is recommended.

4. APPLICATIONS

To illustrate the utility of the above methods to estimate predictive values and their confidence intervals, two examples in molecular diagnostics and prognostics will be examined. The first example is to estimate predictive values using the e4 allele of the apolipoprotein E gene (ApoE.e4) as a molecular diagnostic for AD. The second example is the use of a gene-expression signature as prognosticator for metastasis of breast cancer tumours.

4.1. Alzheimer's disease

Alzheimer's disease (AD) is a common disease among elderly individuals. The prevalence of AD depends upon age which ranges between 3 and 50 per cent among individuals between the ages of 65 and 90 [7, 8]. Improving the diagnostic accuracy of AD will aid in the early and correct identification of proper treatment regimes. It has been well established that ApoE.e4 is associated with an increased susceptibility to AD [9–11]. Studies have also been carried out to evaluate the usefulness of ApoE.e4 genotype in the diagnosis of AD among persons with dementia [12]. The general conclusion is that while ApoE.e4 genotyping does not provide sufficient Se and Sp to be used alone as a diagnostic test for AD, but when in combination with clinical criteria, it improves the Sp of the diagnosis. Diagnostic utility of genotyping tests can also be improved when additional susceptibility genes are found and added to the diagnostic genotyping panel [13].

Li *et al.* [1] have recently performed a case–control study to identify other susceptibility genes for AD. This Washington University study recruited 418 cases and 375 controls, with a known ApoE genotype, of which 240 of the cases and 87 of the controls carried at least one ApoE.e4 allele. A positive molecular diagnostic for AD was defined if the ApoE.e4 allele was present and negative if absent. These data are summarized in Table VIII.

From Table VIII, Se of the APoE.e4 genotyping test is estimated to be 0.574 (95 per cent CI 0.526–0.621) while the Sp is estimated to be 0.768 (95 per cent CI 0.723–0.808) using the standard method. Moreover, PPV and NPV cannot be directly estimated from the above table due to the study design. Cases have been oversampled (52.7 per cent of the whole sample) as compared to the prevalences in the general population, especially in the younger age group.

Using disease prevalence of 3 and 50 per cent, estimates of PPV and NPV, their 95 per cent confidence intervals and interval lengths were calculated using the four described methods. These results are summarized in Table IX.

Table VIII. Alzheimer's disease data.

	Case	Control
ApoE.e4+	240	87
ApoE.e4−	178	288
	418	375

DOI: 10.1002/sim

Table IX. Alzheimer's disease results.

	PPV*			PPV†		
Method	Estimate	95% CI‡	CI length	Estimate	95% CI	CI length
Standard	0.0711	0.0578–0.0844	0.0267	0.7122	0.6709–0.7536	0.0827
Standard logit	0.0711	0.0589–0.0856	0.0268	0.7122	0.6692–0.7518	0.0826
Adjusted	0.0703	0.0572–0.0833	0.0261	0.7096	0.6685–0.7507	0.0823
Adjusted logit	0.0703	0.0583–0.0845	0.0262	0.7096	0.6668–0.7489	0.0821
	NPV*			NPV†		
Standard	0.9831	0.9811–0.9852	0.0041	0.6433	0.6147–0.6719	0.0571
Standard logit	0.9831	0.9809–0.9851	0.0041	0.6433	0.6143–0.6713	0.0570
Adjusted	0.9831	0.9810–0.9851	0.0041	0.6421	0.6137–0.6706	0.0570
Adjusted logit	0.9831	0.9809–0.9850	0.0041	0.6421	0.6132–0.6701	0.0569

*Prevalence = 0.03.
†Prevalence = 0.50.
‡CI = confidence interval.

The standard methods, both transformed and untransformed, usually produced a larger estimate than their adjusted counterparts. These differences were remarkably small, <1 per cent, and did not have a significant effect on the overall interval length (Table IX). The point estimates of NPV were 98 per cent when the prevalence was 3 per cent, thus indicating if a patient tested negative for the ApoE.e4 gene, then the investigators were almost certain that the patient did not have AD. As the prevalence increased to 50 per cent, then the NPV estimate decreased to 64 per cent while the PPV estimate increased from 7 to 71 per cent. Regardless of the estimation method, similar intervals were obtained, which results from the large sample size.

4.2. Breast cancer

Breast cancer is the most common cancer among women in the United States, accounting for almost 30 per cent of all newly diagnosed cancers [14]. It is also the most common cancer in women worldwide. Various treatment options are available once the diagnosis has been made and primary tumours are removed. While adjuvant therapy, such as chemotherapy or hormonal therapy, reduces the risk of distant metastasis, 70–80 per cent of patients receiving the treatment would have survived without it [2]. Accurate prognosis of breast cancer patients, especially those at early stage and node negative, can help to make correct treatment decisions. Moreover, many traditional predictors for metastases, such as tumour grade, fail to accurately classify breast tumours according to their clinical behaviour. Recently, several studies have demonstrated that gene expression profile of tumours removed from breast cancer patients can be utilized for prognosis of metastasis [2, 15, 16]. van 't Veer *et al.* [2] performed a case–control study to develop a gene signature as a prognosticator of breast cancer patients. Cases were defined as those that have metastasis within 5 years of tumour excision, while controls were those that did not. Each tumour was classified as having a good or poor gene signature which was defined as having a signature correlation coefficient above or below the 'optimized sensitivity' threshold, respectively. This 'optimized sensitivity' threshold is defined as the correlation value that would result in a misclassification of at most 10 per cent of the cases. The performance of their 70-gene signature as prognosticator for metastasis is summarized in Table X.

From Table X, the Se of the molecular signature is estimated to be 0.91 (95 per cent CI 0.77–0.97) while Sp is estimated to be 0.73 (95 per cent CI 0.58–0.84) using the standard method. As in the example with AD, PPV and NPV cannot be directly estimated as cases that have metastases in 5 years have been oversampled (43.6 per cent of all samples) as compared to the general population of breast cancer patients. In their follow-up paper to validate their 70-gene prognosticator of survival in breast cancer patients, metastasis-free survival rates in 5 years in node-negative patients with tumour excision can be estimated from the supplementary data on their website [15]. The metastasis-free survival rates in 5 years are estimated to be 70 per cent for all node-negative patients, 93.4 per cent for patients with good signature, and 56.2 per cent for those with bad signature. Based on these estimates, PPV and NPV of the 70-gene signature will be estimated with prevalences of 7, 30, and 44 per cent for those who metastasize in 5 years (see Table XI).

Table X. Breast cancer data.

	Case	Control
Poor signature	31	12
Good signature	3	32
	34	44

Table XI. Breast cancer results.

Method	Estimate	95% CI[§]	CI length	Estimate	95% CI	CI length
	PPV[*]			NPV[*]		
Standard	0.2010	0.1217–0.2803	0.1586	0.9910	0.9811–1.0010	0.0197
Standard logit	0.2010	0.1331–0.2919	0.1588	0.9910	0.9734–0.9970	0.0235
Adjusted	0.1837	0.1148–0.2526	0.1377	0.9864	0.9750–0.9977	0.0227
Adjusted logit	0.1837	0.1245–0.2626	0.1382	0.9864	0.9689–0.9941	0.0252
	PPV[†]			NPV[†]		
Standard	0.5889	0.4694–0.7085	0.2390	0.9506	0.8991–1.0020	0.1029
Standard logit	0.5889	0.4665–0.7013	0.2347	0.9506	0.8654–0.9829	0.1175
Adjusted	0.5617	0.4486–0.6747	0.2261	0.9271	0.8701–0.9841	0.1140
Adjusted logit	0.5617	0.4474–0.6698	0.2224	0.9271	0.8455–0.9673	0.1218
	PPV[‡]			NPV[‡]		
Standard	0.7243	0.6257–0.8229	0.1972	0.9130	0.8259–1.0000	0.1741
Standard logit	0.7243	0.6159–0.8114	0.1956	0.9130	0.7781–0.9691	0.1910
Adjusted	0.7014	0.6053–0.7976	0.1923	0.8740	0.7812–0.9669	0.1858
Adjusted logit	0.7014	0.5975–0.7881	0.1906	0.8740	0.7490–0.9416	0.1926

[*] Prevalence = 0.07.
[†] Prevalence = 0.30.
[‡] Prevalence = 0.44.
[§] CI = confidence interval.

As with the AD example, the standard methods produced larger predictive values than the adjusted estimates ranging in an increase of 0.4–10 per cent depending on the predictive value and prevalence value, as well as slightly wider confidence intervals. As expected when the prevalence increased so did the PPV estimate, while the NPV estimate decreased. There was not a clear relationship between interval length and prevalence for PPV, which coincides with the results from the simulation (Table V). The interval lengths for NPV estimates do clearly increase as p increases. There is considerable overlap among the intervals generated by each method, but intervals associated with the standard estimate should be questioned. The upper bound for each standard NPV interval borders or exceeds 1, which results from the small sample sizes.

5. DISCUSSION

Four methods for estimating confidence intervals of predictive values were compared using coverage probabilities and interval lengths via a simulation study. The methods which were investigated extended the well received ideas of correcting the standard estimation of a binomial proportion by incorporating a continuity correction to that of predictive values as well as performing the logit transformation to minimize the adverse effects of estimating proportions near 0 and 1. Based on the results of the simulation study, the adjustment using the logit transformation outperformed the addition of the continuity correction in terms of coverage probabilities. Interval lengths were fairly uniform across estimation methods, and thus did not enter the decision making process of which method was 'superior'.

Under the evaluated combinations of p, Se, Sp, n_0, and n_1, the standard logit method generated overall coverage probabilities greater than its adjusted counterpart. The adjusted logit method is advantageous of the standard logit method in that the requirement of the predictive value estimate not be 1 is not enforced. The standard logit method is recommended solely on having a 3.6 per cent greater coverage probability. Both logit methods are preferred over the untransformed methods due to the logit method's inability to create interval bounds outside the permitted range.

The two examples demonstrated the effects of sample size on the various interval estimation methods. The Alzheimer's disease example displayed the similarities between estimation results among each of the evaluated methods when both n_0 and n_1 were large. The breast cancer example exemplified the power of the logit function not allowing for the creation of non-sensical results, standard *versus* standard logit, when n_0 and n_1 were small.

Based on these results, if the sample size is 'large enough', then the estimation method for a predictive value's confidence interval varies minutely between the tested estimation methods. As with the binomial proportion, if the sample size is relatively small then further caution needs to be applied in the estimation of the predictive value and its confidence interval. Overall, the standard logit method is recommended for the estimation of the confidence interval for a predictive value. It is superior over the other evaluated methods in terms of interval length and coverage probability, when the point estimate can be calculated, i.e. the estimate of the predictive value is not 1. If the logit transform cannot be applied, then the adjusted method is preferred. The untransformed methods generated nearly identical results for PPV, and thus this recommendation is based solely on the adjusted method's ability to attain the nominal coverage probability level for NPV.

DOI: 10.1002/sim

ACKNOWLEDGEMENTS

Professor Xiao-Hua Zhou is presently a Core Investigator and Senior Biostatistician at the Northwest HSR&D Center of Excellence within the VA Puget Sound Health Care System. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

REFERENCES

1. Li Y *et al.* Association of late-onset Alzheimer's disease with genetic variation in multiple members of the GAPD gene family. *Proceedings of the National Academy of Sciences*, *U.S.A.*, 2004; **101**:15688–15693.
2. van 't Veer L *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**:530–535.
3. Obuchowski N, Zhou XH. Prospective studies of diagnostic test accuracy when disease prevalence is low. *Biostatistics* 2002; **3**:477–492.
4. Brown LD, Cal TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001; **16**:101–133.
5. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
6. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2005.
7. Lautenschlager NT *et al.* Risk of dementia among relatives of Alzheimer's disease patients in the MIRAGE study: what is in store for the oldest old? *Neurology* 1996; **46**:641–650.
8. Evans DA *et al.* Prevalence of Alzheimer's disease in a community population of older persons. *JAMA* 1989; **262**:2551–2556.
9. Strittmatter WJ *et al.* Apolipoprotein E: high-avidity binding to beta-amlyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences* 1993; **90**:1977–1981.
10. Saunders AM *et al.* Association of apolipoprotein E allele e4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* 1993; **43**:1467–1472.
11. Poirier J *et al.* Apolipoprotein E polymorphism and Alzheimer's disease. *Lancet* 1993; **342**:697–699.
12. Mayeux R *et al*. Utility of the Apolipoprotein E genotype in the diagnosis of Alzheimer's disease. *New England Journal of Medicine* 1998; **338**:506–511.
13. Yang Q *et al.* Improving the prediction of complex diseases by testing for multiple disease susceptibility genes. *American Journal of Human Genetics* 2003; **72**:639–649.
14. Schottenfeld D, Fraumeni J (eds). *Cancer Epidemiology and Prevention* (2nd edn). Oxford University Press: New York, 1996.
15. van de Vijver MJ *et al.* A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 2002; **25**:1999–2009.
16. Paik S *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* 2004; **351**:2817–2816.

DOI: 10.1002/sim