EFFECTS OF VERIFICATION AND IMPERFECT REFERENCE STANDARD BIASES ON THE ESTIMATED PREVALENCE RATE

Xiao-Hua Zhou *†

December 2, 1996

Key Words: Missing-data, EM algorithm, Verification bias, Imperfect gold standard

Abstract

Two types of biases may exist when estimating prevalence rates from a two-stage design study: (1) verification bias and (2) imperfect reference bias. In this paper, we study the effects of both types of biases on the estimated prevalence rate, and derive the ML estimator for the prevalence rate, adjusting for both biases.

1. INTRODUCTION

In epidemiologic research, the prevalence rate of a disease is often estimated with data from a two stage design study [1]. The first stage assesses a large sample with a screening test. Based on performance on the screening test, some of the screened subjects are selected for a reference standard test, a more expensive clinical assessment, for diagnosis of the disease. Therefore, not all screened subjects have diagnoses of the disease. Even for those who have the diagnosis of the disease, the diagnosis may be wrong because the clinical assessment may not be 100% accurate.

Therefore, in estimating prevalence rates of the disease, two common problems are (1) verification bias and (2) imperfect reference bias. Verification bias may occur if only a subset of screened subjects have verified disease status; and imperfect reference bias may occur if an imperfect reference standard is used to establish the disease status.

In this paper, we study the effects of both biases on the estimation of the prevalence rate. To emphasize conceptual rather than technical issues we will first concentrate our discussions on a simple setting. This simple setting assumes (1) that the selection probability of a patient for the clinical assessment depends only on his/her screening test result, (2) the screening test and the reference standard test are conditionally independence given the true disease status, and (3) that the sensitivity and specificity of the imperfect reference standard are known. Then, in Section 4, we extend our method to more general settings.

2. METHODS

Denote the screening test result, the result of an imperfect reference standard test, and the true disease status of a patient by T, R, D, respectively. Let V be the verification indicator. That is, $V = 1$ if a patient is selected for the clinical assessment; and $V = 0$ if a patient is not selected for the clinical assessment. The variables T and R are binary. We never observe D, and we only observed R if $V = 1$. Let n_k denote the number of subjects whose screening test result is k. Let u_k be the number of not clinically assessed subjects whose screening test result is k. Let s_k and r_k denote the number of subjects who were clinically assessed as diseased and non-diseased, respectively. Table 1 summarizes the observed data. Denote S_R and SP_R be the known sensitivity and specificity of the reference standard test, respectively. Let θ and η be the sensitivity and specificity of the screening tests T.

2.1 INTUITIVE ESTIMATORS

Let $\#(A)$ denote the number of elements in a set A, and that $n = n_0 + n_1$. Since $P(R, T \mid D) = P(R \mid$

*Division of Biostatistics, Indiana University School of Medicine, and Regenstrief Institute for Health Care.

†Correspondence concerning this article should be addressed to the author at Division of Biostatistics, Indiana University, Riley Research, RR 135, Indianapolis, IN 46202-5200. Supported by PHS grants P30AG10133 and AHCPR R29HS08559

Table 1: Observed Data Y_o

		$T=1$	$T=0$
$V=1$	R=1	s_1	s_0
	R=0	r_1	r_0
$V=0$		u_1	u_0
		n_1	n_0

$D)P(T \mid D)$,

$$\#(R=1,T=1) = S_R\theta\#(D=1)+(1-SP_R)(1-\eta)\#(D=0),$$
$$\#(R=1,T=0) = S_R(1-\theta)\#(D=1)+(1-SP_R)\eta\#(D=0),$$
$$\#(R=0,T=1) = (1-S_R)\theta\#(D=1)+SP_R(1-\eta)\#(D=0),$$

and

$$\#(T=0) = (1-\theta)\#(D=1) + \eta\#(D=0),$$
$$\#(T=1) = \theta\#(D=1) + (1-\eta)\#(D=0).$$

Then, solving the equations above gives us that

$$\theta = \frac{\#(T=1)SP_R - \#(T=1,R=0)}{\#(D=1)(S_R+SP_R-1)},$$
$$\eta = \frac{\#(T=0)S_R - \#(T=0,R=1)}{\#(D=0)(S_R+SP_R-1)},$$

and

$$\#(D=1) = \frac{\#(R=1) - n(1-SP_R)}{S_R+SP_R-1}.$$

Under the assumption that the probability of verifying a patient depend on only T, we have that $P(V \mid R,T) = P(V \mid T)$, which implies that

$$\#(R=1,T=1) = \frac{s_1}{r_1+s_1}n_1$$

and

$$\#(R=1,T=0) = \frac{s_0}{r_0+s_0}n_0.$$

Thus, the intuitive estimators for p, θ, and η are

$$\theta = \frac{\frac{n_1}{n}(SP_R - \frac{r_1}{r_1+s_1})}{\frac{s_1}{r_1+s_1}\frac{n_1}{n} + \frac{s_0}{s_0+r_0}\frac{n_0}{n} - (1-SP_R)},$$
$$\eta = \frac{\frac{n_0}{n}(S_R - \frac{s_0}{r_0+s_0})}{\frac{r_1}{r_1+s_1}\frac{n_1}{n} + \frac{r_0}{s_0+r_0}\frac{n_0}{n} - (1-S_R)},$$
$$\hat{p} = \frac{\frac{s_1}{r_1+s_1}\frac{n_1}{n} + \frac{s_0}{r_0+s_0}\frac{n_0}{n} - (1-SP_R)}{S_R+SP_R-1}. \quad (1)$$

provided that

$$S_R \geq max(\frac{s_0}{r_0+s_0}, \frac{s_1}{r_1+s_1}),$$
$$SP_R \geq max(\frac{r_0}{r_0+s_0}, \frac{r_1}{r_1+s_1}). \quad (2)$$

2.2 ML ESTIMATORS

In this section, we show that the intuitive estimators derived in the previous section are actually ML estimators. When we use an imperfect reference standard to establish the true disease status D, we can treat D missing for all subjects. We will use the EM algorithm to derive the ML estimators. The EM algorithm is a general iterative method for finding ML estimates in the missing-data problem [3]. The E step finds the conditional expectation of the complete data log-likelihood function, given the values of the parameters and the observed data. The M step maximizes the conditional expectation of the log-likelihood derived from the E step.

Under the assumption that

$$P(V \mid T,R,D) = P(V \mid T,R) = P(V \mid T)$$

and that

$$P(T,R \mid D) = P(T \mid D)P(R \mid D),$$

the log-likelihood for the complete data that would be observed if all subjects have clinical assessments and have known true disease status is

$$l(\theta,\eta,p) = \sum_{i=1}^{n} D_iR_i\log S_R + D_i(1-R_i)\log(1-S_R) + (1-D_i)R_i\log(1-SP_R) + (1-D_i)(1-R_i)\log SP_R + D_iT_i\log\theta + D_i(1-T_i)\log(1-\theta) + (1-D_i)T_i\log(1-\eta) + (1-D_i)(1-T_i)\log\eta + D_i\log p + (1-D_i)\log(1-p).$$

Let $(\theta^{(t)},\eta^{(t)},p^{(t)})$ be the current values of (θ,η,p) after t cycles of the EM algorithm. Let $P^{(t)}(.)$ be the conditional probability given the observed data and the current values of $(\theta,\eta,p) = (\theta^{(t)},\eta^{(t)},p^{(t)})$. After some algebraic manipulation, we obtain the next estimates for (θ,η,p):

$$\eta^{(t+1)} = \frac{r_1(1-P_{01}^{(t)}) + s_1(1-P_{11}^{(t)}) + u_1P^{(t)}(D=0 \mid T=1)}{\sum_{j=0}^{1} r_j(1-P_{0j}^{(t)}) + s_j(1-P_{1j}^{(t)}) + u_jP^{(t)}(D=0 \mid T=j)},$$

$$\theta^{(t+1)} = \frac{r_1P_{01}^{(t)} + s_1P_{11}^{(t)} + u_1P^{(t)}(D=1 \mid T=1)}{\sum_{j=0}^{1} r_jP_{0j}^{(t)} + s_jP_{1j}^{(t)} + u_jP^{(t)}(D=1 \mid T=j)},$$

$$p^{(t+1)} = \frac{\sum_{j=0}^{1} r_j P_{0j}^{(t)} + s_j P_{1j}^{(t)} + u_j P^{(t)}(D=1 \mid T=j)}{n},$$

where

$$P_{11}^{(t)} = \frac{S_R \theta^{(t)} p^{(t)}}{S_R \theta^{(t)} p^{(t)} + (1 - SP_R)(1 - \eta^{(t)})(1 - p^{(t)})},$$

$$P_{10}^{(t)} = \frac{S_R (1 - \theta^{(t)}) p^{(t)}}{S_R (1 - \theta^{(t)}) p^{(t)} + (1 - SP_R)\eta^{(t)}(1 - p^{(t)})},$$

$$P_{01}^{(t)} = \frac{(1 - S_R)\theta^{(t)} p^{(t)}}{(1 - S_R)\theta^{(t)} p^{(t)} + SP_R(1 - \eta^{(t)})(1 - p^{(t)})},$$

$$P_{00}^{(t)} = \frac{(1 - S_R)(1 - \theta^{(t)}) p^{(t)}}{(1 - S_R)(1 - \theta^{(t)}) p^{(t)} + SP_R \eta^{(t)}(1 - p^{(t)})},$$

$$P^{(t)}(D=1 \mid T=0) = \frac{(1 - \theta^{(t)}) p^{(t)}}{(1 - \theta^{(t)}) p^{(t)} + \eta^{(t)}(1 - p^{(t)})},$$

$$P^{(t)}(D=1 \mid T=1) = \frac{\theta^{(t)} p^{(t)}}{\theta^{(t)} p^{(t)} + (1 - \eta^{(t)})(1 - p^{(t)})}.$$

If we take our intuitive estimates as our initial estimates for θ,η, and p, then

$$P^{(0)}(D=1 \mid S=1, T=1) = \frac{S_R(SP_R - \frac{r_1}{r_1+s_1})}{\frac{s_1}{r_1+s_1}(S_R + SP_R - 1)},$$

$$P^{(0)}(D=1 \mid S=1, T=0) = \frac{S_R(SP_R - \frac{r_0}{r_0+s_0})}{\frac{s_0}{r_0+s_0}(S_R + SP_R - 1)},$$

$$P^{(0)}(D=1 \mid S=0, T=1) = \frac{(1 - S_R)(SP_R - \frac{r_1}{r_1+s_1})}{\frac{r_1}{r_1+s_1}(S_R + SP_R - 1)},$$

$$P^{(0)}(D=1 \mid S=0, T=0) = \frac{(1 - S_R)(SP_R - \frac{r_0}{r_0+s_0})}{\frac{r_0}{r_0+s_0}(S_R + SP_R - 1)},$$

$$P^{(0)}(D=1 \mid T=0) = \frac{SP_R - \frac{r_0}{r_0+s_0}}{S_R + SP_R - 1},$$

$$P^{(0)}(D=1 \mid T=1) = \frac{SP_R - \frac{r_1}{r_1+s_1}}{S_R + SP_R - 1}.$$

After some calculations, we show that $\theta^{(1)} = \theta^{(0)}$, $\eta^{(1)} = \eta^{(0)}$, and $p^{(1)} = p^{(0)}$. Thus, the EM algorithm converges after 1 iteration. We have shown that the intuitive estimators defined by (1) are the ML estimators for (θ, η, p).

If there were no verification bias, an unbiased estimator for p, correcting for imperfect reference standard bias is

$$\hat{p}_2 = \frac{\frac{s_1+s_0}{s_1+s_0+r_0+r_1} - (PS_R - 1)}{S_R + SP_R - 1} \qquad (3)$$

If there were no imperfect reference standard bias, the ML estimator for p, correcting for verification bias is [4]

$$\hat{p}_3 = \frac{s_1}{r_1 + s_1}\frac{n_1}{n} + \frac{s_0}{r_0 + s_0}\frac{n_0}{n} \qquad (4)$$

If we ignore both verification bias and imperfect reference bias, the resulting estimate for p is

$$\hat{p}_4 = \frac{s_0 + s_1}{r_0 + r_1 + s_0 + s_1}. \qquad (5)$$

3. A REAL EXAMPLE

In this section, we study effects of both verification bias and imperfect reference standard bias on the estimated prevalence rate in a real example. This example comes from a study of dementia (Hall *et al.* (1996)[2]). This study used a two-stage design. In the first stage, a screening test is used on all subjects in the study sample. Based on the results of the screening test, some of subjects are selected for clinical assessment in the second stage. One of goals in the study is to estimate the prevalence rate of dementia. To illustrate our methods, we used a subset of 75 years old Indianapolis residents. Table 2 summaries the observed data.

Table 2: Observed Data

		$T=1$	$T=0$
V=1	R=1	46	6
	R=0	63	110
V=0		53	624
		162	740

When the sensitivity and specificity of the the imperfect reference standard are known, the estimator given by equation(1) is the ML estimator of p. Thus, we may study the effects of verification and imperfect reference biases on the estimated prevalence by comparing the estimator in (3), derived ignoring verification bias, the estimator in (4), derived ignoring imperfect reference standard bias, and the estimator in (5), derived ignoring both biases with the ML estimator, correcting for both biases.

To see the effects of the specificity and sensitivity of the imperfect reference standard on the estimated prevalence rates, we plot four different estimators of p against the values of sensitivity and specificity of the reference standard, respectively. Figures 1 to 4 summarize the results.

FIGURES 1 AND 4 GO HERE

From Figures 1 to 4, we conclude that verification bias has a bigger effect than imperfect reference bias

Table 3: The observed data with $X = g$

Disorder diagnosis		Screening Test, T			
T		1	2	$\cdots$	K
$V = 1$	$S = 1$	s_{1g}	s_{2g}	$\cdots$	s_{Kg}
	$S = 0$	r_{1g}	r_{2g}	$\cdots$	r_{Kg}
$V = 0$		u_{1g}	u_{2g}	$\cdots$	u_{Kg}
Total		n_{1g}	n_{2g}	$\cdots$	n_{Kg}

on the estimated prevalence rate. If both sensitivity and specificity of the imperfect reference standard are high, we can obtain a reasonable estimator for the prevalence rate by only correcting for verification bias. However, when the sensitivity and specificity of the imperfect reference standard are not high, then both biases have big effects on the estimated the prevalence rate.

4. AN EXTENSION

So far, we have assumed that we know the sensitivity and specificity of the imperfect reference test and that the probability of selection for clinical assessment depends on only the screening test. In this section, we extend our results to the setting whether some other observed discrete covariates X also influence the probability of selection for clinical assessment and that we may not know the sensitivity and specificity of the imperfect reference standard.

Let n_{ig} denote the number of subjects with $T = i$ and $X = g$. Let u_{ig} be the number of subjects with $V = 0$, $T = i$, and $X = g$, s_{ig} be the number of subjects with $V = 1$, $T = i$, $S = 1$, and $X = g$, and r_{ig} be the number of subjects with $V = 1$, $T = i$, $S = 0$, and $X = g$. Table 3 illustrates the layout of the data.

Let us assume that the X takes a value from 1 to G. Since $P(V,T,R,D \mid X) = P(V \mid T,X)P(T \mid D,X)P(R \mid D,X)P(D \mid X)$, the number of unknown parameters is $3GK+G$. However, the degrees of freedom our data can offer is only $G(3K-1)$. Therefore, the likelihood with no constraint on the parameters is over-parameterized. The number of inestimable parameters is $2G$. Thus, to find the ML estimator of p, we need to put the constraints on the remaining parameters. We consider two possibilities.

1. We assume that $S_{Rg} = P(R = 1 \mid D = 1, X = g)$ and $SP_{Rg} = P(R = 0 \mid D = 0, X = g)$ are known (for example, $S_{Rg} = SP_{Rg} = 1$).

2. We assume that $P(T = k \mid D = m, X = g) = \alpha_{km}$ and $K \geq 3$.

Under the constraint (1), the number of parameters is equal to the degrees of freedom. Under the constraint (2), the number of parameters doesn't exceed the degree of freedom.

4.1 ML ESTIMATORS UNDER CONSTRAINT (1)

In this subsection, we derive the ML estimator for the prevalence rate p under the constraint (1). Since

$$\#(R = 1, T = k \mid X = g) = S_{Rg}P(T = k \mid D = 1, X = g)\#(D = 1 \mid X = g)$$
$$+(1 - SP_{Rg})P(T = k \mid D = 0, X = g)\#(D = 0 \mid X = g)$$

and

$$\#(R = 0, T = k \mid X = g) =$$
$$(1 - S_{Rg})P(T = k \mid D = 1, X = g)\#(D = 1 \mid X = g)+$$
$$SP_{Rg}P(T = k \mid D = 0, X = g)\#(D = 0 \mid X = g),$$
$$\#(T = k \mid X = g) = P(T = k \mid D = 1, X = g)\#(D = 1 \mid X = g)$$
$$+P(T = k \mid D = 0, X = g)\#(D = 0 \mid X = g),$$
$$\#(R = 1 \mid X = g) = S_{Rg}\#(D = 1 \mid X = g)+(1-SP_{Rg})\#(D = 0 \mid X = g),$$

and

$$\#(R = 0 \mid X = g) = (1-S_{Rg})\#(D = 1 \mid X = g)+SP_{Rg}\#(D = 0 \mid X = g).$$

After some algebraic manipulation, we get an intuitive estimator for the prevalence rate as

$$\hat{p} = \sum_{g=1}^{G} \frac{n_g}{n} \frac{\sum_{k=1}^{K} \frac{s_{kg}}{s_{kg}+r_{kg}} \frac{n_{kg}}{n_g} - (1 - SP_{Rg})}{SP_{Rg} + S_{Rg} - 1},$$

provided that

$$S_{Rg} \geq max_{1 \leq k \leq K}(\frac{s_{kg}}{s_{kg} + r_{kg}}),$$

$$SP_{Rg} \geq max_{1 \leq k \leq K}(\frac{r_{kg}}{s_{kg} + r_{kg}}).$$

Using the EM algorithm as used in Section 2, we may show that the estimator for p above is also the ML estimator.

4.2 ML ESTIMATORS UNDER CONSTRAINT (2)

Define

$$\alpha_{km} = P(T = k \mid D = m),$$

$$\beta_{lmg} = P(R = l \mid D = m, X = g),$$

and

$$\lambda_{jkg} = P(V = j \mid T = k, X = g).$$

The complete-data log-likelihood is

$$l = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{j,l,m=0,1} \sum_{k=1}^{K} I_{[V_{ig}=j]} I_{[R_{ig}=l]} I_{[D_{ig}=m]} \log \lambda_{jkg} \alpha_{km} \beta_{lmg} p_{mg}.$$

Let $(\alpha^{(t)}, \beta^{(t)}, p^{(t)})$ be the current values of (α, β, p) after t cycles of the EM algorithm. Let $P^{(t)}(.)$ be the conditional probability given the observed data and the current values of $(\alpha, \beta, p) = (\alpha^{(t)}, \beta^{(t)}, p^{(t)})$. After some algebraic manipulation, we obtain the next estimates for (α, β, p):

$$\alpha_{km}^{(t+1)} = \frac{\sum_{g=1}^{G} s_{kg} P_{m11kg}^{(t)} + r_{kg} P_{m10kg}^{(t)} + u_{kg} W_{m1kg}^{(t)}}{\sum_{g=1}^{G} \sum_{k=1}^{K} s_{kg} P_{m11kg}^{(t)} + r_{kg} P_{m10kg}^{(t)} + u_{kg} W_{m1kg}^{(t)}},$$

$$\beta_{0mg}^{(t+1)} = \frac{\sum_{k=1}^{K} (r_{kg} P_{m10kg}^{(t)} + u_{kg} W_{m0kg}^{(t)}}{\sum_{k=1}^{K} s_{kg} P_{m11kg}^{(t)} + r_{kg} P_{m10gk}^{(t)} + u_{kg} W_{m1kg}^{(t)}},$$

$$\beta_{1mg}^{(t+1)} = \frac{\sum_{k=1}^{K} (s_{kg} P_{m11kg}^{(t)} + u_{kg} P_{m10kg}^{(t)})}{\sum_{k=1}^{K} s_{kg} P_{m11kg}^{(t)} + r_{kg} P_{m10kg}^{(t)} + u_{kg} W_{m1kg}^{(t)}},$$

$$p_{mx}^{(t+1)} = \frac{\sum_{k=1}^{K} s_{kg} P_{m11kg}^{(t)} + r_{kg} P_{m10kg}^{(t)} + u_{kg} W_{m1kg}^{(t)}}{n_g},$$

where

$$P_{m10kg}^{(t)} = P^{(t)}(D = m \mid V = 1, R = 0, T = k, X = g),$$

and

$$W_{m1kg}^{(t)} = P^{(t)}(D = m \mid V = 1, T = k, X = g).$$

We iterate this process until the estimates converge. The convergent values $\hat{\alpha}$, $\hat{\beta}$, and $\hat{p}$ are the ML estimates for α, β, and p. Although the output of the EM algorithm does not provide a direct estimate for the asymptotic variance σ of $\hat{\theta}$, four approaches are available to estimate σ. The first approach is to calculate the likelihood function based on the observed data. The second approach is to use the Missing Information Principle:

Observed Information = Complete Information -

Missing Information.

The third approach is to use simulation. When it is difficult to compute the Missing Information, we can use simulation to approximate it. The fourth approach is to use EM iterates.

5. DISCUSSION

In this paper, we have studied the effects of both verification and imperfect reference biases on the estimated prevalence rate. Our example suggests that if the sensitivity and specificity of the imperfect reference standard are high, then we only need to correct for verification bias in estimation of the prevalence rate. However, if the sensitivity and specificity of the imperfect reference standard are not high, then we need to correct for both biases. Under some assumptions, we derive the ML estimator for the prevalence rate, correcting for both biases. Since it is much more complicated and involves more assumptions to correct for both biases than to correct for only verification bias, we recommend to concentrate on correcting for verification bias in a two-stage design study when the accuracy of the imperfect reference standard is reasonable high. Only when the assumption of the high accuracy of the imperfect reference standard is questionable, we recommend to use the more complicated procedures to estimate the prevalence rate.

References

[1] B. P. Dohrenwend and P. E. Shrout. Toward the development of a two-stage procedure for case identification and classification in psychiatric epidemiology. *Research in Community and Mental Health (ed. R. Simmons)*, 2:295–323, 1981.

[2] K.S. Hall, H.C. Hendrie, B.O. Osuntokun, A. Ogunniyi, and S.L. Hui. Community screening for dementia in Indianapolis and Ibadan, Nigeria. *The Proceedings of the Scientific Session at Indiana University School of Medicine*, abstract, 1994.

[3] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, New York, NY, 1987.

[4] X.H. Zhou. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Comm in Stat - Theory Meth*, 22:3177–3198, 1993.

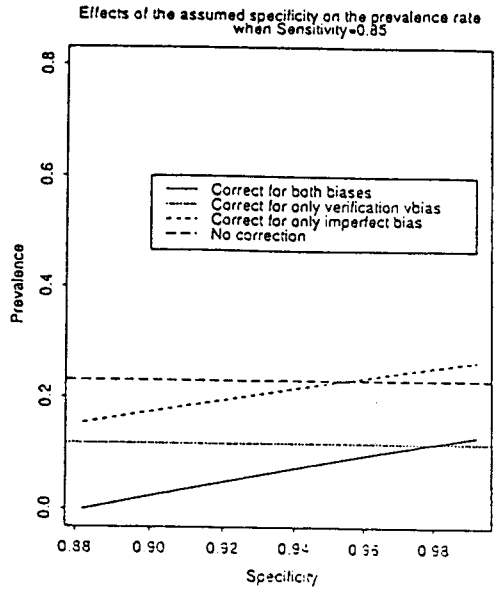

Figure 2

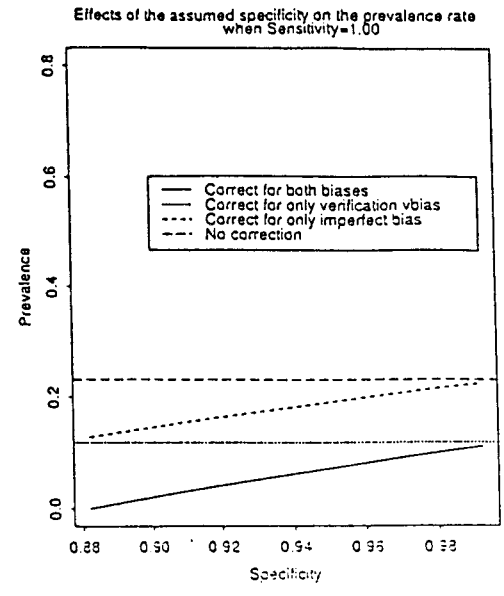

Figure 1

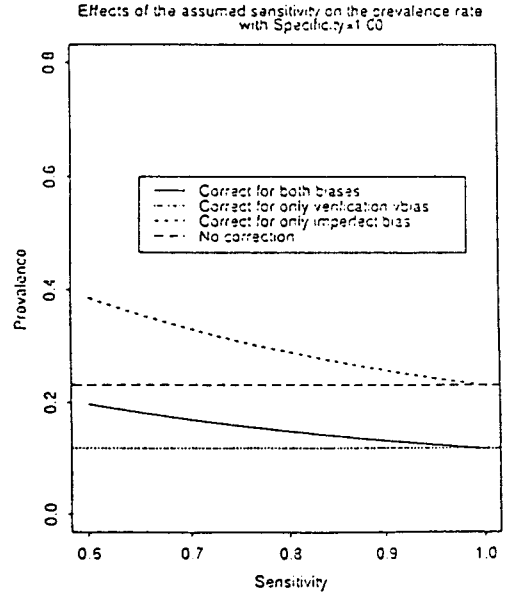

Figure 3

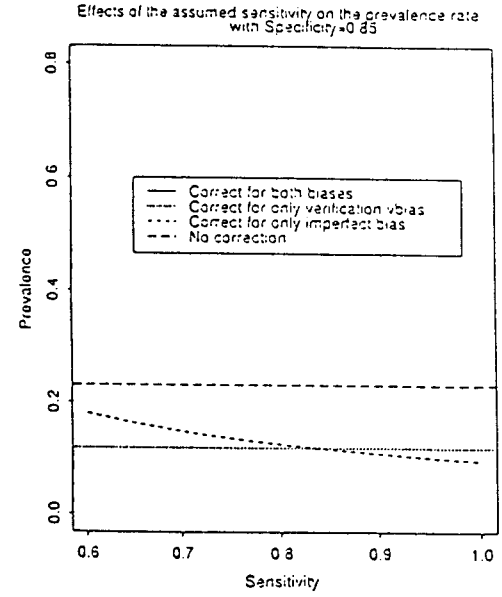

Figure 4