Measurement, Design, and Analytic Techniques in Mental Health and Behavioral Sciences

Lecture 16: Correction of Verification Bias
May 21, 2009

XH Andrew Zhou

azhou@u.washington.edu

Department of Biostatistics, University of Washington

Verification bias

- To estimate sensitivity/specificity, predictive values, and ROC curves, we assume that we know the disease status of each patient under the study.
- In clinical practice, however, some of the patients with test results may not have verified disease status. For example, if disease verification is based on invasive surgery, then patients with negative test results are less likely to receive the disease verification than patients with positive test results.
- Although this approach may be sensible and cost-effective in clinical studies, when it occurs in studies designed to evaluate the accuracy of diagnostic tests, the estimated accuracy of the tests may be biased. This type of bias is called verification bias.

An example

- Let us consider a real data about hepatic scintigraphy for liver disease.
- Hepatic scintigraphy is an imaging scan procedure to detect liver cancer.
- In this study some of the patients were refereed to disease verification process–liver pathology–which was considered as a golden standard.

Hepatic scintigraphy data

		$T=1$	$T=0$
$V=1$	$D=1$	231	27
	$D=0$	32	54
$V=0$		166	140
Total		429	221

The second real example

- Marshall et al introduced the Diaphanography as a test for detecting breast cancer.
- Diaphanography (lightscanning) is a noninvasive method of examining the breast by transillumination using visible or infrared light.
- Gold standard, needle biopsy.
- Data:

		$T=1$	$T=0$
$V=1$	$D=1$	26	7
	$D=0$	11	44
$V=0$		30	782
Total		67	833

Effects of Verification Bias

- To learn how verification bias works, let us look at a hypothetical example with 200 patients.
- Let us suppose that the decision to verify disease status by a gold standard depends on the result of the binary test under the study:
  - a patient with a positive result has a 1/2 chance of receiving the verification procedure, and
  - a patient with a negative result has only 1/5 chance of receiving verification.
- We want to estimate sensitivity and specificity of the test.

The Target Population

- The following Table contains the result that might have been obtained if every patient had been verified.

	T=1	T=0	Total
D=1	80	20	100
D=0	10	90	100
Total	90	110	200

- Sens=0.80 and spec=0.90

The target population, cont

- The following table contains the data with partial verification:

	T=1	T=0	Total
D=1	40	4	44
D=0	5	18	23
unverified	45	88	133
Total	90	110	200

- Sens=0.91 and spec=0.78.
  True sensitivity was overestimated and specificity was underestimated.

small Verification bias in published clinical studies

- Greenes and Begg (1985, Investigative Radiology) reviewed 145 studies published between 1976 and 1980 and found that at least 26% of the articles had verification bias, but failed to recognize it.
- Bates (1993, Journal of Pediatrics) reviewed 54 pediatric studies and found more than one third had verification bias.
- Philbrick (1980,American Journal of Cardiology) reviewed 33 studies on the accuracy of exercise tests for coronary disease and found that 31 might have had verification bias.

How to correct for verification bias

- A patient without disease verification = missing the value of the true disease status.
- The problem of verification bias is a special type of missing-data problems.
- Maximum likelihood (ML) methods for data with missing values; the expectation-maximization (EM) algorithm, a general approach to the iterative computation of ML estimates in a variety of missing-data problems. Little and Rubin (1987) and McLachlan and Krishnan (1997).

Single Test

- T: a binary random variable, indicating whether or not the test was positive $(T = 1)$ or negative $(T = 0)$.
- V: a random variable indicating whether or not the subject was verified using the golden standard procedure ($V = 1$ if verified, $V = 0$ if not).
- D: the true status for those who were verified using the golden standard, such that $D = 1$ if diseased and $D = 0$ if non-diseased (we assume there is no measurement error for the golden standard procedure).

Single Test, Cont

Table 1: Data summary

a. aggregated data

		$T=1$	$T=0$
$V=1$	$D=1$	x^A_{11}	x^A_{10}
	$D=0$	x^A_{01}	x^A_{00}
$V=0$		x^B_{+1}	x^B_{+0}
Total		n_1	n_0

b. complete data

	$T=1$	$T=0$
$D=1$	x_{11}	x_{10}
$D=0$	x_{01}	x_{00}
Total	n_1	n_0

Here

$x^A_{ij} = nP(V=1, D=i, T=j), x^B_{ij} = nP(V=0, D=i, T=j),$ and $x_{ij} = x^A_{ij} + x^B_{ij},$

where $n = n_0 + n_1$, $i, j = 0, 1$.

Conditional independence assumption

- $$P(V = 1 \mid D, T) = P(V = 1 \mid T).$$
- Parameters of interest:

$$Se = P(T = 1 \mid D = 1), Sp = P(T = 0 \mid D = 0).$$

Existing moment estimators

- Note that $Se = \frac{\#(T=1,D=1)}{\#(D=1)}$.
- Since $P(V = 1 \mid D = 1, T = 1) = P(V = 1 \mid T = 1)$ due to the conditional independence assumption,

$$\frac{\#(V = 1, T = 1, D = 1)}{\#(D = 1, T = 1)} = \frac{\#(V = 1, T = 1)}{\#(T = 1)}$$

- Hence

$$\#(D = 1, T = 1) = \#(V = 1, T = 1, D = 1)\frac{\#(T = 1)}{\#(V = 1, T = 1)} = x_{11}^A \frac{n_1}{x_{11}^A + x_{01}^A}.$$

- Similarly, since $P(V = 1 \mid D = 1, T = 0) = P(V = 1 \mid T = 0)$, we obtain that

$$\#(D = 1, T = 0) = \#(V = 1, T = 0, D = 1)\frac{\#(T = 0)}{\#(V = 1, T = 0)} = x_{10}^A \frac{n_0}{x_{10}^A + x_{00}^A}.$$

- We obtain that $\#(D = 1) = x_{11}^A \frac{n_1}{x_{11}^A + x_{01}^A} + x_{10}^A \frac{n_0}{x_{10}^A + x_{00}^A}$.

Moment estimators, cont

- Moment estimator of Se is given as follows:

$$\widehat{Se} = \frac{(x_{11}^A n_1)/(x_{11}^A + x_{01}^A)}{(x_{11}^A n_1)/(x_{11}^A + x_{01}^A) + (x_{10}^A n_0)/(x_{10}^A + x_{00}^A)}.$$

- Similarly we can show that the moment estimator of Sp is given as follows:

$$\widehat{Sp} = \frac{(x_{00}^A n_0)/(x_{10}^A + x_{00}^A)}{(x_{01}^A n_1)/(x_{11}^A + x_{01}^A) + (x_{00}^A n_0)/(x_{10}^A + x_{00}^A)}.$$

- These estimators are denoted by B&G estimators (Begg and Greenes, 1987).
- It can be shown these moment estimates are also ML estimates (Zhou, 1994).

Confidence intervals

- Variance estimate of $\widehat{Se}$:

$$\hat{var}(\widehat{Se}) = (\widehat{Se}(1-\widehat{Se}))^2[\frac{n}{n_1 n_0} + \frac{x_{01}^A}{x_{11}^A(x_{11}^A + x_{01}^A)} + \frac{x_{00}^A}{x_{10}^A(x_{10}^A + x_{00}^A)}].$$

- Variance estimate of $\widehat{Sp}$:

$$\hat{var}(\widehat{Sp}) = (\widehat{Sp}(1-\widehat{Sp}))^2[\frac{n}{n_1 n_0} + \frac{x_{11}^A}{x_{01}^A(x_{11}^A + x_{01}^A)} + \frac{x_{10}^A}{x_{00}^A(x_{10}^A + x_{00}^A)}].$$

- The $100(1-\alpha)\%$ confidence intervals for sensitivity and specificity will be

$$\widehat{Se} \pm \kappa\sqrt{\hat{var}(\widehat{Se})},$$

$$\widehat{Sp} \pm \kappa\sqrt{\hat{var}(\widehat{Sp})},$$

  respectively, where κ is the $(1-\alpha/2)$ percentile of the standard normal distribution.

Confidence intervals, cont

- Instead of assuming normality for $(\widehat{Se} - \pi)$, one may think that the $logit$ transformation of $\widehat{Se}$ is closer to a normal approximation, such that $logit(\widehat{Se}) - logit(Se) \sim N(0, \hat{Var}(logit(\widehat{Se})))$. Using this logit transformation, the $100(1-\alpha)\%$ confidence interval for sensitivity and specificity will be

$$logit^{-1}\left(logit(\widehat{Se}) \pm \kappa\sqrt{\hat{Var}(logit(\widehat{Se}))}\right),$$

$$logit^{-1}\left(logit(\widehat{Sp}) \pm \kappa\sqrt{\hat{Var}(logit(\widehat{Sp}))}\right),$$

respectively, where,

$$\hat{Var}(logit(\widehat{Se})) = \frac{n}{n_1 n_0} + \frac{x_{01}^A}{x_{11}^A(x_{11}^A + x_{01}^A)} + \frac{x_{00}^A}{x_{10}^A(x_{10}^A + x_{00}^A)},$$

and

$$\hat{Var}(logit(\widehat{Sp})) = \frac{n}{n_1 n_0} + \frac{x_{11}^A}{x_{01}^A(x_{11}^A + x_{01}^A)} + \frac{x_{10}^A}{x_{00}^A(x_{10}^A + x_{00}^A)}.$$

Correction Methods Without the MAR Assumption

- The validity of the above methods depends on the MAR assumption for the verification mechanism.
- However, if the verification process depends on unobserved variables that are related to the condition status, the verification process is not MAR.
- This most likely occurs when there is a long time lag between the initial test and verification, when there are multiple investigators at various institutions, when the patient population is very heterogeneous, or when the disease process is not well understood.
- Without the MAR assumption, we need to model the verification process to make inferences about the test's sensitivity and specificity.

Correction methods

- Let λ_{11} be the conditional probability of the selection of a patient for verification given that the patient has a positive test result and has the condition,
- λ_{01} be the conditional probability of the selection of a patient for verification given a positive test result and the absence of the condition,
- λ_{10} be the conditional probability of the selection of a patient for verification given a negative test result and the presence of the condition,
- and λ_{00} be the conditional probability of the selection for verification of a patient given a negative test result and the absence of the condition. Denote

$$\phi_{1t} = P(T = t) \text{ and } \phi_{2t} = P(D = 1 \mid T = t),$$

where $t, \tilde{t} = 0, 1$. Let $\phi_1 = \phi_{11}$ and $\phi_2 = (\phi_{20}, \phi_{21})'$.

Likelihood function

- Based on the observed data given, we may write the log-likelihood function as

$$l = \sum_{j=0}^{1} n_j \log \phi_{1j} + \sum_{j=0}^{1} x_{1j}^{A} \log(\lambda_{1j}\phi_{2j}) + x_{0j}^{B} \log(\lambda_{0j}(1-\phi_{2j}))+$$

$$x_{+j}^{B} \log((1-\lambda_{1j})\phi_{2t} + (1-\lambda_{0j})(1-\phi_{2j})).$$

- Set $e_j = \lambda_{1j}/\lambda_{0j}$. Then, the log-likelihood becomes

$$l = \sum_{j=0}^{1} n_j \log \phi_{1j} + \sum_{j=0}^{1} x_{1j}^{A} \log(e_j\lambda_{0j}\phi_{2j}) + X_{0j}^{A} \log(\lambda_{0j}(1-\phi_{2j}))+$$

$$x_{+j}^{B} \log((1-e_j\lambda_{0j})\phi_{2j} + (1-\lambda_{0j})(1-\phi_{2j})).$$

Likelihood function, cont

- Since the degrees of freedom in the data is 5, not all *7* parameters $\phi_{11}, \phi_{20}, \phi_{21}, \lambda_{11}, \lambda_{12}, e_0,$ and e_1 are estimable.
- If we can assume that two of them are known, the remaining *5* parameters may be estimable.
- Under the assumption that e_0 and e_1 were known, we can show that the resulting ML estimators for sensitivity and specificity are as follows: that the resulting ML estimators for sensitivity and specificity are as follows:

$$\widehat{Se}(e_0, e_1) = \frac{(s_1 m_1)/(s_1 + e_1 r_1)}{(s_1 m_1)/(s_1 + e_1 r_1) + (s_0 m_0)/(s_0 + e_0 r_0)} \tag{1}$$

and

$$\widehat{Sp}(e_0, e_1) = \frac{(e_0 r_0 m_0)/(s_0 + e_0 r_0)}{(e_1 r_1 m_1)/(s_1 + e_1 r_1) + (e_0 r_0 m_0)/(s_0 + e_0 r_0)}, \tag{2}$$

respectively.

Application to the second real example

- Marshall et al introduced the Diaphanography as a test for detecting breast cancer.
- Diaphanography (lightscanning) is a noninvasive method of examining the breast by transillumination using visible or infrared light.
- Gold standard, needle biopsy.
- Data:

		$T=1$	$T=0$
$V=1$	$D=1$	26	7
	$D=0$	11	44
$V=0$		30	782
Total		67	833

The second real example, cont

- Recall that e_0 and e_1 are ratios of two conditional probabilities:
  $e_0 = P(V = 1 \mid T = 1, D = 1)/P(V = 1 \mid T = 1, D = 0)$ and
  $e_1 = P(V = 1 \mid T = 0, D = 1)/P(V = 1 \mid T = 0, D = 0)$.
- Using the observed data, Zhou (1993) showed the ranges of possible values of e_0 and e_1 are as follows.

$$\frac{s_1}{s_1 + u_1} \leq e_1 \leq \frac{r_1 + u_1}{r_1}, \frac{s_0}{s_0 + u_0} \leq e_0 \leq \frac{r_0 + u_0}{r_0}. \qquad (3)$$

- Using these bounds, we can study how sensitive the ML estimators of sensitivity and specificity derived under the MAR assumption are to the departure from the MAR assumption.

Application to Hepatic Scintigraph Example

- Under the MAR, the estimated sensitivity is 0.84 with a 95% confidence interval of (0.79,0.88), and the estimated specificity is 0.74 with a 95% confidence interval of (0.66,0.81).
- Without the MAR assumption, we need to assume two ratios e_1 and e_0 are known to derive the ML estimators for sensitivity and specificity.
- Here, e_1 is the ratio of the probability of verifying a patient who has a positive hepatic scintigraph result and liver disease to that of verifying a patient who has a positive hepatic scintigraph result but does not have liver disease,
- and e_0 is the ratio of the probability of verifying a patient who has a negative hepatic scintigraph result and liver disease to that of verifying a patient who has a negative hepatic scintigraph result and does not have liver disease.

Hepatic Scintigraph Example, Cont

- For given values of e_1 and e_0, ML estimators for sensitivity and specificity are

$$\widehat{Se}(e_0, e_1) = \frac{1}{1 + 0.06(32e_1 + 231)/(54e_0 + 27)}$$

and

$$\widehat{Sp}(e_0, e_1) = \frac{1}{1 + 1.15(e_1(54e_0 + 27))/(e_0(32e_1 + 231))}$$

respectively.

- Using Formula (**??**), we obtain lower and upper bounds for e_1 and e_0,

$$0.57 \leq e_1 \leq 1.72,\ 0.16 \leq e_0 \leq 6.2.$$

Hepatic Scintigraph Example, Cont

- From these bounds, we can derive lower and upper bounds for the estimated sensitivity and specificity.
- Note that for a given e_0 both $\widehat{Se}(e_1, e_0)$ and $\widehat{Sp}(e_1, e_0)$ are decreasing functions of e_1, and for a given e_1 both $\widehat{Se}(e_1, e_0)$ and $\widehat{Sp}(e_1, e_0)$ are increasing functions of e_0.
- Thus,

$$0.68 \leq \frac{1}{1 + 17.16/(54e_0 + 27)} \leq \widehat{Se}(e_1, e_0) \leq \frac{1}{1 + 14.95/(54e_0 + 27)} \leq 0.95$$

and

$$0.37 \leq \frac{1}{1 + (54e_0 + 27)/216.73} \leq \widehat{Sp}(e_1, e_0) \leq \frac{1}{1 + (54e_0 + 27)/248.73} \leq 0.86.$$

- Therefore, the ML estimators for sensitivity and specificity could vary from 0.68 to 0.95 and 0.37 to 0.86 respectively, depending on the values of e_0 and e_1.

A single ordinal scale test

- Let T be the ordinal scale test result; the definitions of random variables D and V are the same as before.
- We can then summarize the observed data in Table below.

		Diagnostic test results		
		T=1	...	T=K
Verified	D=1	s_1	...	s_K
	D=0	s_1	...	s_K
Unverified		u_1	...	u_K
Total		m_1	...	m_K

- We assume that the probability of verifying a patient depends only on the test result T; that is,

$$P(V = 1 \mid T, D) = P(V = 1 \mid T). \tag{4}$$

Estimation of ROC Curves

- For an ordinal scale test, by varying the definition of a positive test, we can calculate K+1 pairs of true positive rates (TPR) and false positive rates (FPR) of the test.
- Specifically, if we define a positive test as the one with $T \geq t$, a corresponding pair of TPR and FPR are

$$TPR(t) = P(T \geq t \mid D = 1),\ FPR(t) = P(T \geq t \mid D = 0),$$

  respectively, for $t = 1, \ldots, K + 1$. Using the trapezoidal rule (Bamber, 1975), we produce an empirical ROC curve by connecting the coordinates, $(FPR(t), TPR(t))$.
- Since $TPR(1) = FPR(1) = 1$ and $TPR(K + 1) = FPR(K + 1) = 0$, to provide a unbiased estimator of an empirical ROC curve we need to find unbiased estimators for $(FPR(t), TPR(t))$, $t = 2, \ldots, K$.

Further notation

- Define $\phi_{1t} = P(T = t)$ and $\phi_{2t} = P(D = 1 \mid T = t)$, where $t = 1, \ldots, K$. Then, $\phi_{1K} = 1 - \phi_{11} - \ldots - \phi_{1(K-1)}$.
- Denote $\phi_1 = (\phi_{11}, \ldots, \phi_{1(K-1)})$ and $\phi_2 = (\phi_{21}, \ldots, \phi_{2K})$.
- Under the assumption that the verification mechanism is MAR,valid likelihood-based inferences on ϕ_{1t} and ϕ_{2t} can be made based on observed data without specifying a distribution for the verification mechanism.

Likelihood function

- The log-likelihood function based on the observed data is

$$l(\phi_1, \phi_2) = \sum_{t=1}^{K} m_t log(\phi_{1t}) + \sum_{t=1}^{K} (s_t log(\phi_{2t}) + r_t log(1 - \phi_{2t})). \quad (5)$$

- Let $l_1(\phi_1) = \sum_{t=1}^{K} m_t log(\phi_{1t})$ and $l_2(\phi_2) = \sum_{t=1}^{K} (s_t log(\phi_{2t}) + r_t log(1 - \phi_{2t}))$.
- We can write $l(\phi_1, \phi_2)$ as the sum of $l_1(\phi_1)$ and $l_2(\phi_2)$.
- Since ϕ_1 and ϕ_2 are distinct parameters and both l_1 and l_2 are the log-likelihood functions for multinomial distributions, the ML estimators for ϕ_1 and ϕ_2 are

$$\hat{\phi}_{1t} = \frac{m_t}{N}, t = 1, \ldots, K-1, \hat{\phi}_{2t} = \frac{s_t}{s_t + r_t}, t = 1, \ldots, K. \quad (6)$$

- The corresponding observed Fisher information matrix on (ϕ_1, ϕ_2) is

$$\text{diag}(I_1(\phi_1), I_2(\phi_2)) \quad (7)$$

where $I_1(\phi_1)$ and $I_2(\phi_2)$ are the observed Fisher information matrices on the log-likelihood $l_1(\phi_1)$ and $l_1(\phi_2)$, respectively.

ML estimates for ROC curves

- Next, we derive the ML estimators for the empirical ROC curve.
- Note that the coordinates of the empirical ROC curve can be written as functions of ϕ_1 and ϕ_2,

$$TPR(t) = \frac{\sum_{\tilde{t}=t}^{K} \phi_{1\tilde{t}}\phi_{2\tilde{t}}}{\sum_{\tilde{t}=1}^{K} \phi_{1\tilde{t}}\phi_{2\tilde{t}}} \text{ and } FPR(t) = \frac{\sum_{\tilde{t}=t}^{K} \phi_{1\tilde{t}}(1-\phi_{2\tilde{t}})}{\sum_{\tilde{t}=1}^{K} \phi_{1\tilde{t}}(1-\phi_{2\tilde{t}})}.$$

- For $2 \leq t \leq K$ the ML estimators of $TPR(t)$ and $FPR(t)$ are defined as follows:

$$\widehat{TPR}(t) = \frac{\sum_{\tilde{t}=t}^{K} \frac{m_{\tilde{t}}}{N} \frac{s_{\tilde{t}}}{s_{\tilde{t}}+r_{\tilde{t}}}}{\sum_{\tilde{t}=1}^{K} \frac{m_{\tilde{t}}}{N} \frac{s_{\tilde{t}}}{s_{\tilde{t}}+r_{\tilde{t}}}} \text{ and } \widehat{FPR}(t) = \frac{\sum_{\tilde{t}=t}^{K} \frac{m_{\tilde{t}}}{N} \frac{r_{\tilde{t}}}{s_{\tilde{t}}+r_{\tilde{t}}}}{\sum_{\tilde{t}=1}^{K} \frac{m_{\tilde{t}}}{N} \frac{r_{\tilde{t}}}{s_{\tilde{t}}+r_{\tilde{t}}}}. \quad (8)$$

Estimation of the ROC curve area

- We first observe that the area under the empirical ROC curve is a function of the parameters ϕ_1 and ϕ_2,

$$A = \frac{\sum_{t=1}^{K-1} \sum_{\tilde{t}=t+1}^{K} (1-\phi_{2t})\phi_{1t}\phi_{2\tilde{t}}\phi_{1\tilde{t}} + (1/2)\sum_{t=1}^{K}(1-\phi_{2t})\phi_{2t}\phi_{1t}^2}{\sum_{t=1}^{K}(1-\phi_{2t})\phi_{1k}\sum_{\tilde{t}=1}^{K}\phi_{2\tilde{t}}\phi_{1\tilde{t}}}, \quad (9)$$

- the ML estimator for the area under the ROC curve is

$$\hat{A} = \frac{\sum_{t=1}^{K-1} \sum_{\tilde{t}=t+1}^{K} \frac{r_t m_t}{s_t+r_t}\frac{s_{\tilde{t}} m_{\tilde{t}}}{s_{\tilde{t}}+r_{\tilde{t}}} + (1/2)\sum_{t=1}^{K}\frac{s_t r_t m_t^2}{(s_t+r_t)^2}}{\sum_{t=1}^{K}\frac{r_t m_t}{s_t+r_t}\sum_{\tilde{t}=1}^{K}\frac{s_{\tilde{t}} m_{\tilde{t}}}{s_{\tilde{t}}+r_{\tilde{t}}}}. \quad (10)$$

Fever of Uncertain Origin Example

- Gray et al (1984) reported data from a study on the accuracy of computed tomography in differentiating focal from nonfocal sources of sepsis among patients with fever of uncertain origin.
- In this study only some patients were verified, depending on their CT results. Hence, this study had verification bias. Table below displays the data.

		T=1	T=2	T=3	T=4	T=5
V=1	D=1	7	7	2	3	37
	D=0	8	0	1	1	4
V=0		40	11	3	5	12
Total		55	18	6	9	53

Fever of Uncertain Origin Example, cont

- We obtained the nonparametric ML estimate for the ROC area as 0.75. The corresponding SD estimate was 0.066 using the information method.