A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error Without a Gold Standard

Paul S. Albert and Lori E. Dodd
Biometric Research Branch
National Cancer Institute

FDA/Industry Statistics Workshop
September 2003

Objective

Estimating diagnostic error (sensitivity and specificity) without a gold standard from repeat tests on a given patient.

Examples

- Handelman's (1986) dentistry dataset where 5 dentists evaluated dental x-rays from 3,869 incipient carries.
- Alvord's (1988) HIV dataset where 428 subjects were tested by 4 conventional bioassays.
- Holmquist's (1967) uterine cancer dataset where 7 pathologists independently evaluated 118 histological slides from biopsies of the uterine cervix.

Handleman's Dentistry Dataset

Test result	Obs. freq.	Test result	Obs. freq.
00000	1880	10000	22
00001	789	10001	26
00010	43	10010	6
00011	75	10011	14
00100	23	10100	1
00101	63	10101	20
00110	8	10110	2
00111	22	10111	17
01000	188	11000	2
01001	191	11001	20
01010	17	11010	6
01011	67	11011	27
01100	15	11100	3
01101	85	11101	72
01110	8	11110	1
01111	56	11111	100

Latent Class Modeling Approaches

- Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iJ})'$.

- Let d_i be the true binary disease status.

$$P(Y_{i1}, Y_{i2}, ..., Y_{iJ}) = \sum_{l=0}^{1} P(Y_{i1}, Y_{i2}, ..., Y_{iJ}|d_i = l)P(d_i = l)$$

- Different models for $P(Y_{i1}, Y_{i2}, ..., Y_{iJ}|d_i = l)$

Conditional Independence (Hui and Walters, 1980)

- $Y_{ij}|d_i \sim$ Bernoulli with probability $P(Y_{ij} = 1|d_i)$
- $P(Y_{i1}, Y_{i2}, ..., Y_{iJ}|d_i = l) = \prod_{j=1}^{J} P(Y_{ij}|d_i = l)$
- Sensitivity=$P(Y_{ij} = 1|d_i = 1)$
- Specificity=$P(Y_{ij} = 0|d_i = 0)$

Gaussian Random Effects Model (Qu et al., 1996)

- $Y_{ij}|d_i, b \sim$ Bernoulli with probability $\Phi(\beta_{d_i} + \sigma_{d_i} b)$ where $b \sim \ N(0,1)$ is a subject-specific random effect

- $P(Y_{i1}, Y_{i2}, ..., Y_{iJ}|d_i = l) = \int \{\prod_{j=1}^{J} P(Y_{ij}|d_i, b)\}\phi(b)db$ where the integral can be evaluated using Gaussian Quadrature

- Sensitivity$=P(Y_{ij} = 1|d_i = 1) = E_b\{P(Y_{ij} = 1|d_i = 1)\}$
  $= \Phi(\beta_1/\sqrt{1+\sigma_1^2})$

- Specificity$=P(Y_{ij} = 0|d_i = 0) = E_b\{P(Y_{ij} = 0|d_i = 0)\}$
  $= 1 - \Phi(\beta_0/\sqrt{1+\sigma_0^2})$

Beta-Binomial Model

- $\sum_j Y_{ij} = a | d_i = 0 \sim$ Beta-binomial(α_0, β_0)

  and $\sum_j Y_{ij} = a | d_i = 1 \sim$Beta-binomial$(\alpha_1, \beta_1)$

- $P(Y_{i1}, Y_{i2}, ..., Y_{iJ} | d_i = l) = P(\sum_j Y_{ij} = a | d_i = l) / \binom{J}{a}$

- Sensitivity=$\alpha_1/(\alpha_1 + \beta_1)$

- Specificity= $1/(\alpha_0 + \beta_0)$

Finite Mixture Model

- $P_0 = P(\text{Not Subject to Misclassification}|d_i = 0)$
- $P_1 = P(\text{Not Subject to Misclassification}|d_i = 1)$
- $P(Y_{i1}, ..., Y_{iJ}|d_i = 1) =$
$$\begin{cases} P_1 + (1 - P_1) \prod_j P_{d_i=1}(Y_{ij} = 1) & \text{Tests all ones} \\ (1 - P_1) \prod_j P_{d_i=1}(Y_{ij}) & \text{Test not all ones} \end{cases}$$
where $P_{d_i}(Y_{ij} = 1)$ is the probability of $Y_{ij} = 1$ given the patient is subject to misclassification and d_i is the true binary disease status.
- Sensitivity=$P_1 + (1 - P_1)P_{d_i=1}(Y_{ij} = 1)$
- Specificity=$P_0 + (1 - P_0)P_{d_i=0}(Y_{ij} = 0)$

Identifiability and Estimation

- Conditional independence model identifiable when $J \geq 3$.
- Gaussian random effects, Beta-binomial, and Finite Mixture identifiable when $J \geq 5$.
- Maximum-likelihood Estimation
- Bootstrap for standard errors of sensitivity and specificity estimates.

Comparison of Methods on Handelman's Dentistry Data

		Expected Frequency			
Pos. Tests	Freq.	*Indep*	*FM*	*BB*	*GRE*
0	1880	1821.5	1879.5	1882.5	1880.4
1	1065	1132.9	1065.1	1058.8	1062.8
2	404	376.2	404.2	411.4	408.8
3	247	244.5	247.2	239.4	242.3
4	173	211.2	172.9	178.0	176.5
5	100	82.7	100.0	98.9	99.2
Total	3869				
$\widehat{SENS}$		0.658	0.645	0.518	0.457
		(0.017)	(0.026)	(0.076)	(0.088)
$\widehat{SPEC}$		0.894	0.895	0.904	0.912
		(0.004)	(0.006)	(0.006)	(0.010)
logL		-8726.5	-8717.1	-8717.8	-8717.8
χ^2		20.773	1.293	2.317	1.979
df		3	1	1	1

Estimation of rater-specific sensitivity and specificity

		Indep	*FM*	*GRE*
Rater		Est.(SE[1])	Est.(SE)	Est. (SE)
1	Sens	0.40(0.026)	0.45(0.038)	0.54(0.120)
	Spec	0.99(0.002)	0.99(0.003)	0.97(0.013)
2	Sens	0.71(0.025)	0.74(0.034)	0.77(0.100)
	Spec	0.89(0.007)	0.88(0.008)	0.85(0.026)
3	Sens	0.60(0.028)	0.66(0.040)	0.81(0.190)
	Spec	0.99(0.003)	0.98(0.005)	0.96(0.021)
4	Sens	0.49(0.022)	0.51(0.026)	0.50(0.060)
	Spec	0.97(0.005)	0.96(0.007)	0.93(0.022)
5	Sens	0.92(0.014)	0.92(0.018)	0.93(0.070)
	Spec	0.69(0.011)	0.67(0.012)	0.64(0.032)
logLik		-7427.0	-7421.8	-7465.4

[1] standard errors were estimated using a bootstap with 1000 bootstrap samples.

MLEs of Diagnostic Error: Asymptotic Bias

- The mispecified MLE denoted by $\widehat{\boldsymbol{\theta}}^*$ converges to the value $\boldsymbol{\theta}^*$, where

$$\boldsymbol{\theta}^* = \max_{\boldsymbol{\theta}} E_T[\text{logL}(\boldsymbol{Y}_i, \boldsymbol{\theta})]$$

- $E_T(logL_M) = E_T[logL(\boldsymbol{Y}_i, \boldsymbol{\theta})]|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$

- $SENS^* = g_1(\boldsymbol{\theta}^*)$ and $SPEC^* = g_2(\boldsymbol{\theta}^*)$.

- Estimates of sensitivity and specificity converge to $SENS^*$ and $SPEC^*$ under mis-specified models.

Asymptotic Results

Large sample robustness of the assumed latent class beta-binomial (BB) model to the true dependence structure between tests. The true model is a finite mixture model (FM) with $P_0 = P_1 = 0.2$, SENS=0.75 and SPEC=0.9 for differing P_d.

P_d	J	$SENS^*$	$SPEC^*$	$E_T[\log L_{FM}]$	$E_T[\log L_{BB}]$
0.05	5	0.78	0.90	-1.82684	-1.82684
	6	0.64	0.90	-2.17092	-2.171269
	10	0.68	0.90	-3.52125	-3.52758
0.1	5	0.55	0.90	-1.98481	-1.98481
	6	0.53	0.90	-2.34536	-2.34586
	10	0.66	0.90	-3.74875	-3.75775

Asymptotic Results (continued)

Large sample robustness of the Gaussian random effects (GRE) assumption for four tests with different diagnostic errors when the true model is a finite mixture model with $P_0 = 0.5$, $P_1 = 0.5$, and $P_d = 0.2$.

	Diagnostic Error			
	True Model		Miss-specified Model.	
Test	$SENS$	$SPEC$	$SENS^*$	$SPEC^*$
1	0.80	0.95	0.73	0.95
2	0.85	0.95	0.78	0.95
3	0.90	0.95	0.83	0.95
4	0.95	0.95	0.89	0.96

$E_{FM}[\log L_{FM}] = E_{FM}[\log L_{GRE}] = -1.35814.$

Simulation Results

Simulated under the finite mixture model with $P_d = 0.2$, $P_0 = P_1 = 0.2$, SENS=0.75, and SPEC=0.90.

		Avg. Est. GRE		Reject χ^2		
I	J	SENS	SPEC	*Indep*	*FM*	*GRE*
250	5	0.62	0.89	16.0	0.8	3.2
		(0.17)	(0.03)			
250	10	0.56	0.90	79.4	2.3	35.4
		(0.17)	(0.02)			
1000	5	0.62	0.89	88.7	0.1	4.9
		(0.17)	(0.03)			
1000	10	0.54	0.90	100	2.3	98.7
		(0.16)	(0.01)			

Simulation Results (Continued)

Simulated under Gaussian random effects model with $P_d = 0.2$, $\sigma_0 = \sigma_1 = 1.5$, SENS=0.75, and SPEC=0.90.

		Avg. Est. FM		Reject χ^2		
I	J	SENS	SPEC	*Indep*	*FM*	*GRE*
250	5	0.84	0.94	95.1	0.2	4.6
		(0.07)	(0.02)			
250	10	0.83	0.94	100	50.2	3.6
		(0.04)	(0.01)			
1000	5	0.84	0.94	100	0	3.6
		(0.07)	(0.02)			
1000	10	0.83	0.94	100	99.7	3.5
		(0.02)	(0.01)			

Simulation Results (continued)

Simulation with four tests in which test-specific sensitivity and specificity were estimated. Data were simulated under the finite mixture model (FM) with $P_0 = P_1 = P_d = 0.5$, and $I = 1000$.

			Avg. Est.	
Test		Truth	FM	GRE
1	SENS	0.80	0.80	0.64
	SPEC	0.95	0.95	0.79
2	SENS	0.85	0.85	0.72
	SPEC	0.90	0.90	0.72
3	SENS	0.90	0.90	0.77
	SPEC	0.85	0.85	0.68
4	SENS	0.95	0.95	0.79
	SPEC	0.80	0.80	0.64

Concluding Remarks

- Sensitivity and Specificity are asymptotically biased when dependence structure is mis-specified.
- $E_T(logL_M) \approx E_T(logL_T)$ for small number of tests.
- Example and simulations demonstrate that it is difficult to distinguish between models for the dependence structure with few numbers of tests.
- Problem remains for rater-specific sensitivity and specificity. In addition, ranking is not often preserved under a mispecified model.
- Recommendations:
  1. collect gold standard information (even on a subset of data) whenever possible
  2. Perform sensitivity analysis
  3. Perform as many tests as possible
- Future Research: Gain in robustness when we collect gold standard information on a subset of patients?