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 $\boldsymbol{Y}_{i}=\left(Y_{i 1}, Y_{i 2}, \ldots, Y_{i J}\right)^{\prime}$.
- Let $d_{i}$ be the true binary disease status.

$$
P\left(Y_{i 1}, Y_{i 2}, \ldots, Y_{i J}\right)=\sum_{l=0}^{1} P\left(Y_{i 1}, Y_{i 2}, \ldots, Y_{i J} \mid d_{i}=l\right) P\left(d_{i}=l\right)
$$

- Different models for $P\left(Y_{i 1}, Y_{i 2}, \ldots, Y_{i J} \mid d_{i}=l\right)$


## Conditional Independence (Hui and Walters, 1980)

- $Y_{i j} \mid d_{i} \sim$ Bernoulli with probability $P\left(Y_{i j}=1 \mid d_{i}\right)$
- $P\left(Y_{i 1}, Y_{i 2}, \ldots, Y_{i J} \mid d_{i}=l\right)=\prod_{j=1}^{J} P\left(Y_{i j} \mid d_{i}=l\right)$
- Sensitivity $=P\left(Y_{i j}=1 \mid d_{i}=1\right)$
- Specificity $=P\left(Y_{i j}=0 \mid d_{i}=0\right)$


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

- $Y_{i j} \mid d_{i}, b \sim$ Bernoulli with probability $\Phi\left(\beta_{d_{i}}+\sigma_{d_{i}} b\right)$ where $b \sim N(0,1)$ is a subject-specific random effect
- $P\left(Y_{i 1}, Y_{i 2}, \ldots, Y_{i J} \mid d_{i}=l\right)=\int\left\{\prod_{j=1}^{J} P\left(Y_{i j} \mid d_{i}, b\right)\right\} \phi(b) d b$ where the integral can be evaluated using Gaussian Quadrature
- Sensitivity $=P\left(Y_{i j}=1 \mid d_{i}=1\right)=E_{b}\left\{P\left(Y_{i j}=1 \mid d_{i}=1\right)\right\}$

$$
=\Phi\left(\beta_{1} / \sqrt{1+\sigma_{1}^{2}}\right)
$$

- Specificity $=P\left(Y_{i j}=0 \mid d_{i}=0\right)=E_{b}\left\{P\left(Y_{i j}=0 \mid d_{i}=0\right)\right\}$

$$
=1-\Phi\left(\beta_{0} / \sqrt{1+\sigma_{0}^{2}}\right)
$$

## Beta-Binomial Model

- $\sum_{j} Y_{i j}=a \mid d_{i}=0 \sim \operatorname{Beta-\operatorname {binomial}(\alpha _{0},\beta _{0})}$ and $\sum_{j} Y_{i j}=a \mid d_{i}=1 \sim \operatorname{Beta-binomial}\left(\alpha_{1}, \beta_{1}\right)$
- $P\left(Y_{i 1}, Y_{i 2}, \ldots, Y_{i J} \mid d_{i}=l\right)=P\left(\sum_{j} Y_{i j}=a \mid d_{i}=l\right) /\binom{J}{a}$
- Sensitivity $=\alpha_{1} /\left(\alpha_{1}+\beta_{1}\right)$
- Specificity $=1 /\left(\alpha_{0}+\beta_{0}\right)$


## Finite Mixture Model

- $P_{0}=P\left(\right.$ Not Subject to Misclassification $\left.\mid d_{i}=0\right)$
- $P_{1}=P\left(\right.$ Not Subject to Misclassification $\left.\mid d_{i}=1\right)$
- $P\left(Y_{i 1}, \ldots, Y_{i J} \mid d_{i}=1\right)=$
$\begin{cases}P_{1}+\left(1-P_{1}\right) \prod_{j} P_{d_{i}=1}\left(Y_{i j}=1\right) & \text { Tests all ones } \\ \left(1-P_{1}\right) \prod_{j} P_{d_{i}=1}\left(Y_{i j}\right) & \text { Test not all ones }\end{cases}$
where $P_{d_{i}}\left(Y_{i j}=1\right)$ is the probability of $Y_{i j}=1$ given the patient is subject to misclassification and $d_{i}$ is the true binary disease status.
- Sensitivity $=P_{1}+\left(1-P_{1}\right) P_{d_{i}=1}\left(Y_{i j}=1\right)$
- Specificity $=P_{0}+\left(1-P_{0}\right) P_{d_{i}=0}\left(Y_{i j}=0\right)$


## 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.

| Expected Frequency |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Comparison of Methods on Handelman's Dentistry Data |  |  |  |  |  |
|  |  |  |  |  |  |
| Pos. Tests | Freq. | Indep | $F M$ | $B B$ | 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 |  |  |  |  |
| $S \widehat{E N S}$ |  | 0.658 | 0.645 | 0.518 | 0.457 |
|  |  | $(0.017)$ | $(0.026)$ | $(0.076)$ | $(0.088)$ |
| $S \widehat{P E} C$ |  | 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 |
| $\chi^{2}$ |  | 20.773 | 1.293 | 2.317 | 1.979 |
| $d f$ |  | 3 | 1 | 1 | 1 |
|  |  |  |  |  |  |

Estimation of rater-specific sensitivity and specificity

|  |  | Indep | $F M$ | GRE |
| :---: | :---: | :---: | :---: | :---: |
| Rater |  | Est.(SE $\left.{ }^{1}\right)$ | 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 | $0.69(0.011)$ | $0.67(0.012)$ | $0.64(0.032)$ |
| $\operatorname{logLik}$ |  | -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

$$
\boldsymbol{\theta}^{*}=\max _{\boldsymbol{\theta}} E_{T}\left[\log \mathrm{~L}\left(\boldsymbol{Y}_{i}, \boldsymbol{\theta}\right)\right]
$$

- $E_{T}\left(\log L_{M}\right)=\left.E_{T}\left[\log L\left(\boldsymbol{Y}_{i}, \boldsymbol{\theta}\right)\right]\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*}}$
- SENS $^{*}=g_{1}\left(\boldsymbol{\theta}^{*}\right)$ and $S P E C^{*}=g_{2}\left(\boldsymbol{\theta}^{*}\right)$.
- Estimates of sensitivity and specificity converge to $S E N S^{*}$ and $S P E C^{*}$ under mis-specified models.


## Asymptotic Results

Large sample robustness of the assumed latent class beta-binomial $(B B)$ model to the true dependence structure between tests. The true model is a finite mixture model $(F M)$ with $P_{0}=P_{1}=0.2$, $\mathrm{SENS}=0.75$ and $\mathrm{SPEC}=0.9$ for differing $P_{d}$.

| $P_{d}$ | $J$ | SENS $^{*}$ | SPEC $^{*}$ | $E_{T}\left[\log L_{F M}\right]$ | $E_{T}\left[\log L_{B B}\right]$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 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 |

$$
E_{F M}\left[\log L_{F M}\right]=E_{F M}\left[\log L_{G R E}\right]=-1.35814 .
$$

## Simulation Results

Simulated under the finite mixture model with $P_{d}=0.2$, $P_{0}=P_{1}=0.2$, SENS $=0.75$, and $\mathrm{SPEC}=0.90$.

|  | Avg. Est. GRE |  |  | Reject $\chi^{2}$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| I | J | SENS | SPEC | Indep | $F M$ | $G R E$ |
| 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 $P_{d}=0.2, \sigma_{0}=\sigma_{1}=1.5, \mathrm{SENS}=0.75$, and $\mathrm{SPEC}=0.90$.

|  |  | Avg. Est. FM |  | Reject $\chi^{2}$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| I | J | SENS | SPEC | Indep | $F M$ | $G R E$ |
| 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 $(F M)$ 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.
- $E_{T}\left(\log L_{M}\right) \approx E_{T}\left(\log L_{T}\right)$ 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?

