Published in final edited form as:
J Am Stat Assoc. 2008 December ; 103(484): 1518–1519. doi:10.1198/016214508000000940.

A few remarks on "A capture-recapture approach for screening using two diagnostic tests with availability of disease status for the test positives only" by Böhning and Patilea

Haitao Chu [Research Associate Professor] and
Department of Biostatistics and the Lineberger Comprehensive Cancer Center, The Univerity of North Carolina, Chapel Hill, NC 27516 (Email: hchu@bios.unc.edu).

Lei Nie [Mathematical Statistician]
Office of Biostatistics, Food and Drug Administration, Silver Spring, MD 20993 (Email: lei.nie@fda.hhs.gov).

Using a capture-recapture approach, Böhning and Patilea (2008) proposed two useful estimators for unobserved cell counts, assuming homogeneous association of the screening tests over disease status. However, they are mistaken in claiming that the maximum likelihood estimators (MLEs) are difficult to obtain. The point of this note is to present closed-form MLEs for, in their notation: 1) the α model where $\alpha=p_{11}^{(i)}p_{00}^{(i)}/\left(p_{01}^{(i)}p_{10}^{(i)}\right)$ is assumed to be identical for all $i = 1, 2, \ldots, d$; and 2) the θ model where $\theta=p_{1|1}^{(i)}/p_{1+}^{(i)}$ is assumed to be identical for all i.

One way to write the likelihood function (ignoring constant terms) in this setting is in terms of q_i and $p_{jk}^{(i)}(i=1,2,\ldots,d;j=0,1;k=0,1)$, as the authors did:

$$x_{00}^{(+)}\log\left(\sum_i p_{00}^{(i)}q_i\right)+\sum_i x_{11}^{(i)}\log\left(p_{11}^{(i)}q_i\right)+\sum_i x_{10}^{(i)}\log\left(p_{10}^{(i)}q_i\right)+\sum_i x_{01}^{(i)}\log\left(p_{01}^{(i)}q_i\right). \tag{1}$$

This parameterization involves a mixture likelihood, preventing closed-form solution for the MLEs. To obtain closed-form MLEs, we consider an alternative parameterization in terms of π_{jk} and $\pi_{jk}^{(i)}(j,k=0,1$ and $i=1,2,\ldots,d)$ where $\pi_{jk}=P(T_1=j,T_2=k)$, $\pi_{jk}^{(i)}=P(D=i|T_1=j,T_2=k)$. The log-likelihood function is (ignoring constant terms),

$$\begin{aligned}\log L&=\sum_j\sum_k x_{jk}^{(+)}\log\left(\pi_{jk}\right)+\sum_i x_{11}^{(i)}\log\left(\pi_{11}^{(i)}\right)+\sum_i x_{10}^{(i)}\log\left(\pi_{10}^{(i)}\right)+\sum_i x_{01}^{(i)}\log\left(\pi_{01}^{(i)}\right)\\&=x_{00}^{(+)}\log(\pi_{00})+\sum_d x_{11}^{(i)}\left\{\log\left(\pi_{11}^{(i)}\right)+\log(\pi_{11})\right\}+\sum_i x_{10}^{(i)}\left\{\log\left(\pi_{10}^{(i)}\right)+\log(\pi_{10})\right\}+\sum_i x_{01}^{(i)}\left\{\log\left(\pi_{01}^{(i)}\right)+\log(\pi_{01})\right\}\\&=x_{00}^{(+)}\log(\pi_{00})+\sum_i x_{11}^{(i)}\log\left(\pi_{11}^{(i)}\pi_{11}\right)+\sum_i x_{10}^{(i)}\log\left(\pi_{10}^{(i)}\pi_{10}\right)+\sum_i x_{01}^{(i)}\log\left(\pi_{01}^{(i)}\pi_{01}\right)\end{aligned} \tag{2}$$

This representation relates to previous work in some other settings (Satten and Kupper 1993; Lyles 2002; Pepe and Janes 2007). Note that,

$\pi_{jk}^{(i)}\pi_{jk}=P(D=i|T_1=j,T_2=k)P(T_1=j,T_2=k)=P(T_1=j,T_2=k|D=i)P(D=i)=p_{jk}^{(i)}q_i$, and

$\pi_{00}=P(T_1=0,T_2=0)=\sum_i P(T_1=0,T_2=0,D=i)=\sum_i P(T_1=0,T_2=0|D=i)P(D=i)=\sum_i p_{00}^{(i)}q_i$

Therefore equation (2) is equivalent to equation (1).

These equations are tractable and yield closed-form MLEs of π_{jk} ($j, k = 0, 1$) and $\pi_{jk}^{(i)}$ if $j + k > 0$. Omitting the algebra, we obtain the MLEs as $\hat{\pi}_{jk}=x_{jk}/n(j,k=0,1)$ and $\hat{\pi}_{jk}^{(i)}=x_{jk}^{(i)}/x_{jk}$ if $j + k > 0$. Therefore, the MLEs of q_i s, which can be written as functions of π_{jk} ($j, k = 0, 1$) and $\pi_{jk}^{(i)}(j+k>0)$ under the α or θ model assumptions, have closed-form solutions. The details are given below.

Under the α model, $\alpha=\frac{p_{11}^{(i)}p_{00}^{(i)}}{p_{01}^{(i)}p_{10}^{(i)}}$ is assumed to be identical for all $i = 1, 2, \ldots, d$; by Bayes' rule,

$$\begin{aligned}\alpha&=\frac{p_{11}^{(i)}p_{00}^{(i)}}{p_{01}^{(i)}p_{10}^{(i)}}=\frac{P\,(T_1=1,T_2=1|D=i)\;P\,(T_1=0,T_2=0|D=i)}{P\,(T_1=0,T_2=1|D=i)\;P\,(T_1=1,T_2=0|D=i)}\\&=\frac{P\,(D=i|T_1=1,T_2=1)\;P\,(T_1=1,T_2=1)\;P\,(D=i|T_1=0,T_2=0)\;P\,(T_1=0,T_2=0)}{P\,(D=i|T_1=1,T_2=0)\;P\,(T_1=1,T_2=0)\;P\,(D=i|T_1=0,T_2=1)\;P\,(T_1=0,T_2=1)}\\&=\frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}\times\frac{\pi_{11}^{(i)}\pi_{00}^{(i)}}{\pi_{01}^{(i)}\pi_{10}^{(i)}},\end{aligned}$$

Thus

$$\alpha=\frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}\times\left[\sum_i\frac{\pi_{01}^{(i)}\pi_{10}^{(i)}}{\pi_{11}^{(i)}}\right]^{-1},\quad \pi_{00\alpha}^{(i)}=\frac{\pi_{01}^{(i)}\pi_{10}^{(i)}}{\pi_{11}^{(i)}}\times\left[\sum_i\frac{\pi_{01}^{(i)}\pi_{10}^{(i)}}{\pi_{11}^{(i)}}\right]^{-1},$$

$$q_{i\alpha}=\pi_{11}\pi_{11}^{(i)}+\pi_{10}\pi_{10}^{(i)}+\pi_{01}\pi_{01}^{(i)}+\pi_{00}\frac{\pi_{01}^{(i)}\pi_{10}^{(i)}}{\pi_{11}^{(i)}}\left[\sum_i\frac{\pi_{01}^{(i)}\pi_{10}^{(i)}}{\pi_{11}^{(i)}}\right]^{-1},$$

where the subscript α indicates the α model assumption. Since the MLEs of the parameters π_{jk} and $\pi_{jk}^{(i)}$ are $\hat{\pi}_{jk}=x_{jk}/n$ $(j,k=0,1)$ and $\hat{\pi}_{jk}^{(i)}=x_{jk}^{(i)}/x_{jk}$ if $j + k > 0$, the closed-form MLE of $n_{i\alpha}$ under the α model is

$$\widehat{n}_{i\alpha}=n\widehat{q}_{i\alpha}=x_{11}^{(i)}+x_{10}^{(i)}+x_{01}^{(i)}+x_{00}\frac{x_{01}^{(i)}x_{10}^{(i)}}{x_{11}^{d}}\left[\sum_i\frac{x_{01}^{(i)}x_{10}^{(i)}}{x_{11}^{(i)}}\right]^{-1},\tag{3}$$

which is essentially the same as the equation (15) in Böhning and Patilea (2008) without the stability correction. In other words, the estimator obtained in equation (15) is the MLE under the α model assumption with the stability correction.

Under the θ model $\theta=\frac{p_{1|1}^{(i)}}{p_{1+}^{(i)}}$ is assumed to be identical for all $i = 1, 2, \ldots, d$; by Bayes' rule

$$\begin{aligned}\theta&=\frac{p_{1|1}^{(i)}}{p_{1+}^{(i)}}=\frac{P\,(T_1=1|T_2=1,D=i)}{P\,(T_1=1|D=\mathrm{i})}=\frac{P\,(D=i|T_1=1,T_2=1)\;P\,(T_1=1,T_2=1)}{P\,(T_2=1,D=i)\;P\,(T_1=1,D=i)}\times P\,(D=i)\\&=\frac{\pi_{11}\pi_{11}^{(i)}}{\left(\pi_{01}\pi_{01}^{(i)}+\pi_{11}\pi_{11}^{(i)}\right)\left(\pi_{10}\pi_{10}^{(i)}+\pi_{11}\pi_{11}^{(i)}\right)}\times P\,(D=i),\end{aligned}$$

$$\text{Thus } \theta=\left[\sum_i\left(\frac{\pi_{10}\pi_{10}^{(i)}}{\pi_{11}\pi_{11}^{(i)}}+1\right)\left(\pi_{01}\pi_{01}^{(i)}+\pi_{11}\pi_{11}^{(i)}\right)\right]^{-1} \text{ and}$$

$$\pi_{00\theta}^{(i)}=\frac{1}{\pi_{00}}\left\{\left(\frac{\pi_{10}\pi_{10}^{(i)}}{\pi_{11}\pi_{11}^{(i)}}+1\right)\left(\pi_{01}\pi_{01}^{(i)}+\pi_{11}\pi_{11}^{(i)}\right)\left[\sum_i\left(\frac{\pi_{10}\pi_{10}^{(i)}}{\pi_{11}\pi_{11}^{(i)}}+1\right)\left(\pi_{01}\pi_{01}^{(i)}+\pi_{11}\pi_{11}^{(i)}\right)\right]^{-1}-\pi_{11}\pi_{11}^{(i)}-\pi_{10}\pi_{10}^{(i)}-\pi_{01}\pi_{01}^{(i)}\right\}$$

$$q_{i\theta}=\left(\frac{\pi_{10}\pi_{10}^{(i)}}{\pi_{11}\pi_{11}^{(i)}}+1\right)\left(\pi_{01}\pi_{01}^{(i)}+\pi_{11}\pi_{11}^{(i)}\right)\left[\sum_i\left(\frac{\pi_{10}\pi_{10}^{(i)}}{\pi_{11}\pi_{11}^{(i)}}+1\right)\left(\pi_{01}\pi_{01}^{(i)}+\pi_{11}\pi_{11}^{(i)}\right)\right]^{-1},$$

where the subscript θ indicates the θ model assumption. Similarly, the closed-form MLE of under $n_{i\theta}$ under the θ model is

$$\widehat{n}_{i\theta}=n\widehat{q}_{i\theta}=\left(\frac{x_{10}^{(i)}}{x_{11}^{(i)}}+1\right)\left(x_{01}^{(i)}+x_{11}^{(i)}\right)\left[\sum_i\left(\frac{x_{10}^{(i)}}{x_{11}^{(i)}}+1\right)\left(x_{01}^{(i)}+x_{11}^{(i)}\right)\right]^{-1}=\frac{x_{1+}^{(i)}x_{+1}^{(i)}}{x_{11}^{(i)}}\left[\sum_i\frac{x_{1+}^{(i)}x_{+1}^{(i)}}{x_{11}^{(i)}}\right]^{-1}, \quad (4)$$

which is essentially the same as the equation (10) in Böhning and Patilea (2008) without the stability correction.

As a byproduct of this alternative parameterization, we can test the difference between $\hat{}q_{i\theta}$ and $\hat{}q_{i\alpha}$ (or equivalently, the difference between $\hat{}n_{i\theta}$ and $\hat{}n_{i\alpha}$) to make inference on whether these two assumptions provide statistically significantly different predictions for the probability (or equivalently, the number) of individuals with certain disease class *i*. Although the formula for $se(\hat{}q_{i\theta}-\hat{}q_{i\alpha})$ is tedious, its numerical value can be obtained easily through statistical software using the delta method. We note that the difference between estimated probabilities of disease classes under the α and θ models can be statistically different and potentially meaningful for the same study. For example, in the Health Insurance Plan Study for breast cancer screening in New York (Strax, Venet Shapiro and Gross 1967), the estimated probability of having cancer assuming the α model is 4.8% with a 95% confidence interval (CI) of 0.3% to 9.3%, while the estimated probability of having cancer assuming the θ model is 7.5% with 95% CI of 2.8% to 12.2%. The difference is 2.7% (95% CI: 1.4% to 4%) with a p-value less than 0.001. This difference can have a big impact on the cancer surveillance and prevention. Unfortunately, the data does not contain information to differentiate the α model versus the θ model.

The alternative parameterization in (2) sheds lights on maximum likelihood approaches in the setting considered here; the corresponding closed-form ML estimators under the α and θ models allow tests of the difference between the estimated probabilities of a specific disease class using the α versus the θ model. Our results complements the estimators obtained in equations (10) and (15) by Böhning and Patilea (2008) using a capture-recapture approach, and ensure the usual MLE properties.

Acknowledgments

Dr. Chu was supported in part by the Lineberger Cancer Center Core Grant CA16086 from the U.S. National Cancer Institute. The authors are very grateful to the editor for his helpful comments and suggestions.

References

Böhning D, Patilea V. A capture-recapture approach for screening using two diagnostic tests with availability of disease status for the test positives only. Journal of the American Statistical Association 2008;103:212–221.

Lyles RH. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. Biometrics 2002;58:1034–1036. [PubMed: 12495160]

Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics 2007;8:474–484. [PubMed: 17085745]

Satten GA, Kupper LL. Inferences About Exposure-Disease Associations Using Probability-Of-Exposure Information. Journal of the American Statistical Association 1993;88:200–208.

Strax P, Venet L, Shapiro S, Gross S. Mammography and Clinical Examination in Mass Screening for Cancer of the Breast. Cancer 1967;20:2184–2188. [PubMed: 6073895]