Adjusting Diagnostic Test Accuracy for Numerous Confounders: The Mantel-Haenszel Approach

Gene Pennello*[1], Huan Wang[2], and Norberto Pantoja-Galicia[3]

Disclaimer: This poster reflects the views of the authors and should not be construed to represent the FDA's views or policies. [1]Food and Drug Administration, Division of Imaging, Diagnostic and Software Reliability; [2]The George Washington University, Department of Biostatistics; [3]Foundation Medicine; *gene.pennello@fda.hhs.gov

Abstract

Background: Diagnostic test accuracy studies are typically observational and thus subject to confounding. Logistic regression could be used to adjust the odds ratio between test result and disease status for confounding strata, but the maximum likelihood (ML) estimator is inconsistent for the true value when the number of strata grows with sample size, i.e., the *sparse data limiting model* (SDLM). In contrast, the Mantel-Haenszel (MH) estimator of the odds ratio adjusted for stratum effects (**Mantel, Haenszel**, 1959) is consistent under SLDM. However, diagnostic tests are typically evaluated not with the odds ratio, but with pairs of measures such as specificity and sensitivity, negative and positive predictive value (NPV, PPV), and negative and positive likelihood ratio (NLR, PLR).

Purpose: To improve clinical evaluation of diagnostic test accuracy studies, we develop MH estimators of NLR and PLR that adjust for confounding stratum effects. Given disease prevalence, we convert MH estimators of NLR and PLR to NPV and PPV using Bayes Theorem.

Methodology: We derive MH estimators of PLR and NLR, which are consistent under SDLM. Confidence intervals are based on consistent estimators of the variances of the MH estimates. MH and ML estimators are compared on three hypothetical datasets: (1) data with the same PLR in each stratum yet a different marginal PLR (marginalization paradox), (2) matched pairs data with two observations per stratum, and (3) neonatal audiology test data.

Results: ML and MH estimates of PLR were identical for the first dataset. For the matched pairs data, ML and MH estimates of PLR diverged, indicating that the ML estimator of PLR is inconsistent under SDLM. For the neonatal audiology test data, MH and ML estimates of NLR diverged for very fine strata, suggesting that the ML estimator is inconsistent under SDLM. When adjusting for 973 subject strata, MH and ML estimates of PPV and NPV were worse than the marginal estimates based on collapsing data over strata, suggesting that subject strata are confounders.

Conclusion: MH estimators can be used to adjust diagnostic test accuracy for numerous confounding strata and are valid even in sparse data settings..

Diagnostic Likelihood Ratios

Let variables $T = t$ and $D = d$ be the binary test result and binary disease status, where $t = 0$ and 1 denote negative and positive test results and $d = 0$ and 1 denote disease absence and presence. Given $T = t$, the diagnostic likelihood ratio is

$$l_t = \frac{Pr(T = t|D = 1)}{Pr(T = t|D = 0)}$$

The **positive likelihood ratio (PLR)** is

$$\boldsymbol{l_1 = \frac{p_1}{p_0}},$$

where $p_1 = Pr(T = 1|D = 1)$ is the **sensitivity** of the test and $p_0 = Pr(T = 1|D = 0)$ is one minus the **specificity** of the test. The **negative likelihood ratio (NLR)** is

$$\boldsymbol{l_0 = \frac{q_1}{q_0}},$$

where $q_t = 1 - p_t$. By **Bayes Theorem**,

$$o_t = l_t o$$

where

$$o_t = \frac{\pi_t}{1-\pi_t}, o = \frac{\pi}{1-\pi},$$
$$\pi_t = Pr(D = 1|T = t),$$
$$\pi = \Pr(D = 1),$$

$\pi_1 = \Pr(D = 1|T = 1)$ is the **positive predictive value (PPV)** of the test, $\pi_0 = \Pr(D = 1|T = 0)$ is one minus the **negative predictive value (NPV)** of the test, and π is **disease prevalence** (pre-test probability of disease). Thus, l_t is the change in the odds of disease from pre-test to post-test conferred by test result $T = t$.

Mantel-Haenszel Estimators

Consider K pairs of independent binomial variables n_{10k} and n_{11k} with

$$n_{1dk} \sim Bin(n_{\bullet dk}, p_{dk}),$$

$n_{\bullet dk} = n_{0dk} + n_{1dk}$ is the sample size for response variable n_{1dk}, p_{dk} is the probability of response, $d = 0$ or 1 indicates absence or presence of a disease condition (e.g.,cancer), and $k = 1,2, \ldots K$ are strata. For a binary test accuracy study, n_{tdk} is the count for the negative or positive test result $t = 0$ or 1. The strata may be defined by combinations of categorical covariates that may be associated with disease, test result, or both.

Within stratum $S = k$, the positive likelihood ratio is

$$l_{1k} = \frac{\Pr(T = 1|D = 1, S = k)}{\Pr(T = 1|D = 0, S = k)} = \frac{p_{1k}}{p_{0k}}$$

Provided n_{10k} is non-zero, the sample estimate is

$$\hat{l}_{1k} = \frac{\hat{p}_{1k}}{\hat{p}_{0k}} = \frac{n_{11k}/n_{\bullet 1k}}{n_{10k}/n_{\bullet 0k}} = \frac{n_{11k}n_{\bullet 0k}}{n_{10k}n_{\bullet 1k}}$$

Assuming a common value of l_{1k} in every stratum, that is, $l_{1k} \equiv l_1$ for all $k = 1,2, \ldots K$, the MH estimator of l_1 adjusted for stratum effects is

$$\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{MH}} = \frac{\sum_{k=1}^{K} n_{11k}n_{\bullet 0k}/n_{\bullet\bullet k}}{\sum_{k=1}^{K} n_{10k}n_{\bullet 1k}/n_{\bullet\bullet k}}$$

Analogously, the MH estimator of a common value l_0 for the negative likelihood ratio adjusted for stratum effects is

$$\hat{\boldsymbol{l}}_{\boldsymbol{0}}^{\boldsymbol{MH}} = \frac{\sum_{k=1}^{K} n_{01k}n_{\bullet 0k}/n_{\bullet\bullet k}}{\sum_{k=1}^{K} n_{00k}n_{\bullet 1k}/n_{\bullet\bullet k}}$$

Generalized Linear Model Estimators

Consider the linear predictor

$$\log(p_{dk}) = \alpha_{1k} + \beta_1 d,$$

where $\{\alpha_{1k}\}$ are stratum effects and $\beta_1 = \log(l_1)$ is common to all strata. Similarly, consider the linear predictor is

$$\log(q_{dk}) = \alpha_{0k} + \beta_0 d$$

where $\{\alpha_{0k}\}$ are stratum effects and $\beta_0 = \log(l_0)$ is common to all strata. We use the two models to obtain ML estimates of l_1 and l_0 adjusted for stratum effects.

Table 1. Hypothetical Data for a test and three strata.

Stratum 1

Test	$D = 0$	$D = 1$	Total
$T = 0$	5	40	45
$T = 1$	1	20	21
Total	6	60	66

Stratum 2

Test	$D = 0$	$D = 1$	Total
$T = 0$	36	1	37
$T = 1$	24	4	28
Total	60	5	65

Stratum 3

Test	$D = 0$	$D = 1$	Total
$T = 0$	2	1	3
$T = 1$	1	2	3
Total	3	3	6

Margin

Test	$D = 0$	$D = 1$	Total
$T = 0$	43	42	85
$T = 1$	26	26	52
Total	69	68	137

Numerical Studies

EXAMPLE 1. ***Hypothetical data with Common Positive Likelihood Ratio***

Consider hypothetical data in which the per stratum sample estimate of l_{1k} is 2.0 for all three strata $k = 1,2,3$ (Table 1, Figure 1). Paradoxically, the marginal sample estimate of l_1 is 1.0, indicating that it is confounded by stratum effects. Adjusting for stratum effects, the MH and ML estimates of l_1 are both 2.0, in agreement with the per stratum sample estimates.

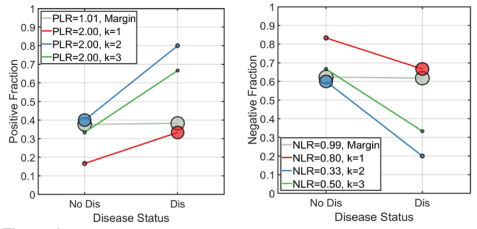

Figure 1. Plot of test positive and test negative fractions for the data in Table 1. Size of co-ordinates is proportional to sample size

EXAMPLE 2. ***Hypothetical Matched Pairs Data***

Consider hypothetical matched pairs data with one test positive and one test negative result in each of 7 strata. In 1 stratum, both subjects are diseased. In 3 strata, both subjects are disease free. In the remaining 3 strata, 1 subject had disease, the other did not, and for both subjects the test agreed with disease status in 2 strata and disagreed with disease status in 1 stratum. The marginal estimate of l_1, based on pooling the data across strata, is 1.35. Adjusting for stratum effects, the MH and ML estimates are $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{MH}} = 2.0$ and $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{ML}} = 3.0$.

The estimator $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{MH}}$ is consistent for l_1 under the sparse data limiting model (proof in **Pennello, Wang, and Pantoja-Galicia**, 2020). Under this model, strata grow with sample size. To simulate data under this model, a new dataset can be formed by replicating the strata a large number of times. By noticing that neither $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{MH}}$ nor $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{ML}}$ will change when the strata are replicated, and that the values of $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{MH}}$ and $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{ML}}$ disagree, we conclude that $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{ML}}$ is inconsistent for l_1 under the sparse data limiting model. This result is similar to the well-known result that the ML estimator of the common odds ratio $\bar{o} = l_1/l_0$ based on logistic regression is inconsistent in matched pairs data, tending to $\bar{o}^2$ instead of $\bar{o}$ (**Andersen**, 1980; **Breslow**, 1981).

EXAMPLE 3: ***Hypothetical Neonatal Audiology Data***

Consider hypothetical data in which newborn babies are screened for impaired hearing, described in **Pepe** (2003), available at DABS Datasets, and based on the design of a study of a distortion product otoacoustic emission (DPOAE) diagnostic test for detecting hearing impairment in adults (**Leisenring, Pepe**, 1998). In the study, 973 babies ranging in age from 19 .55 to 54 .49 days were tested in with 3 passive hearing tests (A, B, C) in each ear (left, right) and at each location (room, booth), yielding 3152 observations. Incomplete testing was common because testing is impossible when babies start to fuss. The reference standard was visual reinforcement audiometry (VRA) test, performed at age 9-12 months.

For analysis, we did not distinguish between the 3 tests, treating them as if they were the same test, thus evaluating their average performance. We assumed multiple test results per subject are independent, which is not true, but this assumption doesn't affect the point estimate, only its uncertainty.

We performed 3 analyses. First, we considered combinations of 1-day increment age groups (i.e., 19-20, 20-21, …, 55-56) with ear and location levels, which yielded 104 strata. Adjusting for stratum effects, the MH and ML estimates are similar for both l_1 ($\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{MH}} = 1.56$, $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{ML}} = 1.58$) and l_0 ($\hat{\boldsymbol{l}}_{\boldsymbol{0}}^{\boldsymbol{MH}} = 0.56$, $\hat{\boldsymbol{l}}_{\boldsymbol{0}}^{\boldsymbol{ML}} = 0.57$).

Second, we considered combinations of 0.5-day increment age groups (i.e., 19-19.5, 19.5-20, …, 55.5-56) with ear and location levels, which yielded 189 strata. Adjusting for the finer strata, MH and ML estimates are again similar for l_1 ($\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{MH}} = 1.56$, $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{ML}} = 1.59$), but diverge for l_0 ($\hat{\boldsymbol{l}}_{\boldsymbol{0}}^{\boldsymbol{MH}} = 0.56$, $\hat{\boldsymbol{l}}_{\boldsymbol{0}}^{\boldsymbol{ML}} = 1.23$). The ML estimate of $\hat{\boldsymbol{l}}_{\boldsymbol{0}}^{\boldsymbol{ML}} = 1.23$ seems clearly wrong, as $l_0 > 1$ indicates a test is worse than random. The divergence of the ML and MH estimates for l_0 is another indication that the ML estimates are inconsistent under the sparse data limiting model.

Finally, we stratified by subject (973 strata). Adjusting for subject effects, the MH and ML estimates for l_1 are $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{MH}} = 1.36$ and $\hat{\boldsymbol{l}}_{\boldsymbol{1}}^{\boldsymbol{ML}} = 1.13$ and for l_0 are $\hat{\boldsymbol{l}}_{\boldsymbol{0}}^{\boldsymbol{MH}} = 0.69$ and $\hat{\boldsymbol{l}}_{\boldsymbol{0}}^{\boldsymbol{ML}} = 0.88$. Collapsing over strata, the marginal estimates of l_1 and l_0 are 1.56, and 0.56, suggesting that subject effects confound the marginal estimates.

The per-observation prevalence of hearing impairment is $\hat{\pi} = 0.3985$ $(1256/3152)$. For this prevalence, when adjusting for subject effects, the MH and ML estimates of PPV are $\hat{\pi}_1^{MH} = 0.47$ and $\hat{\pi}_1^{ML} = 0.44$. The MH and ML adjusted estimates of NPV are $1 - \hat{\pi}_0^{MH} = 0.69$ and $1 - \hat{\pi}_0^{ML} = 0.63$. Collapsing over strata, the marginal estimates of PPV and NPV are 0.51 and 0.73.

Discussion

Diagnostic test accuracy data may not be collapsible over strata. MH estimators of l_0, l_1, NPV and PPV adjust for stratum effects and are consistent under the sparse data limiting model, whereas ML estimators are not. Thus, MH estimators are preferred for finely stratified or matched pairs data.

The MH and ML estimators assume l_{0k} and l_{1k} have common values l_0 and l_1 for all strata $k = 1,2, \ldots K$. However, except in trivial cases l_{0k} and l_{1k} cannot both have common values. Nonetheless, the MH estimates are useful summaries of average l_0 and l_1 if the per stratum values are all in the same direction, i. e., $l_{1k} > 1$ and $l_{0k} < 1$ for all k, which is often expected for a good diagnostic test.

ML estimates of l_0 and l_1 are model-based. Our ML estimate of l_1 (l_0) was based on assuming the log of the probability of a test positive (negative) result is linear in stratum effect and disease status. Alternatively, **Gu and Pepe** (2011) obtain ML estimates of l_0 and l_1 using a logistic model for the probability of disease.

References

Andersen EB. (1980). *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.

Breslow NE. (1981). Odds ratio estimators when the data are sparse. *Biometrika* 68, 73-84.

Gu W, Pepe MS (2011) Estimating the diagnostic likelihood ratio of a continuous marker. *Biostatistics*; 12 (1): 87–101.

Leisenring W, Pepe MS. (1998). Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. Biometrics, 54, 444-452.

Mantel N, Haenszel W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst,* 22, 719-48

Pennello G, Wang H, Pantoja-Galicia N. (2020) Mantel-Haenszel Estimators of Positive and Negative Likelihood Ratios *Statist Biopharm Res*, submitted.

Pepe M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford, 2003.