Methods for Diagnostic Performance Comparison with Correlated Data

Xuan Ye
Center for Devices and Radiological Health, Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Abstract
In new diagnostic device application, the new measurement method or diagnostic device is often compared to a predicate device or a comparator. We need to demonstrate that the diagnostic performance of the subject device is substantially equivalent to the predicate's. In clinical trials, both methods are often applied to the same set of patients producing correlated dataset. In this article, we compare various parametric and non-parametric methods to address the issue and discuss the analysis results through simulation studies.

Key Words: Diagnostic, Correlated, Z-test, Bootstrap, Bayesian, Power

1. Introduction

Based on a dichotomous diagnostic test output and the true disease status, a subject in a clinical trial can be classified into one of the four categories: true positive, true negative, false positive and false negative.

		Disease Status	
		Positive(D=1)	Negative(D=0)
Test Result	Positive(X=1)	True Positive	False Positive
	Negative(X=0)	False Negative	True Negative

The diagnostic performances such as sensitivity and specificity can be defined as conditional probabilities. For example, Sensitivity = P(X = 1 | D = 1); Specificity = P(X = 0 | D = 0). In addition to binary results that a diagnostic test may produce, ordinal results or continuous results are often seen in diagnostic device outputs. In the ordinal or quantitative outputs, often a cut-off value will be pre-specified and applied in determining disease status. The sensitivity and specificity in these cases can be defined as Sensitivity(c) = P(X > c | D = 1); Specificity(c) = P(X $\leqslant$ c |D = 0), where c is the chosen cut-off value and any subject with measurement above c is diagnosed as diseased according to the test.

In a new diagnostic device application, the new measurement method or diagnostic device is often compared to a predicate device as comparator. We need to demonstrate that the diagnostic performance of the subject device is substantially equivalent to the predicate's. In clinical trials, both methods are often applied to the same set of patients producing correlated dataset. In detail, from a diseased subject i, we have both test 1 and test 2 measurements $X_{1,i}$ and $X_{2,i}$. Similarly, for a non-diseased subject j, we have measurements $Y_{1,i}$ and $Y_{2,i}$ from test 1 and test 2, respectively. It is reasonable to assume that measurements from different subjects are independent. However, the measurements of test 1 and test 2 from the same subject are probably correlated. In this article, we consider the case that both tests have quantitative measurement outputs, and that cut-off values c1 and

c2 will be applied for test1 and test2 measurements respectively in medical practice. The clinical study's objective is to compare the sensitivities, specificities of the two diagnostic tests. Since the comparison is based on the underlying correlated test data. We want to find an appropriate statistical method to achieve optimal power while controlling the type I error rate at the nominal α level. We will investigate several statistical methods, such as Z-test for two proportions that ignores or accounts for the correlation, Bootstrap method, Bayesian method, and compare their characteristics.

2. Methods for Correlated Data

2.1 Simulation Study Setting

Suppose we want to conduct hypothesis test on the sensitivities comparison, with H0: $\text{Sensitivity}_1 \leqslant \text{Sensitivity}_2$ vs. Ha: $\text{Sensitivity}_1 > \text{Sensitivity}_2$. In this article, we address the sensitivities comparison only, the specificities comparison is similar.

We use bivariate normal model to simulate two sets of diagnostic test data for test 1 and test 2. We generate data from $(\mathbf{X}1, \mathbf{X}2)^T \sim N\{ (0, 0)^T, \Sigma\}$ under H0 condition, and $(\mathbf{X}1, \mathbf{X}2)^T \sim N\{ (0.3, 0)^T, \Sigma\}$ under Ha condition such that test 1 has superior sensitivity. The covariance matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ with ρ = {0, 0.25, 0.5, 0.75, 0.9}. The cutoff values are 0 for both tests. We simulate data with sample sizes n = 200 for both tests, and with 10,000 simulations. Type I error rates are to be controlled at α = 0.025 level for one-sided hypothesis tests.

2.2 Z-test Ignoring the Correlation

First, we apply the common Z-test method for comparing two independent proportions. The method ignores the possible correlation between tests by assuming independence. The Z test statistic is $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ under H_0 condition, where p_1 and p_2 are sensitivities for Test 1 and Test 2. Under the aforementioned simulation setting, we calculate the observed type I error rate under H0 condition and the observed power under Ha condition, respectively, with various ρ values.

Table 1: Z-test Ignoring the Correlation: Observed type I error rate under H0 and Observed Power under Ha condition.

ρ	Type I error rate under H0	Powers under the Ha condition
0	2.54 %	66.83 %
0.25	1.54 %	66.91 %
0.5	0.78 %	69.15 %
0.75	0.17 %	72.80 %
0.9	0.01 %	76.61 %

It appears that this method tends to overestimate the variance when ρ is greater than 0, hence under value the test statistic. It is too conservative in type I error rate control, with the observed type I error rates smaller than the nominal α (0.025) as ρ becomes larger. Probably it is conservative in type I error control at the cost of losing power. Next, we will apply Bootstrap method on the same dataset and compare the achieved powers.

2.3 Bootstrap Method

We apply the nonparametric method on the same simulation dataset, and control the type I error rate at α=0.025. We apply resampling technique on the subject level such that the underlying within-subject correlation is preserved. As shown in the following table, Bootstrap method achieves more power than the Z test method that ignores the correlation under the alternative hypothesis condition, while the observed type I error rates are around the nominal α level under H0 condition. Bootstrap method is computation burdensome. Hence, we will look again at the Z-test method in the next section and derive a closed form formula for the variance estimation that accounts for the correlation.

Table 2: Bootstrap Method: Observed type I error rate under H0 and Observed Power under Ha condition.

ρ	Type I error rate under H0	Power under Ha condition	Compared to Power of the Z-test that ignores correlation
0	2.69 %	66.42 %	66.83 %
0.25	2.34 %	73.62 %	66.91 %
0.5	2.37 %	83.33 %	69.15 %
0.75	2.26 %	93.70 %	72.80 %
0.9	2.61 %	99.09 %	76.61 %

2.4 Z-test Method that Accounts for the Correlation

We know that the variance of $(\widehat{p_1} - \widehat{p_2})$ is $Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) - 2\mathrm{Cov}(\hat{p}_1, \hat{p}_2)$. Here the empirical estimator of p_1 and p_2 are:

$\hat{p}_1 = \frac{1}{n} \sum_{i=1}^{n} I(X_{1,i} > c_1)$ and

$\hat{p}_2 = \frac{1}{n} \sum_{i=1}^{n} I(X_{2,i} > c_2)$.

Based on our assumptions, we can derive the formula for the covariance as following,

$\mathrm{Cov}(\hat{p}_1, \hat{p}_2) = \frac{1}{n}(S(c_1, c_2) - p_1 \cdot p_2)$,

where function S is the joint tail distribution (survival) function S(x,y) of test 1 and test 2 data, which can be estimated empirically.

Since we now have the closed form formula for variance of $(\hat{p}_1 - \hat{p}_2)$ that accounts for the correlation and can be estimated empirically, we apply the Z-test method with more accurate variance estimation. This method is less computation burdensome compared to Bootstrap method. The simulation study results are displayed in the following table. We can see that the observed type I error rates are controlled around α=0.025, and that it achieves more power compared to the Z test method that ignores the correlation.

Table 3: Z-test Method that Accounts for the Correlation: Observed type I error rate under H0 and Observed Power under Ha condition.

ρ	Type I error rate under H0	Power under Ha condition	Compared to Power of the Z-test that ignores the correlation
0	2.65 %	67.51 %	66.83 %
0.25	2.45 %	73.52 %	66.91 %
0.5	2.86 %	82.24 %	69.15 %
0.75	2.58 %	93.62 %	72.80 %
0.9	2.51 %	99.35 %	76.61 %

2.4 Bayesian Method

We further investigate the Bayesian method for the correlated data issue. If we assume the correlated test data for test 1 and test 2 follow bivariate Gaussian distribution, we can apply conjugate Bayesian analysis on the multivariate Gaussian distribution. We utilize a Normal-Inverse-Wishart conjugate prior for the multivariate normal distribution with unknown means and unknown covariance matrix. The following are the assumptions on the covariance matrix and means:

$\Sigma \sim IW_{\nu_0}(\Lambda_0^{-1})$,

$\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$,

$\mathrm{p}(\mu, \Sigma) \stackrel{\text{def}}{=} NIW(\mu_0, \kappa_0, \Lambda_0, \nu_0)$.

Then we have the following posterior distributions:

$\mathrm{p}(\mu, \Sigma|D, \mu_0, \kappa_0, \Lambda_0, \nu_0) = NIW(\mu_n, \kappa_n, \Lambda_n, \nu_n)$,

and the posterior marginals are:

$\Sigma|D \sim IW(\Lambda_n^{-1}, \nu_n)$,

$\mu|\Sigma, D \sim N(\mu_n, \ \Sigma/\kappa_n)$.

In addition, for the problem under investigation, we apply non-informative prior distribution (multivariate Jeffreys prior density), i.e.:

$\kappa_0 \to 0, \nu_0 \to -1, |\Lambda_0| \to 0$

then we have,

$\mu_n = \bar{x}, \ \kappa_n = n, \nu_n = n - 1,$

$\Lambda_n = S = \sum_i (x_i - \bar{x})(x_i - \bar{x})^T$ where S is the sum of squares matrix about the sample mean.

Consequently, the corresponding posterior distribution is:

$\Sigma|x \sim IW(S, n-1); \quad \mu|\Sigma, x \sim N(\bar{x}, \ \Sigma/n).$

With the above close-form formula, Σ and μ samples can be drawn from the posterior distributions. We can then apply the cutoffs to estimate $p_1, \ p_2$. We will reject the null hypothesis if the posterior probability of p_1 greater than p_2 is more than 97.5%.

We apply the method to the same simulation data set and have the following results (Table 4). We see that the observed type I error rate controlled at about the nominal α and it achieves even more power compared with the Bootstrap method. Furthermore, since the predicate device might have been in the market for a long time, we might have reliable information about the device such as its measurement variance. We can incorporate the information in the Bayesian model and utilize an informative prior. Consequently, we might have an even more efficient test method.

Table 4: Bayesian Method: Observed type I error rate under H0 and Observed Power under Ha condition.

ρ	Type I error rate under H0	Power under Ha condition	Compared to Power of the Bootstrap method
0	2.50 %	84.27 %	66.42 %
0.25	2.53 %	92.81 %	73.62 %
0.5	2.48 %	98.58 %	83.33 %
0.75	2.56 %	99.99 %	93.70 %
0.9	2.55 %	99.99 %	99.09 %

3. Methods for Cluster-Correlated Data

3.1 Cluster-Correlated Test Data

In addition to the correlated test data mentioned above, we sometimes see cluster-correlated diagnostic tests data. From the same subject, we might have multiple measurements for each test. For example, we can have measurements from both left and right eyes for each of the diagnostic tests, or have multiple measurements from multiple tissues of the same subject. From a diseased subject i, we have both test 1 and test 2 measurements. For test 1, we have multiple measurements $X_{i,1}^{(1)}, X_{i,2}^{(1)}, X_{i,3}^{(1)}$, and for test 2, multiple measurements $X_{i,1}^{(2)}, X_{i,2}^{(2)}, X_{i,3}^{(2)}$. It is similar for a non-diseased subject j. Again, we assume that measurements from different subjects are independent. Measurements of test 1 and test 2 from the same subject are probably correlated, i.e. between-test correlation. And the multiple measurements from the same subject of the same test are also correlated, i.e. within-test correlation. Cut-off values c1 and c2 were set for the test1 and test2 measurements respectively in medical practice.

3.2 Simulation Study Setting

Similar to the correlated data study setting, we conduct hypothesis tests of H0: $\text{Sensitivity}_1 \leq \text{Sensitivity}_2$ vs. Ha: $\text{Sensitivity}_1 > \text{Sensitivity}_2$, with type I error rate controlled at $\alpha = 0.025$. Since we have within-test and between-test correlations, we use two parameters λ and ρ to represent the between-test and within-test correlations, respectively. Two sets of diagnostic test data with underlying correlation are simulated using the following procedure. First we draw sample vectors $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$ and $\boldsymbol{Y}_3$, where $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})^T$, $i = 1,2,3;$ with cluster size m. $\boldsymbol{Y}_i \sim \text{N}\{(0, \ldots, 0)^{\text{T}}, \Sigma\}$ and

$$\Sigma = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix} \text{ with } \rho = \{0, 0.2, 0.4, 0.6, 0.8\},\ \lambda = \{0, 0.2, 0.4, 0.6, 0.8\}.$$

Then for test 1 data, we have $\boldsymbol{X}^{(1)} = \boldsymbol{Y}_1\sqrt{\lambda} + \boldsymbol{Y}_2\sqrt{1-\lambda}$. And for test 2 data, $\boldsymbol{X}^{(2)} = \boldsymbol{Y}_1\sqrt{\lambda} + \boldsymbol{Y}_3\sqrt{1-\lambda}$. We set both cutoff values c_1 and c_2 at 0. In addition, we let $\boldsymbol{X}^{(1)} = \boldsymbol{X}^{(1)} + \delta$ for the alternative Ha condition, where $\delta = \{0.1, 0.2, 0.3, 0.4\}$, such that test 1 is superior to test 2. We simulated sample sizes n = 50, cluster size m=5 for both tests with 10,000 simulation repetition.

3.3 Simulation Results

The following table shows the clustered data simulation results under H0 condition. The observed type I error rates (%) for various λ (between-test) and ρ (within-test) setting, with nominal α=0.025.

Table 5: Observed type I error rate under H0 (%) for various λ and ρ

ρ	ρ = 0.2				ρ = 0.6				ρ = 0.8			
$\boldsymbol{\lambda}$	λ=0	0.2	0.6	0.8	λ=0	0.2	0.6	0.8	λ=0	0.2	0.6	0.8
Z-test (ignore)	5.64	3.90	1.24	0.36	11.12	9.04	4.00	1.26	14.00	12.06	5.96	2.32
Bootstrap method	2.85	2.79	2.70	2.74	2.87	2.82	2.80	2.81	2.94	2.96	2.87	3.04

Clustered data simulation results under Ha condition are presented in the following table. We take ρ=0.2 and λ=0.6 as an example and investigate the achieved power though simulation. We let $\boldsymbol{X}^{(1)} = \boldsymbol{X}^{(1)} + \delta$ under the alternative hypothesis (Ha) condition, δ = {0.1, 0.2, 0.3, 0.4}.

Table 5: Observed Powers at Different δ (ρ=0.2, λ=0.6)

δ	Z-test (ignore correlations) Power (%)	Bootstrap Method Power (%)
0.1	11.6 %	17.6 %
0.2	42.9 %	53.0 %
0.3	79.2 %	85.3 %
0.4	96.1 %	97.4 %

4. Summary

We compared various methods for correlated/clustered diagnostic performance comparison. It appears that ignoring the between-test correlation causes conservativeness in type I error rate control at the cost of losing power under Ha condition. And ignoring within-test correlation might cause type I error rate inflation. We can apply an appropriate statistical method, such as a non-parametric method based on re-sampling (Bootstrap), Z-test that accounts for the correlation, or Bayesian method incorporating reliable information, for this kind of problem. We can then achieve powerful statistical method while still control type I error rate at the nominal α level.

References

Emir, B., S. Wieand, S.-H. Jung, and Z. Ying (2000). Comparison of diagnostic markers with repeated measurements: a non-parametric roc curve approach. Statistics in Medicine 19 (4).

Ye, Xuan, and L. Larry Tang. "Group Sequential Methods for Comparing Correlated Receiver Operating Characteristic Curves." Applied Statistics in Biomedicine and Clinical Trials Design. Springer, Cham, 2015. 89-108.

Efron, Bradley, and Robert J. Tibshirani. An introduction to the bootstrap. CRC press, 1994.

Gelman, Andrew, et al. Bayesian data analysis. Chapman and Hall/CRC, 2013.

Pepe, Margaret Sullivan. The statistical evaluation of medical tests for classification and prediction. Medicine, 2003.

Johnson, Richard A., and Dean W. Wichern. "Applied Multivariate Statistical Analysis, Printice-Hall." Inc., Upper Saddle River, NJ (1998).