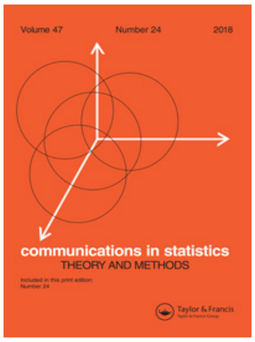

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Optimal nonparametric estimator of the area under ROC curve based on clustered data

Yougui Wu

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2018.1563176

Optimal nonparametric estimator of the area under ROC curve based on clustered data

Yougui Wu

Department of Epidemiology and Biostatistics College of Public Health, University of South Florida, Tampa, Florida, USA

ABSTRACT

In diagnostic trials, clustered data are obtained when several subunits of the same patient are observed. Intracluster correlations need to be taken into account when analyzing such clustered data. A nonparametric method has been proposed by Obuchowski (1997) to estimate the Receiver Operating Characteristic curve area (AUC) for such clustered data. However, Obuchowski's estimator is not efficient as it gives equal weight to all pairwise rankings within and between cluster. In this paper, we propose a more efficient nonparametric AUC estimator with two sets of optimal weights. Simulation results show that the loss of efficiency of Obuchowski's estimator for a single AUC or the AUC difference can be substantial when there is a moderate intracluster test correlation and the cluster size is large. The efficiency gain of our weighted AUC estimator for a single AUC or the AUC difference is further illustrated using the data from a study of screening tests for neonatal hearing.

ARTICLE HISTORY
Received 25 July 2018
Accepted 10 December 2018

KEYWORDS
Clustered data; diagnostic test; area under ROC curve; weighted nonparametric AUC; optimal weights; relative efficiency

1. Introduction

The clustered data are multiple measurements of a diagnostic test from a single subject (or cluster). Such type of data often occur in diagnostic trials where the multiple measurements of a diagnostic test are obtained from several subunits of the same patient, for example, two measurements of the distortion product otoacoustic emissions (DPOAE) were taken from both left and right eyes from each patient (Norton et al. 2000). In such cases, the multiple measurements from the same subject are naturally correlated to each other. The intracluster correlation must be considered when estimating and drawing inferences about the accuracy of the diagnostic test.

When the diagnostic test is continuous or ordinal, the area under the ROC curve is often used as a global measure of the accuracy of the test. An ROC curve is a plot of a test's sensitivity versus 1-specificity. The curve is constructed by changing the cut-point that defines a positive test result. The area under the ROC curve (AUC) summarizes the test's overall diagnostic ability and it is preferred over simple estimates of sensitivity or specificity because AUC incorporates both of these measures of accuracy and accounts for the inherent tradeoffs between them as the decision criterion changes (Metz 1978).

CONTACT Yougui Wu ywu@health.usf.edu

Obuchowski (1997) proposed a nonparametric AUC estimator that can be used for the general clustered data where there are arbitrary numbers of unaffected and affected units in a cluster. Obuchowski derived an asymptotic variance estimator for the AUC estimator, taking into account of intracluster correlations. However, Obuchowski's AUC estimator is not efficient as it gives equal weight to all pairwise rankings within and between clusters. Clusters can be different in terms of cluster size, the number of unaffected units, and the number of affected units. In the presence of various intracluster correlations, these differences would affect the contribution of a cluster to the overall variance of the AUC estimator and hence weights should vary across clusters.

In the present paper, we first propose a nonparametric weighted AUC estimator that assigns an arbitrary weight to each unaffected unit as well as an arbitrary weight to each affected unit in a cluster. Our proposed weighted AUC estimator is the extension of its counterpart proposed by Emir et al. (2000) in that our proposed weighted AUC estimator can be applied to the most general clustered data where there are arbitrary numbers of unaffected and affected units in a cluster, and the test results can be continuous or ordinal, whereas Emir et al.'s AUC estimator can be only applied to the continuous clustered data with each cluster having at most one affected unit.

We then propose an optimal weighting scheme that minimizes the variance of our proposed AUC estimator. Our proposed optimal weighting scheme assigns weights to unaffected units depending on the number of unaffected units and the intracluster correlation among unaffected units, and similarly, assign weights to affected units depending on the the number of affected units and the intracluster correlation among affected units. For the clusters that has both unaffected units and affected units, our proposed optimal weighting scheme also takes into account of the intracluster correlation between unaffected units and affected units. The optimal weighting scheme considered by Wu et al. (2011) was based on Emir et al.'s AUC estimator, and thus it can not be applied to the continuous clustered data where there are multiple affected units in a cluster. Therefore, our proposed optimal weighting scheme includes Wu et al.'s as a special case.

The rest of this paper is organized as follows. In Section 2, our weighted AUC estimator is presented and its asymptotic normality is established under some mild boundary conditions. In Section 3, the two sets of optimal weights which minimize the variance of our proposed weighted AUC estimator are derived, and the asymptotic normality of our weighted AUC estimator with estimated optimal weights are presented. In Section 4, the optimal weighted estimate of the AUC difference and its asymptotic normality are discussed. the finite sample performance of our asymptotic normality results are assessed in Section 5. A data example is presented in Section 6 and conclusions are provided in Section 7.

2. Weighted estimator of a single ROC curve area

2.1. Notation

Assume that there are n clusters of units sampled from the underline cluster population. Let X_{ik} denote the test result of the kth unaffected unit in the ith cluster $(j = 1, 2, \ldots, m_{0i}), (i = 1, 2, \ldots, n)$. Similarly, let Y_{ik} denote the test result of the kth affected unit in the ith cluster $(k = 1, 2, ..., m_{1i})$. These test results can be continuous or ordered discrete random variables. Assume further that test results from different subjects are independent and measurements within a subject are possibly correlated.

Let $M_0 = \sum_{i=1}^{n} m_{0i}$ be the total number of unaffected units and $M_1 = \sum_{i=1}^{n} m_{1i}$ be the total number of affected units. Both m_{0i} and m_{1i} can be 0 ($m_{0i} = 0$ when the units in the ith cluster are all affected; $m_{1i} = 0$ when the units in the ith cluster are all unaffected). Let δ_{0i} be the indicator of whether the ith cluster has at least one unaffected unit or not, i.e., δ_{0i} takes value 1 if $m_{0i} \neq 0$, and 0 otherwise. Similarly, let δ_{1i} be the indicator of whether the ith cluster has at least one affected unit or not, i.e., δ_{1i} takes value 1 if $m_{1i} \neq 0$, and 0 otherwise.

2.2. Weighted estimator of a single ROC curve area using arbitrary weights

Bamber (1975) has shown that, under the trapezoidal rule, the area under the ROC curve, θ, is equal to $\Pr(X<Y) + \frac{1}{2}\Pr(X=Y)$. Note that θ can be expressed as

$$\theta = E\left(\frac{1}{2}\left[I(X \leq Y) + I((X < Y)\right] = \mathrm{E}\bar{F}(Y) = \int \bar{F}(y) d\bar{G}(y), \tag{1}$$

where $I(\cdot)$ is the indicator function, $\bar{F}(x)$ is the normalized distribution function of X_{ik} and $\bar{G}(y)$ is the normalized distribution function of Y_{ik} (Edgar, Ullrich, and Madan 2002). Specifically, $\bar{F}(x)$ and $\bar{G}(y)$ are defined as

$$\bar{F}(x) = \frac{1}{2}\left[\Pr(X \leq x) + \Pr(X<x)\right]$$

and

$$\bar{G}(y) = \frac{1}{2}[\Pr(Y \leq y) + \Pr(Y<y)].$$

Obuchowski (1997) proposed a nonparamtric AUC estimate based on clustered data:

$$\widehat{\theta} = \frac{1}{M_0 M_1} \sum_{i=1}^{n} \sum_{i'=1}^{n} \sum_{k=1}^{m_{0i}} \sum_{k'=1}^{m_{1i'}} \psi(X_{ik}, Y_{i'k'}). \tag{2}$$

where $\psi(X, Y) = \frac{1}{2}[I(X \leq Y) + I(X < Y)]$. Note that Obuchowski's AUC estimator can be written as

$$\widehat{\theta} = \int \widehat{\bar{F}}(y) d\widehat{\bar{G}}(y),$$

where $\widehat{\bar{F}}(x)$ and $\widehat{\bar{G}}(y)$ are the empirical versions of $\bar{F}(x)$ and $\bar{G}(y)$, respectively, i.e.,

$$\begin{aligned} \widehat{\bar{F}}(x) &= \frac{1}{M_0} \sum_{i=1}^{n} \sum_{k=1}^{m_{0i}} \psi(X_{ik}, x) \\ \widehat{\bar{G}}(y) &= \frac{1}{M_1} \sum_{i=1}^{n} \sum_{k=1}^{m_{1i}} \psi(Y_{ik}, y). \end{aligned} \tag{3}$$

Therefore, Obuchowski's AUC estimator (Equation 2) can be viewed as the result of giving equal weight to unaffected units in estimating $\bar{F}(x)$ and to affected units in estimating $\bar{G}(y)$. Clearly, this would be the right way to weight the observations if the observations were independent within clusters. However, other weighting schemes might be preferable when observations are correlated within clusters and when cluster sizes are different.

We are now ready to define a generalized version of (Equation 2). Let w_{0i} be the non-negative weight assigned to each unaffected unit in the ith cluster, where $w_{0i} = 0$ if $\delta_{0i} = 0$ and $\sum_{i=1}^{n} \delta_{0i} m_{0i} w_{0i} = M_0$. Similarly, let w_{1i} be the non-negative weight assigned to each affected unit in the ith cluster, where $w_{1i} = 0$ if $\delta_{1i} = 0$ and $\sum_{i=1}^{n} \delta_{1i} m_{1i} w_{1i} = M_1$. We propose to estimate $\bar{F}(x)$ by assigning an arbitrary weight w_{0i} to each unaffected unit in the ith cluster:

$$\widehat{\bar{F}}(x) = \frac{1}{M_0} \sum_{i=1}^{n} w_{0i} \left\{ \sum_{k=1}^{m_{0i}} \psi(X_{ik}, x) \right\},$$

and estimate $\bar{G}(y)$ by assigning an arbitrary weight w_{1i} to each affected unit in the ith cluster:

$$\widehat{\bar{G}}(y) = \frac{1}{M_1} \sum_{i=1}^{n} w_{1i} \left\{ \sum_{k=1}^{m_{1i}} \psi(Y_{ik}, y) \right\}.$$

The weighted nonparametric estimator of θ is then obtained by plugging the above two weighted empirical distribution functions into (Equation 1):

$$\widehat{\theta}_w = \int \widehat{\bar{F}}(t) d\widehat{\bar{G}}(t) = \frac{1}{M_0 M_1} \sum_{i=1}^{n} \sum_{i'=1}^{n} w_{0i} w_{1i'} \sum_{k=1}^{m_{0i}} \sum_{k'=1}^{m_{1i'}} \psi(X_{ik}, Y_{i'k'}), \tag{4}$$

where the weight $w_{0i} w_{1i}$ is attached to all possible pairs of ranking. In general, $w_{0i} = \phi_0(m_{0i})$ is a function of m_{0i}, and $w_{1i} = \phi_1(m_{1i})$ is a function of m_{1i}. The special choice $w_{0i} = 1, w_{1i} = 1$ gives unit weight to each individual observation and the resulting estimator is equivalent to Obuchowski's AUC estimator (Equation 2).

To present our asymptotic normality results of our proposed weighted AUC estimator, we now extend the Obuchowski's notations of the X- and Y-components for clustered data to their weighted versions. Specifically, the weighted X-components are defined as

$$V_{10}(X_{ik}) = \frac{1}{M_1} \sum_{i'=1}^{n} w_{1i} \sum_{k'=1}^{m_{1i'}} \psi(X_{ik}, Y_{i'k'})$$

for all X_{ik}, and the weighted Y-components are defined as

$$V_{01}(Y_{i'k'}) = \frac{1}{M_0} \sum_{i=1}^{n} w_{0i} \sum_{k=1}^{m_{0i}} \psi(X_{ik}, Y_{i'k'})$$

for all $Y_{i'k'}$. Let $V_{10}(X_{i.})$ and $V_{01}(Y_{i.})$ be the sum of the weighted X- and Y-components, respectively, for the ith cluster. Then, the we can state the asymptotic normality of our proposed weighted AUC estimator using the notations of the weighted X- and Y-components as in the following theorem:

Theorem 2.1. *Assume that cluster size is bounded by a finite number,* $w_{0i} = \phi_0(m_{0i})$, *and* $w_{1i} = \phi_1(m_{1i}), i = 1, ..., n$, *where* $\phi_0(\cdot)$ *and* $\phi_1(\cdot)$ *are bounded non-negative functions satisfying* $\phi_0(0) = \phi_1(0) = 0$. *Then*

$$\frac{\widehat{\theta}_w - \theta}{\sqrt{\widehat{var}\left(\widehat{\theta}_w\right)}} \xrightarrow{d} N(0, 1)$$

where

$$\widehat{var}\left(\widehat{\theta}_w\right) = \frac{1}{M_0} S_{10} + \frac{1}{M_1} S_{01} + \frac{2}{M_0 M_1} S_{11}, \tag{5}$$

with

$$S_{10} = \frac{1}{M_0} \sum_{i=1}^{n} w_{0i}^2 \left[V_{10}(X_{i.}) - m_{0i}\widehat{\theta}\right]^2, \quad S_{01} = \frac{1}{M_1} \sum_{i=1}^{n} w_{1i}^2 \left[V_{01}(Y_{i.}) - m_{1i}\widehat{\theta}\right]^2$$

$$S_{11} = \sum_{i=1}^{n} w_{0i} w_{1i} \left[V_{10}(X_{i.}) - m_{0i}\widehat{\theta}\right] \left[V_{01}(Y_{i.}) - m_{1i}\widehat{\theta}\right].$$

The proof of this theorem can be found in Appendix. The asymptotic variance expression (Equation 5) in the special case of $w_{0i} = w_{1i} = 1$ is identical to Obuchowski's. However, it is more general than Obuchowski's as it can be applied to other choices of weighting scheme, for example, our proposed optimal weighting scheme to be discussed in next section.

2.3. Weighted estimator of a single ROC curve area using optimal weights

The optimal weights problem that only involves one set of weights, w_{0i}, was considered in Wu et al. (2011) when each cluster has at most one affected unit and w_{1i} is fixed $(= \delta_{1i})$. In this section, we consider the optimal weights problem that involves two sets of weights, $(w_{0i}, i = 1, ..., n)$ and $(w_{1i}, i = 1, ..., n)$. To derive the optimal weights, we assume that the correlation coefficients within unaffected units, within affected units, and between unaffected units and affected units, are constant. Specifically, we define the following two transformations:

$$U_{ik} = \bar{G}(X_{ik}), \quad V_{ik} = \bar{F}(Y_{ik}). \tag{6}$$

and assume that $\text{Corr}(U_{ik}, U_{i,k'}) = \rho_U, \text{Corr}(V_{ik}, V_{i,k'}) = \rho_V$, for $k \neq k'$, and $\text{Corr}(U_{ik}, V_{i,k'}) = \rho_{UV}$. Let $\sigma_U^2 = \text{Var}(U_{ik})$ and $\sigma_U^2 = \text{Var}(U_{ik})$. Then using the fact established in the proof of Theorem 2.1 in the Appendix, we have

$$\widehat{\theta}_w - \theta = \frac{1}{M_1} \sum_{i=1}^{n} w_{1i} \sum_{k=1}^{n_i} (V_{ik} - \theta) - \frac{1}{M_0} \sum_{i=1}^{n} w_{0i} \sum_{k=1}^{m_i} \left(U_{ik} - (1 - \theta)\right) + o_p\left(n^{-1/2}\right).$$

and we can approximate the variance of $\widehat{\theta}_w$ given m_{0i} and $m_{1i}, i = 1, ..., n$ in terms of the weights as

$$\text{Var}\left(\widehat{\theta}_w | m_{0i}, m_{1i}, i = 1, ..., n\right) = \sum_{i=1}^{n} \left(a_i w_{0i}^2 - 2b_i w_{0i} w_{1i} + c_i w_{1i}^2\right) \tag{7}$$

where $a_i = \sigma_U^2(m_{0i} + m_{0i}(m_{0i}-1)\rho_U)/M_0^2$, $b_i = m_{0i}m_{1i}\sigma_U\sigma_V\rho_{UV}/M_0M_1$, and $c_i = \sigma_V^2(m_{1i} + m_{1i}(m_{1i}-1)\rho_V)/M_1^2$. The optimal weights of w_{0i} and w_{1i} can be found by minimizing (Equation 7) with respect to w_{0i} and w_{1i} with constraints $w_{0i} \geq 0, w_{1i} \geq 0, i = 1, \ldots, n$, $\sum_{i=1}^{n} \delta_{0i}w_{0i}m_{0i} = M_0$, and $\sum_{i=1}^{n} \delta_{1i}w_{1i}m_{1i} = M_1$. In the Appendix, we show that the optimal weights can be expressed as

$$
\begin{aligned}
w_{0i} &= \delta_{0i}(1-\delta_{1i})\frac{m_{0i}\mu_0}{a_i} + \delta_{0i}\delta_{1i}\left[\frac{c_i m_{0i}}{a_i c_i - b_i^2}\mu_0 + \frac{b_i m_{1i}}{a_i c_i - b_i^2}\mu_1\right], \\
w_{1i} &= \delta_{1i}(1-\delta_{0i})\frac{m_{1i}\mu_1}{c_i} + \delta_{0i}\delta_{1i}\left[\frac{b_i m_{0i}}{a_i c_i - b_i^2}\mu_0 + \frac{a_i m_{1i}}{a_i c_i - b_i^2}\mu_1\right],
\end{aligned}
\tag{8}
$$

where

$$
\begin{aligned}
\mu_0 &= \frac{CM_0 - BM_1}{\Delta}, \quad \mu_1 = \frac{AM_1 - BM_0}{\Delta}, \quad \Delta = AC - B^2, \\
A &= \sum_{i=1}^{n}\delta_{0i}(1-\delta_{1i})m_{0i}^2 a_i^{-1} + \sum_{i=1}^{n}\delta_{0i}\delta_{1i}\frac{m_{0i}^2 c_i}{a_i c_i - b_i^2}, \\
B &= \sum_{i=1}^{n}\delta_{0i}\delta_{1i}\frac{m_{0i}m_{1i}b_i}{a_i c_i - b_i^2}, \\
C &= \sum_{i=1}^{n}(1-\delta_{0i})\delta_{1i}m_{1i}^2 c_i^{-1} + \sum_{i=1}^{n}\delta_{0i}\delta_{1i}\frac{m_{1i}^2 a_i}{a_i c_i - b_i^2}.
\end{aligned}
$$

One of the special cases of (Equation 8) is when the intracluster correlation between unaffected units and affected units is 0, i.e., $\rho_{UV} = 0$. In this special case, the optimal weights (Equation 8) has a simpler expressions, given by

$$
\begin{aligned}
w_{0i} &= \frac{M_0\delta_{0i}(1 + (m_{0i}-1)\rho_U))^{-1}}{\sum_{i'=1}^{n}\delta_{0i}m_{0i'}(1+(m_{0i'}-1)\rho_U))^{-1}}, \\
w_{1i} &= \frac{M_1\delta_{1i}(1 + (m_{1i}-1)\rho_V))^{-1}}{\sum_{i'=1}^{n}\delta_{1i}m_{1i'}(1 + (m_{1i'}-1)\rho_V))^{-1}}.
\end{aligned}
\tag{9}
$$

Note that the expression for w_{0i} in (Equation 9) is the same as that obtained by minimizing the first term of (Equation 7) with respect to w_{0i} with constraints $w_{0i} \geq 0, i = 1, \ldots, n$, and $\sum_{i=1}^{n} \delta_{0i}w_{0i}m_{0i} = M_0$. Similarly, the expression for w_{1i} in (Equation 9) is the same as that obtained by minimizing the third term of (Equation 7) with respect to w_{1i} with constraints $w_{1i} \geq 0, i = 1, \ldots, n$, and $\sum_{i=1}^{n} \delta_{1i}w_{1i}m_{1i} = M_1$. This can be explained by the fact that the interaction term of w_{0i} and w_{1i} in (Equation 7) disappears when $\rho_{UV} = 0$, and, as a result, solving the joint minimization problem with respect to two sets of weights is equivalent to solving two separate minimization problems with respect to each set of weights.

The other special case of (Equation 8) is when all correlations within cluster are equal to 0, i.e., $\rho_U = \rho_V = \rho_{UV} = 0$. In this special case, the optimal weights (Equation 8) reduces to $w_{0i} = 1$ and $w_{1i} = 1$, which means that the weights involved in Obuchowski's estimator is optimal only when the test results within the same cluster are independent.

The optimal weights involve unknown parameters, $\sigma_U^2, \sigma_V^2, \rho_U, \rho_V$ and $\rho_{UV,}$ which can be estimated by first obtaining the estimated transformed data, $\widehat{U}_{ik} = \widehat{\widehat{G}}(X_{ik})$ and $\widehat{V}_{ik} = \widehat{\widehat{F}}(Y_{ik})$, and then using the estimated transformed data to obtain the second moment estimates of those unknown parameters:

$$\widehat{\sigma}_U^2 = \frac{\sum_{i,k} \delta_{0ik}\left(\widehat{U}_{ik} - \bar{U}_k\right)^2}{\sum_{i,k} \delta_{0ik}}, \quad \widehat{\sigma}_V^2 = \frac{\sum_{i,k} \delta_{1ik}\left(\widehat{V}_{ik} - \bar{V}_k\right)^2}{\sum_{i,k} \delta_{1ik}},$$
$$\widehat{\rho}_{UV} = \frac{\sum_{i,k,k'} \delta_{0ik}\delta_{1ik'}\left(\widehat{U}_{ik} - \bar{U}_k\right)\left(\widehat{V}_{ik'} - \bar{V}_{k'}\right)}{\widehat{\sigma}_U\widehat{\sigma}_V \sum_{i,k,k'} \delta_{0ik}\delta_{1ik'}}, \quad \widehat{\rho}_U = \frac{\sum_{i,k,k'} \delta_{0ik}\delta_{0ik'}\left(\widehat{U}_{ik} - \bar{U}_k\right)\left(\widehat{U}_{ik'} - \bar{U}_{k'}\right)}{\widehat{\sigma}_U^2 \sum_{i,k,k'} \delta_{0ik}\delta_{0ik'}},$$
$$\widehat{\rho}_V = \frac{\sum_{i,k,k'} \delta_{1ik}\delta_{1ik'}\left(\widehat{V}_{ik} - \bar{V}_k\right)\left(\widehat{V}_{ik'} - \bar{V}_{k'}\right)}{\widehat{\sigma}_V^2 \sum_{i,k,k'} \delta_{1ik}\delta_{1ik'}}, \tag{10}$$

where $\delta_{0jk} = 1$ if the jth cluster has the kth unaffected unit and $=0$ otherwise, $\delta_{1jk} = 1$ if the jth cluster has the kth affected unit and $=0$ otherwise, $\bar{U}_k = \sum_{j=1}^n \delta_{0jk}\widehat{U}_{jk} / \sum_{j=1}^n \delta_{0jk}$ and $\bar{V}_k = \sum_{j=1}^n \delta_{1jk}\widehat{V}_{jk} / \sum_{j=1}^n \delta_{1jk}$.

Let $(\widehat{w}_{0i}, \widehat{w}_{1i})$ be the estimated optimal weights obtained by replacing $\sigma_U^2, \sigma_V^2, \rho_U, \rho_V$ and ρ_{UV} in (Equation 8) by their corresponding estimates given by (Equation 10), and $\widehat{\theta}_{op}$ be the estimate of θ using the estimated optimal weights $(\widehat{w}_{0i}, \widehat{w}_{1i})$. Then the following theorem establishes the asymptotic normality of $\widehat{\theta}_{op}$.

Theorem 2.2. *Suppose that the conditions stated in Theorem 2.1 are satisfied. Then it follows that* $(\widehat{\theta}_{op} - \theta)/\sqrt{\widehat{var}(\widehat{\theta}_{op})}$ *follows an asymptotically standard normal distribution, where* $\widehat{var}(\widehat{\theta}_{op})$ *is obtained by (Equation 5) using the estimated optimal weights.*

The proof of this theorem can be found in Appendix.

3. Weighted estimator of the difference of two correlated ROC curve areas using optimal weights

When comparing two ROC curve areas estimated from the same sample of clustered data, the covariance between the two estimated areas must be taken into account. Obuchowski (1997) proposed an estimator for the covariance between the two estimated areas AUC estimators. In this section, we extend Obuchowski's estimator to accommodate the case where the optimal weights are applied to estimating each AUC.

Let X_{ik}^r denote the rth diagnostic test result of the jth unaffected unit in the ith cluster($r = 1, 2, k = 1, ..., m_{0i}, i = 1, ..., n$). Similarly, let Y_{ik}^r denote the rth diagnostic test result of the jth affected unit in the ith cluster($r = 1, 2, k = 1, ..., m_{1i}, i = 1, ..., n$). Let θ^r be the AUC of the rth diagnostic test, $(\widehat{w}_{0i}^r, \widehat{w}_{1i}^r)$ be the estimated optimal weights using the rth diagnostic test data (X_{ij}^r, Y_{ij}^r). Let $\widehat{\theta}_{op}^r (r = 1, 2)$ be the estimator of θ^r using the optimal weights, $(\widehat{w}_{0i}^r, \widehat{w}_{1i}^r)$. Define

$$S_{10}^{1,2} = \frac{1}{M_0}\sum_{i=1}^{n}\widehat{w}_{0i}^{1}\widehat{w}_{0i}^{2}\left[V_{10}^{1}(X_{i.}) - m_{0i}\widehat{\theta}_{op}^{1})\right]\left[V_{10}^{2}(Y_{i.}) - m_{0i}\widehat{\theta}_{op}^{2})\right]$$
$$S_{01}^{1,2} = \frac{1}{M_1}\sum_{i=1}^{n}\widehat{w}_{1i}^{1}\widehat{w}_{1i}^{2}\left[V_{01}^{1}(Y_{i.}) - m_{1i}\widehat{\theta}_{op}^{1})\right]\left[V_{01}^{2}(Y_{i.}) - m_{1i}\widehat{\theta}_{op}^{2})\right]$$
$$S_{11}^{1,2} = \sum_{i=1}^{n}\widehat{w}_{0i}^{1}\widehat{w}_{1i}^{2}\left[V_{10}^{1}(X_{i.}) - m_{0i}\widehat{\theta}_{op}^{1})\right]\left[V_{01}^{2}(Y_{i.}) - m_{1i}\widehat{\theta}_{op}^{2}\right]$$
$$S_{11}^{2,1} = \sum_{i=1}^{n}\widehat{w}_{0i}^{2}\widehat{w}_{1i}^{1}\left[V_{10}^{2}(X_{i.}) - m_{0i}\widehat{\theta}_{op}^{2})\right]\left[V_{01}^{1}(Y_{i.}) - m_{1i}\widehat{\theta}_{op}^{1})\right]$$

where $V_{10}^{r}(X_{i.})$ is the sum of the X-components in cluster i from the rth ROC curve, and $V_{01}^{r}(Y_{i.})$ is the sum of the Y-components in cluster i from the rth ROC curve. The asymptotic normality of $\widehat{\theta}_{op}^{1} - \widehat{\theta}_{op}^{2}$ is established in the following theorem:

Theorem 3.1. *Suppose that the conditions stated in Theorem 2.1 are satisfied for both diagnostic tests. Then it follows that*

$$\frac{\widehat{\theta}_{op}^{1} - \widehat{\theta}_{op}^{2} - (\theta^{1} - \theta^{2})}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\theta}_{op}^{1} - \widehat{\theta}_{op}^{2}\right))}} \to N(0,1)$$

where

$$\widehat{\mathrm{Var}}\left(\widehat{\theta}_{op}^{1} - \widehat{\theta}_{op}^{2}\right)) = \widehat{\mathrm{Var}}\left(\widehat{\theta}_{op}^{1}\right) + \widehat{\mathrm{Var}}\left(\widehat{\theta}_{op}^{2}\right) - 2\widehat{\mathrm{Cov}}\left(\widehat{\theta}_{op}^{1}, \widehat{\theta}_{op}^{2}\right), \tag{11}$$

with $\widehat{\mathrm{Var}}(\widehat{\theta}_{op}^{r}), r = 1,2$ given by Equation (5), and $\widehat{\mathrm{Cov}}(\widehat{\theta}_{op}^{1}, \widehat{\theta}_{op}^{2})$ given by

$$\widehat{\mathrm{Cov}}\left(\widehat{\theta}_{op}^{1}, \widehat{\theta}_{op}^{2}\right) = \frac{S_{10}^{1,2}}{M_0} + \frac{S_{01}^{1,2}}{M_1} + \frac{S_{11}^{1,2}}{M_0 M_1} + \frac{S_{11}^{2,1}}{M_0 M_1}$$

The proof of this theorem can be found in Appendix.

4. Monte Carlo simulation study

We conducted a Monte Carlo simulation study to assess the finite sample performance of the asymptotic results as stated in Theorem 2.2 for the optimal weighted AUC estimator and Theorem 3.1 for the optimal weighted AUC difference estimator. Specifically, the coverage of 95% confidence intervals derived from Theorem 2.2 were assessed for the AUC of one single diagnostic test, and the size and power of Wald tests derived from Theorem 3.1 were assessed for comparing the AUCs of two diagnostic tests.

The clustered data for an continuous test were generated as follows. The cluster size $m_i = m_{0i} + m_{1i}$ for ith cluster was drawn from a cluster size distribution. Given m_i, we randomly generate disease indicators $d_{ij}, j = 1, \ldots, m_i$ for each cluster by first drawing a multivariate normal random vector from m_i-variate normal distribution with mean vector zero and a compound symmetry covariance matrix with intracluster correlation ρ_D, and then turning each random variable into a binary indicator using a percentile corresponding to a specified disease prevalence. Test results for units within each cluster were then generated from a second m_i-variate normal distribution with mean vector zero and a compound symmetry covariance matrix with intracluster correlation ρ_T. A constant, d, was added to the test results of units identified as affected, where d was chosen in order to make the area under the ROC curve equal to the corresponding AUC value. To generate data for an ordinal test on a 1-5 scale, we simulated continuous clustered data as described above, and then the regroup all realized values into 5 groups using q% percentile, as cutting point, where $q = 20, 40, 60, 80$.

The parameter settings are designed as follows. The number of clusters, the disease prevalence is fixed at 0.5, and the disease intracluster correlation were fixed at $100, 0.5$, and 0.5, respectively. The simulation results are similar under other different values, and, for saving space, were not presented here. The parameters that vary were the cluster size, and the intracluster correlation of test within a cluster, ρ_T. The values for cluster size were 2, 4 and 10, and the values used for ρ_T are $0, 0.4$, and 0.8.

Under each parameter setting, we simulated 2000 data sets and the optimal weighted AUC estimates were calculated from each data set. The optimal weights were obtained by implementing a four-step procedure: Step 1: Use weights $w_{0i} = w_{1i} = 1$ to (Equation 3) to obtain the empirical distributions: $\widehat{\bar{F}}$ and $\widehat{\bar{G}}$; Step 2: replace $\bar{F}$ and $\bar{G}$ in (Equation 6) by their empirical counterparts to obtain the transformed data $(\widehat{U}_{jk}, \widehat{V}_{jk})$; Step 3: use the transformed data $(\widehat{U}_{jk}, \widehat{V}_{jk})$ and Equation (10) to obtain the estimates of $\sigma_U^2, \sigma_V^2, \rho_U, \rho_V$ and $\rho_{UV;}$ Step 4: plug the estimates of $\sigma_U^2, \sigma_V^2, \rho_U, \rho_V$ and ρ_{UV} into Equation (8) to obtain the optimal weight estimates. The estimated optimal weights were then plugged into Equation (4) to calculate the point estimate of the AUC.

For the single ROC curve, the coverage of the 95% CIs for the true area was assessed. The true AUC values considered were 0.5 and 0.7 for continuous test results and 0.7 for ordinal test results. The 95% confidence interval were constructed from the variance estimator given in Equation (5). Table 1 summarizes the estimated standard error from (Equation 5), the empirical standard error, and the coverage of the 95% confidence interval. The results were compared to those obtained when Obuchowski's estimators were used. To facilitate the comparison, the ratio of the empirical variance of our proposed optimal weighted AUC estimator to that of Obuchowski's estimator was also included in Table 1. It can be seen that the asymptotic 95% confidence interval based on (Equation 5) provides appropriate coverage for the true AUC for various cluster sizes, types of test results, a wide range of correlations between test results. In addition, our proposed optimal weighted estimators are more efficient than their counterparts of Obuchowski's estimators with larger efficiency gain as cluster size and intracluster correlation increase.

For two correlated ROC curves, the size and power of the Wald test comparing the areas and the coverage of the 95% confidence interval for the difference in areas were assessed. The Wald tests and CIs were constructed from the estimated variance of the difference, given in Equation (11). When assessing the size of the Wald tests, the true areas under the two curves were 0.7; a significance level of 0.05 was used. When assessing the power of the Wald tests and the coverage of the 95% confidence interval, the true areas under the two curves were 0.7 and 0.8. The correlations between the results of the two diagnostic tests are 0 and 0.5 for continuous test results and 0.5 ordinal test

Table 1. Standard error and Coverage rate of 95% confidence interval of a single AUC derived from (Equation 5).

			Obuchowski's estimator				Proposed estimator				
AUC	CS[a]	ρ_T^b	Est[c]	ASE[d]	SE[e]	Coverage	Est	ASE	SE	Coverage	RE[f]
					Normal test results						
0.5	2	0.0	0.500	0.041	0.041	0.948	0.500	0.041	0.041	0.942	1.02
		0.4	0.501	0.043	0.045	0.942	0.501	0.04	0.042	0.939	0.88
		0.8	0.500	0.046	0.045	0.944	0.501	0.031	0.031	0.950	0.47
	4	0.0	0.500	0.029	0.029	0.944	0.500	0.029	0.029	0.944	1.01
		0.4	0.500	0.034	0.035	0.942	0.500	0.028	0.030	0.944	0.71
		0.8	0.501	0.039	0.039	0.944	0.500	0.019	0.019	0.954	0.25
	10	0.0	0.501	0.018	0.018	0.947	0.501	0.018	0.018	0.948	1.02
		0.4	0.500	0.027	0.028	0.942	0.501	0.018	0.018	0.940	0.44
		0.8	0.500	0.034	0.033	0.953	0.500	0.012	0.011	0.958	0.12
0.7	2	0.0	0.701	0.037	0.036	0.948	0.700	0.036	0.037	0.947	1.02
		0.4	0.700	0.039	0.039	0.946	0.701	0.036	0.037	0.938	0.91
		0.8	0.701	0.041	0.043	0.940	0.705	0.030	0.032	0.930	0.56
	4	0.0	0.701	0.026	0.026	0.948	0.701	0.026	0.026	0.942	1.02
		0.4	0.700	0.030	0.031	0.946	0.701	0.026	0.026	0.947	0.73
		0.8	0.701	0.035	0.036	0.938	0.703	0.021	0.022	0.951	0.36
	10	0.0	0.700	0.016	0.017	0.944	0.700	0.016	0.017	0.940	1.03
		0.4	0.701	0.024	0.024	0.932	0.702	0.017	0.018	0.944	0.52
		0.8	0.702	0.031	0.030	0.945	0.703	0.016	0.016	0.948	0.26
					Ordinal test results						
0.7	4	0.0	0.699	0.026	0.026	0.944	0.699	0.026	0.026	0.945	1.02
		0.4	0.700	0.030	0.031	0.944	0.700	0.026	0.027	0.938	0.75
		0.8	0.700	0.035	0.034	0.948	0.703	0.022	0.022	0.949	0.40

[a]Cluster size.
[b]Correlation of test results within the same cluster.
[c]Mean estimated AUCs.
[d]Mean estimated standard errors.
[e]Empirical standard deviation of the estimated AUC.
[f]The ratio of empirical variance of our proposed estimator to that of Obuchowski's estimator.

results. Table 2 summarizes the size and power of the Wald test, as well as the coverage of the 95% confidence interval of the AUC difference. It can be seen that the size of the Wald test constructed from the estimator in (Equation 11) is close to the nominal level, and the power is much higher than its counterpart based on Obuchowski's AUC difference estimator.

5. Date example

We applied our proposed weighted AUC estimators to data from the identification of neonatal hearing impairment (INHI) study, details of which have been previously published (Norton et al. 2000). Visual reinforcement audiometry (VRA) was used as the gold standard test to determine the true neonatal hearing loss. VRA is a behavioral test that cannot be administered until infants are between 8 and 12 months old. The objective of the INHI study was to assesse the accuracy of three passive electronic devices, the auditory brainstem response (ABR) the distortion product otoacoustic emissions (DPOAE) and the tran- sient evoked otoacoustic emissions (TEOAE) tests, administered soon after birth. For illustration purpose, we only selected both ABR test and DPOAE in our analysis.

The ABR and DPOAE tests use electronic devices that emit sounds in the ear canal and record continuous measures of the strength of response to these auditory stim- uli.

Table 2. Size and power of the Wald test for comparing two AUCs constructed from (Equation 11).

			Obuchowski's estimator			Proposed estimator			
γ[a]	CS	ρ	Size	Power	Coverage	Size	Power	Coverage	RE[b]
				Normal test results					
0.0	2	0.0	0.056	0.170	0.942	0.064	0.176	0.940	1.02
		0.4	0.053	0.160	0.934	0.055	0.190	0.926	0.91
		0.8	0.056	0.147	0.945	0.063	0.243	0.942	0.58
	4	0.0	0.054	0.302	0.952	0.055	0.304	0.948	1.02
		0.4	0.055	0.218	0.947	0.059	0.289	0.937	0.77
		0.8	0.055	0.180	0.946	0.053	0.362	0.939	0.42
	10	0.0	0.052	0.612	0.951	0.054	0.610	0.950	1.02
		0.4	0.062	0.332	0.948	0.058	0.544	0.944	0.52
		0.8	0.053	0.221	0.950	0.049	0.586	0.946	0.29
0.5	2	0.0	0.052	0.280	0.954	0.056	0.278	0.948	1.02
		0.4	0.056	0.266	0.940	0.064	0.292	0.930	0.95
		0.8	0.054	0.226	0.954	0.055	0.328	0.942	0.66
	4	0.0	0.049	0.500	0.947	0.051	0.492	0.946	1.03
		0.4	0.051	0.370	0.945	0.053	0.456	0.940	0.78
		0.8	0.051	0.304	0.938	0.048	0.522	0.945	0.46
	10	0.0	0.052	0.865	0.948	0.052	0.854	0.940	1.07
		0.4	0.058	0.560	0.948	0.058	0.804	0.938	0.56
		0.8	0.055	0.359	0.952	0.051	0.740	0.943	0.36
				Ordinal test results					
0.5	4	0.0	0.0506	0.478	0.944	0.060	0.474	0.944	1.01
		0.4	0.060	0.378	0.946	0.062	0.458	0.946	0.80
		0.8	0.054	0.284	0.950	0.052	0.486	0.948	0.52

[a]Correlation between two test results.
[b]The ratio of empirical variance of the proposed estimator of the AUC difference to that of Obuchowski's estimator.

ABR was performed using a brief click transmitted from an acoustic transducer. DPOE was performed using input sounds with different frequencies and intensities. We considered DPOAE tests at the frequency of 2000 MHz with stimulus intensity levels of 65 dB.

In this data example, there are 2742 infants(clusters), out of which 2614 infants had only unaffected ears, 73 patients had both unaffected and affected ears, and 55 patients had only affected ears. The ABR and DPOAE test results of the unaffected ears and affected ears were treated as the original data of X_{1ik} and Y_{1ik} and X_{2ik} and Y_{2ik}, respectively.

The estimated optimal weights for estimating the AUCs of ABR and DPOAE were obtained by implementing the four-step procedure described in simulation section. The estimated AUC values for ABR and DPOAE were obtained by plugging the estimated optimal weights into Equation (4). In obtaining variance estimates, (Equation 5) was used to calculate the variance estimates of the estimated AUCs for ABR and DPOAE, and Equation (11) was used to calculate the variance estimates of the estimated AUC difference.

Based on the data from this data example, the proposed estimates using optimal weights for ABR's AUC, DPOAE's AUC and the AUC difference were $0.594, 0.608$, and 0.014, respectively, and the corresponding variance estimates for these three estimators were 0.00056, 0.00055 and 0.00089. For comparison purpose, we also obtained Obuchowski's estimates for ABR's AUC, DPOAE's AUC, and the AUC difference as $0.608, 0.631$, and 0.023, respectively, and the corresponding variance estimates for these three estimators as $0.00071, 0.00066$ and 0.00118. It can be seen that our proposed optimal weighted estimators are more efficient than their counterparts of Obuchowski's estimators with significant efficiency gains. The efficiency gains were about 21% for ABR, 17% for DPOAE and 25% for the AUC difference.

6. Conclusions

In this paper, we extended the weighted AUC estimator considered by Emir et al. (2000) in the longitudinal study setting where the test results are continuous and each subject contribute at most one case observation, to the general clustered data setting where the test results can be continuous or ordinal, and clusters can have have multiple unaffected units as well as multiple affected units. Our weighted AUC estimator assigns two sets of arbitrary cluster weights, one for estimating specificity and the other for estimating sensitivity; Emir et al.'s weighted AUC estimator only involves one set of arbitrary cluster weights for estimating specificity. Unlike Emir et al.'s weighted AUC estimator, the asymptotic normality of our weighted AUC estimator is derived using the concept of normalized distribution function so that it holds true for both continuous and ordinal cases as long as the weights satisfy some mild boundary conditions. In addition, we also extend our AUC weighted estimator with only one set of optimal weights (Wu and Wang 2011) to the general clustered data where two sets of optimal weights are involved. Simulation results show that the loss of efficiency using equal weights assigned to units or clusters instead of our optimal weights can be severe when there is a moderate intracluster test correlation and the cluster size is large, in which case, our proposed AUC estimator with two sets of optimal weights is recommended.

Our proposed optimal weights requires to estimate covariance parameters based on the transformed data. The estimated covariance parameters might lead to negative optimal weights when the number of clusters is small. One alternative is to use the optimal weights given in (Equation 9), which always yields positive optimal weights regardless of the number of clusters even though the obtained weights are not optimal when the correlation between unaffected units and affected units within a cluster is not equal to 0. Our extensive simulation experience, however, suggests that negative optimal weights won't occur as long as the number of clusters is above 100.

References

Bamber, D. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 12 (4):387–415. doi:10.1016/0022-2496(75)90001-2.

Edgar, B., M. Ullrich, and L. P. Madan. 2002. The multivariate nonparametric Behrens-Fisher problem. *Journal of Statistical Planning and Inference* 108:37–53.

Emir, B., S. Wieand, S. Jung, and Z. Ying. 2000. Comparison of diagnostic markers with repeated measurements: A non-parametric ROC curve approach. *Statistics in Medicine* 19 (4):511–23.

Metz, C. E. 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 8 (4): 283–98.

Norton, S. J., M. P. Gorga, J. E. Widen, R. C. Folsom, Y. Sininger, B. Cone-Wesson, B. R. Vohr, K. Mascher, and K. Fletcher. 2000. Identification of neonatal hearing impairment: Evaluation

of transient evoked otoacoustic emission, distortion product otoacoustic emission, and auditory brain stem response test performance. *Ear and Hearing* 21 (5):508–28.
Obuchowski, N. A. 1997. Nonparametric analysis of clustered ROC curve data. *Biometrics* 53 (2): 567–78.
Wu, Y., and X. Wang. 2011. Optimal weight in estimating and comparing areas under the receiver operating characteristic curve using longitudinal data. *Biometrical Journal* 53 (5): 764–78. doi:10.1002/bimj.201100033.

Appendix

Proof of theorem 2.1. First we prove that $\widehat{\theta}_w - \theta$ is asymptotically equivalent to an independent sum:

$$\widehat{\theta}_w - \theta = \frac{1}{M_1}\sum_{i=1}^{n} w_{1i}\sum_{k=1}^{m_{1i}}\left[\bar{F}(Y_{ik}) - \theta\right] - \frac{1}{M_0}\sum_{i=1}^{n} w_{0i}\sum_{k=1}^{m_{0i}}\left[\bar{G}(X_{ik}) - (1-\theta)\right] + o_p\left(n^{-1/2}\right) \tag{12}$$

To show (12), note that

$$\begin{aligned}
&\int \widehat{\bar{F}}(y)d\widehat{\bar{G}}(y) - \int \bar{F}(y)d\bar{G}(y)\\
&= \int \left(\widehat{\bar{F}}(y) - \bar{F}(y)\right)d\left(\widehat{\bar{G}}(y) - \bar{G}(y)\right) + \int \bar{F}(y)d\left[\widehat{\bar{G}}(y) - \bar{G}(y)\right] + \int \left[\widehat{\bar{F}}(y) - \bar{F}(y)\right]d\bar{G}(y)'\\
&= W_1 + W_2 + W_3.
\end{aligned}$$

Now, we show that $W_1 = o_p(n^{-1/2})$. It suffices to show that $E(nW_1^2) = o(1)$. We express W_1 in terms of sums as:

$$\begin{aligned}
W_1 &= \int \left(\widehat{\bar{F}}(y) - \bar{F}(y)\right)d\left(\widehat{\bar{G}}(y) - \bar{G}(y)\right)\\
&= \int \left(\widehat{\bar{F}}(y) - \bar{F}(y)\right)d\widehat{\bar{G}}(y) - \int \left(\widehat{\bar{F}}(y) - \bar{F}(y)\right)d\bar{G}(y)\\
&= \frac{1}{M_1}\sum_{i'=1}^{n} w_{1i'}\sum_{k'=1}^{m_{1i'}}\left[\widehat{\bar{F}}(Y_{i'k'}) - \bar{F}(Y_{i'k'}) - \int \left(\widehat{\bar{F}}(y) - \bar{F}(y)\right)d\bar{G}(y)\right]\\
&= \frac{1}{M_0M_1}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{k=1}^{m_{0i}}\sum_{k'=1}^{m_{1i'}} w_{0i}w_{1i'}\left[\psi(X_{ik}, Y_{i'k'}) - \bar{F}(Y_{i'k'}) - \int \left(\psi(X_{ik}, y) - \bar{F}(y)\right)d\bar{G}(y)\right]\\
&= \frac{1}{M_0M_1}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{k=1}^{m_{0i}}\sum_{k'=1}^{m_{1i'}} w_{0i}w_{1i'}\left[\psi(X_{ik}, Y_{i'k'}) - \bar{F}(Y_{i'k'}) + \bar{G}(X_{ik}) + \theta - 1\right]
\end{aligned}$$

Define

$$H(X, Y) = \psi(X, Y) - \bar{F}(Y) + \bar{G}(X_{ik}) + \theta - 1.$$

Then we have $\mathrm{E}H(X, Y) = E\psi(X, Y) - \mathrm{E}\bar{F}(Y) + \mathrm{E}\bar{G}(X_{ik}) + \theta - 1 = 0$, and for $i \neq j$,

$$\begin{aligned}
\mathrm{E}\left[H(X_{ik}, Y_{i'k'})H(X_{jl}, Y_{i'k'})\right] &= \mathrm{E}\left\{\mathrm{E}\left[H(X_{ik}, Y_{i'k'})H(X_{jl}, Y_{i'k'})|Y_{i'k'}\right]\right\}\\
&= \mathrm{E}\left\{\mathrm{E}\left[H(X_{ik}, Y_{i'k'})|Y_{i'k'}\right]\mathrm{E}\left[H(X_{jl}, Y_{i'k'})|Y_{i'k'})\right]\right\} = 0.
\end{aligned}$$

and, similarly, $\mathrm{E}[H(X_{ik}, Y_{i'k'})H(X_{ik}, Y_{j'l'})] = 0$, for $i' \neq j'$. Consequently,

$$\begin{aligned}
\mathrm{E}\left(nW_1^2\right) &= \frac{n}{M_0^2M_1^2}\mathrm{E}\left\{\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{k=1}^{m_{0i}}\sum_{k'=1}^{m_{1i'}} w_{0i}w_{1i'}H(X_{ik}, Y_{i'k'})|m_{0i}, m_{1i'}\right\}^2 \\
&= \frac{n}{M_0^2M_1^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\mathrm{E}\left\{w_{0i}^2w_{1i'}^2\sum_{k}^{m_{0i}}\sum_{k'=1}^{m_{1i'}}\sum_{k''=1}^{m_{0i}}\sum_{k'''=1}^{m_{1i'}}\mathrm{E}\left[H(X_{ik}, Y_{i'k'})H(X_{ik''}, Y_{i'k'''})|m_{0i}, m_{1i'}\right]\right\} \\
&\le \frac{n}{M_0^2M_1^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\mathrm{E}\left\{w_{0i}^2w_{1i'}^2\sum_{k}^{m_{0i}}\sum_{k'=1}^{m_{1i'}}\sum_{k''=1}^{m_{0i}}\sum_{k'''=1}^{m_{1i'}}4\right\} \\
&= \frac{4n}{M_0^2M_1^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\mathrm{E}(w_{0i}m_{0i})^2(w_{1i'}m_{1i'})^2 \\
&= \frac{4n}{M_0^2M_1^2}\mathrm{O}(n^2) = 0_p(n^{-1/2}).
\end{aligned}$$

The last equality holds because m_{0i}, m_{1i}, w_{0i}, and w_{1i} are bounded. Therefore, $\widehat{\theta}-\theta$ is asymptotically equivalent to $W_2 + W_3$ and (12) follows as

$$W_2 = \int \bar{F}(y)d\left[\widehat{\bar{G}}(y) - \bar{G}(y)\right] = \frac{1}{M_1}\sum_{i=1}^{n} w_{1i}\sum_{k=1}^{m_{1i}}\left[\bar{F}(Y_{ik}) - \theta\right]$$

and

$$W_3 = \int\left[\widehat{\bar{F}}(y) - \bar{F}(y)\right]d\bar{G}(y) = -\frac{1}{M_0}\sum_{i=1}^{n} w_{0i}\sum_{k=1}^{m_{0i}}\left[\bar{G}(X_{ik}) - (1-\theta)\right].$$

Note that $W_2 + W_3$ is an independent sum with mean 0. Applying Central Limit Theorem along with the fact that $\widehat{\bar{F}}(Y_{ik}) = V_{01}(Y_{ik})$ and $\widehat{\bar{G}}(X_{ik}) = 1-V_{10}(X_{ik})$ completes the proof of Theorem 2.1.

Proof of optimal weights expression (8)

The optimal weights can be found via Lagrange multipliers. Let

$$H = \sum_{i=1}^{n}\left[a_iw_{0i}^2 - 2b_iw_{0i}w_{1i} + c_iw_{1i}^2\right] - 2\mu_0\left[\sum_{i=1}^{n}\delta_{0i}m_{0i}w_{0i} - M_0\right] - 2\mu_1\left[\sum_{i=1}^{n}\delta_{1i}m_{1i}w_{1i} - M_1\right],$$

where μ_0 and μ_1 are Langrange multipliers. Taking derivatives of H with respect to w_{0i} and w_{1i} and setting to 0, we have

$$\begin{cases} w_{0i} = a_i^{-1}m_{0i}\mu_0 & \text{if } \delta_{0i} = 1 \text{ and } \delta_{1i} = 0 \\ w_{1i} = c_i^{-1}m_{1i}\mu_1 & \text{if } \delta_{1i} = 1 \text{ and } \delta_{0i} = 0 \end{cases}$$

and

$$\begin{cases} a_iw_{0i} - b_iw_{1i} - \mu_0m_{0i} = 0 & \text{if } \delta_{0i} = 1 \text{ and } \delta_{1i} = 1 \\ c_iw_{1i} - b_iw_{0i} - \mu_1m_{1i} = 0 & \text{if } \delta_{0i} = 1 \text{ and } \delta_{1i} = 1 \end{cases} \tag{13}$$

Solving the system Equation (13) for w_{0i} and w_{1i} gives

$$\begin{cases} w_{0i} = \dfrac{c_im_{0i}}{a_ic_i - b_i^2}\mu_0 + \dfrac{b_im_{1i}}{a_ic_i - b_i^2}\mu_1 & \text{if } \delta_{0i} = 1 \text{ and } \delta_{1i} = 1 \\ w_{1i} = \dfrac{b_im_{0i}}{a_ic_i - b_i^2}\mu_0 + \dfrac{a_im_{1i}}{a_ic_i - b_i^2}\mu_1 & \text{if } \delta_{0i} = 1 \text{ and } \delta_{1i} = 1 \end{cases}$$

Therefore, the unified expression for both weights are

$$\begin{aligned} w_{0i} &= \delta_{0i}(1-\delta_{1i})\frac{m_{0i}\mu_0}{a_i} + \delta_{0i}\delta_{1i}\left[\frac{c_i m_{0i}}{a_i c_i - b_i^2}\mu_0 + \frac{b_i m_{1i}}{a_i c_i - b_i^2}\mu_1\right], \\ w_{1i} &= \delta_{1i}(1-\delta_{0i})\frac{m_{1i}\mu_1}{c_i} + \delta_{0i}\delta_{1i}\left[\frac{b_i m_{0i}}{a_i c_i - b_i^2}\mu_0 + \frac{a_i m_{1i}}{a_i c_i - b_i^2}\mu_1\right] \end{aligned}$$

Plugging the above expressions for w_{0i} and w_{1i} into the two constrains, $\sum_{i=1}^n \delta_{0i}m_{0i}w_{0i} = M_0$ and $\sum_{i=1}^n \delta_{1i}m_{1i}w_{1i} = M_1$, we have the following system of equations for μ_0 and μ_1:

$$\begin{aligned} &\left[\sum_{i=1}^n \delta_{0i}(1-\delta_{1i})m_{0i}^2 a_i^{-1} + \sum_{i=1}^n \delta_{0i}\delta_{1i}\frac{m_{0i}^2 c_i}{a_i c_i - b_i^2}\right]\mu_0 + \left[\sum_{i=1}^n \delta_{0i}\delta_{1i}\frac{m_{0i}m_{1i}b_i}{a_i c_i - b_i^2}\right]\mu_1 = M_0 \\ &\left[\sum_{i=1}^n \delta_{0i}\delta_{1i}\frac{m_{0i}m_{1i}b_i}{a_i c_i - b_i^2}\right]\mu_0 + \left[\sum_{i=1}^n (1-\delta_{0i})\delta_{1i}m_{1i}^2 c_i^{-1} + \sum_{i=1}^n \delta_{0i}\delta_{1i}\frac{m_{1i}^2 a_i}{a_i c_i - b_i^2}\right]\mu_1 = M_1 \end{aligned}$$

and solving this system of equations for μ_0 and μ_1 yields

$$\mu_0 = \frac{CM_0 - BM_1}{\Delta}$$
$$\mu_1 = \frac{AM_1 - BM_0}{\Delta}$$

which completes the proof of (8).

Proof of theorem 2.2. Let $\beta = (\sigma_U^2, \sigma_V^2, \rho_U, \rho_V \text{ and } \rho_{UV})^t$ and $\widehat{\beta}$ be the estimator of β. Then have

$$\widehat{\theta}_{op} - \theta = \left[\widehat{\theta}_w\left(\widehat{\beta}\right) - \widehat{\theta}_w(\beta)\right] + \left[\widehat{\theta}_w(\beta) - \theta\right]$$

Using Tylor's Expansion, It follows that

$$n^{1/2}\left[\widehat{\theta}_w\left(\widehat{\beta}\right) - \widehat{\theta}_w(\beta)\right] = \frac{\partial\widehat{\theta}_w(\beta)}{\partial\beta}n^{1/2}\left(\widehat{\beta} - \beta\right) + o_p(1).$$

Notice that θ does not depend on β and $E\widehat{\theta}_w(\beta) = \theta$. We have $\mathrm{E}\frac{\partial\widehat{\theta}_w(\beta)}{\partial\beta} = \frac{\partial \mathrm{E}\widehat{\theta}_w(\beta)}{\partial\beta} = \frac{\partial\theta}{\partial\beta} = 0$, which implies that $\widehat{\theta}_w(\widehat{\beta}) - \widehat{\theta}_w(\beta) = o_p(n^{-1/2})$. Finally, applying Theorem 1 to $\widehat{\theta}_w(\beta)$ completes the proof.

Proof of theorem 3.1. Let $\epsilon_i^r = \sum_{k=1}^{m_{1i}}[\bar{F}(Y_{ik}^r) - m_{1i}\theta^r]$ and $\xi_i^r = \sum_{k=1}^{m_{1i}}[1 - \bar{G}(Y_{ik}^r) - m_{0i}\theta^r]$, $r = 1, 2$. Then applying (12) to both $\widehat{\theta}^1$ and $\widehat{\theta}^2$, we have

$$\begin{aligned} \widehat{\theta}_w^1 - \theta^1 &= \frac{1}{M_1}\sum_{i=1}^n w_{1i}^1\epsilon_i^1 + \frac{1}{M_0}\sum_{i=1}^n w_{0i}^1\xi_i^1 + o_p(n^{-1/2}) \\ \widehat{\theta}_w^2 - \theta^2 &= \frac{1}{M_1}\sum_{i=1}^n w_{1i}^2\epsilon_i^2 + \frac{1}{M_0}\sum_{i=1}^n w_{0i}^2\xi_i^2 + o_p(n^{-1/2}) \end{aligned}$$

and hence $\widehat{\theta}_w^1 - \widehat{\theta}_w^2 - (\theta^1 - \theta^2)$ is asymptotically equivalent to an independent sum:

$$\widehat{\theta}_w^1 - \widehat{\theta}_w^2 - (\theta^1 - \theta^2) = \sum_{i=1}^n \left[\frac{1}{M_1}w_{1i}^1\epsilon_i^1 + \frac{1}{M_0}w_{0i}^1\xi_i^1 - \frac{1}{M_1}w_{1i}^2\epsilon_i^2 - \frac{1}{M_0}w_{0i}^2\xi_i^2\right]$$

Applying CLT to the independent sum and replacing ϵ_i^r and ξ_i^r by their estimates $\widehat{\epsilon}_i^r = \sum_{k=1}^{m_{1i}}[\widehat{\bar{F}}(Y_{ik}^r) - m_{1i}\widehat{\theta}^r]$ and $\widehat{\xi}_i^r = \sum_{k=1}^{m_{0i}}[(1 - \widehat{\bar{G}}(X_{ik}^r)) - m_{0i}\widehat{\theta}^r]$, respectively, it follows that

$$\frac{\widehat{\theta}^1 - \widehat{\theta}^1 - (\theta^1 - \theta^2)}{\sqrt{\widehat{\text{var}}\left(\widehat{\theta}^1 - \widehat{\theta}^2\right))}} \to N(0,1)$$

where

$$\begin{aligned}\widehat{\text{var}}\left(\widehat{\theta}^1 - \widehat{\theta}^2\right)) &= \sum_{i=1}^{n}\left[\frac{1}{M_1}w_{1i}^1\widehat{\epsilon}_i^1 + \frac{1}{M_0}w_{0i}^1\widehat{\xi}_i^1 - \frac{1}{M_1}w_{1i}^2\widehat{\epsilon}_i^2 - \frac{1}{M_0}w_{0i}^2\widehat{\xi}_i^2\right]^2 \\ &= \widehat{\text{var}}\left(\widehat{\theta}^1\right) + \widehat{\text{var}}\left(\widehat{\theta}^2\right) - 2\widehat{\text{cov}}\left(\widehat{\theta}^1, \widehat{\theta}^2\right)\end{aligned}$$