Weighted mean difference statistics for paired data in the presence of missing values

Yuntong Li¹, Brent J Shelton²,³, William St Clair⁴, Heidi L Weiss²,³, John L Villano³, Arnold J Stromberg⁵, Chi Wang²,³,⁵ and Li Chen²,³

Abstract
Missing data is a common issue in many biomedical studies. Under a paired design, some subjects may have missing values in either one or both of the conditions due to loss of follow-up, insufficient biological samples, etc. Such partially paired data complicate statistical comparison of the distribution of the variable of interest between the two conditions. In this article, we propose a general class of test statistics based on the difference in weighted sample means without imposing any distributional or model assumption. An optimal weight is derived from this class of tests. Simulation studies show that our proposed test with the optimal weight performs well and outperforms existing methods in practical situations. Two cancer biomarker studies are provided for illustration.

Keywords
Paired data, missing data, biomarker data, nonparametric tests, mean difference tests

I Introduction
Comparing the distribution of a variable of interest between two experimental conditions/groups is a common problem in many biomedical studies. A paired design is frequently used because it can reduce the variability in the observations due to differences between individuals. However, in practice, the data are often partially paired, where only a proportion of individuals have measurements under both conditions while other individuals have missing measurements in either one or both of the conditions. For example, a study enrolled a total of 18 prostate cancer patients who underwent surgery. The study longitudinally measured three biomarkers, including PSA, CRP, and IL8, at baseline (prior to surgery), 3-month and 6-month follow-ups with a goal of comparing those biomarker levels before and after surgery. Due to missed visits or loss of follow-up, 61% of patients had missing values at 3-month and 50% had missing values at 6-month follow-up. As another example, a breast cancer biomarker study was conducted to compare expression of five proteins of interest between primary and recurrent tumors.¹ Tissue blocks of paired primary and recurrent tumors from 22 breast cancer patients were selected as part of a tissue microarray construction. Immunohistochemistry (IHC) was performed to measure the protein expressions. A 0 to 3 intensity score was assigned to each tissue core by a pathologist. Missing values were present in primary and/or recurrent tumor tissue samples due to the lack of remaining tissue, lack of invasive carcinoma cells or...
difficulty in interpreting tissues. The missing proportion in primary tumor tissue samples ranged from 5% to 14%, and that in recurrent tumor tissue samples ranged from 0% to 19%.

Let \((Y_1, Y_2)\) denote the paired variables from groups 1 and 2, and \((Y_{1i}, Y_{2i})\), \(i = 1, \ldots, n\), be a simple random sample of \((Y_1, Y_2)\). For partially paired data, suppose \((Y_1, Y_2)\) are both observed in individuals \(1, \ldots, n_g\), \(Y_1\) alone is observed in individuals \(n_0 + 1, \ldots, n_0 + n_1\), and \(Y_2\) alone is observed in individuals \(n_0 + n_1 + 1, \ldots, n_0 + n_1 + n_2\). Let \(R_g (g = 1, 2)\) be the missing indicator of \(Y_g\), where \(R_g = 1\) if \(Y_g\) is observed or \(0\) if \(Y_g\) is missing. In this article, we assume \((R_1, R_2)\) is independent of \((Y_1, Y_2)\), that is, missing completely at random (MCAR).\(^2\) The observed data can be divided into the following four blocks: group 1 data from complete pairs \(Y^C_1 = (Y_{1,1}, \ldots, Y_{1,n_g})\), group 1 data from incomplete pairs \(Y^I_1 = (Y_{1,0} + 1, \ldots, Y_{1,n_0 + n_1})\), group 2 data from complete pairs \(Y^C_2 = (Y_{2,1}, \ldots, Y_{2,n_0})\), and group 2 data from incomplete pairs \(Y^I_2 = (Y_{2,0} + n_1 + 1, \ldots, Y_{2,n_0 + n_1 + n_2})\). The data structure is illustrated below.

| Both \(Y_1\) and \(Y_2\) are observed | \(Y_1\) alone is observed | \(Y_2\) alone is observed | Both \(Y_1\) and \(Y_2\) are missing |
|-------------------------------------|--------------------------|--------------------------|--------------------------|
| \(Y_{1,1}, \ldots, Y_{1,n_g}\)     | \(Y_{1,0} + 1, \ldots, Y_{1,n_0 + n_1}\) | \(Y_{2,0} + n_1 + 1, \ldots, Y_{2,n_0 + n_1 + n_2}\) | |
| \(Y_{2,1}, \ldots, Y_{2,n_0}\)     |                          |                          |                          |

Naive methods to handle partially paired data are either only using data of complete pairs or imputing the missing values by sample mean or median and then applying standard statistical tests. However, only using data of complete pairs is inefficient because it discards part of the data. Imputing the missing values by sample mean will underestimate the variances of estimates, and lead to erroneous results.\(^2\) To fully utilize all available data, the mixed models, including the linear mixed model and cumulative link mixed model, can be applied. However, those models require data normality assumption and distributional assumption on random effects. In addition to the naive methods and mixed models, statistical tests specifically designed for partially paired data, including rank-based tests, mean difference-based tests and \(P\)-value pooling methods, have been proposed. The rank-based tests considered weighted linear combinations of Wilcoxon signed-rank and Wilcoxon rank-sum tests. Dubnicka et al.\(^3\) combined a Wilcoxon signed-rank test statistic on complete pairs, that is, comparing \(Y^C_1\) and \(Y^C_2\), and a Wilcoxon rank-sum test statistic on incomplete pairs, that is, comparing \(Y^I_1\) and \(Y^I_2\). Brunner et al.\(^4\) Konietschke et al.\(^5\) and Fong et al.\(^6\) considered all possible comparisons of blocks between groups, including \(Y^C_1\) and \(Y^C_2\), \(Y^C_1\) and \(Y^C_2\), \(Y^I_1\) and \(Y^I_2\), \(Y^I_1\) and \(Y^I_2\), \(Y^I_1\) and \(Y^I_2\). One shortcoming of rank-based tests is that they might not be sensitive to the magnitude of the difference between groups. In contrast, mean difference-based tests took into account the absolute values of observations and could be more powerful. Bhöj\(^7\) considered a weighted linear combination of the paired and unpaired \(t\)-test statistics under the assumption of bivariate normal distribution. Martinez-Cambler et al.\(^8\) proposed a test based on the difference in sample means between groups and obtained asymptotic normal distribution of the test without any distribution assumption. In addition, the \(P\)-value pooling methods considered combining \(P\)-values from paired and unpaired tests. Kuan and Huang\(^9\) proposed a weighted inverse normal approach to combine independent \(P\)-values for the case that missing occurs in both groups. Qi et al.\(^10\) proposed a Fisher’s pooling method and a weighted inverse normal method to combine dependent \(P\)-values for the case that missing occurs only in one group. Qi et al.\(^11\) also compared \(P\)-value pooling methods with naive methods only using data of complete pairs and provided guidance on how to choose the best one among those tests in practical situations.

In this article, we propose a general class of test statistics based on the difference in weighted sample means of the two groups for partially paired data. This general class of test statistics include the test statistics proposed by Bhöj\(^7\) and Martinez-Cambler et al.\(^8\) as special cases. We further derive an optimal weight for this class of tests. Our tests are constructed without imposing any distributional or model assumption. Therefore, they are applicable to continuous data with an arbitrary distribution as well as discrete interval data. We evaluate the performance of the proposed test with optimal weight and compare it with several existing tests by using simulation studies and real data analyses.

## 2 Methods

### 2.1 Weighted mean difference test statistics

Let \(\mu_1 = EY_1\) and \(\mu_2 = EY_2\). To test whether there is a difference in distribution between paired outcomes \(Y_1\) and \(Y_2\), we focus on examining whether there is a mean difference, that is, \(H_0: D = \mu_1 - \mu_2 = 0\) versus \(H_1: D \neq 0\). We are going to first propose a general class of estimators for \(\mu_1 - \mu_2\), and then construct Wald tests for \(H_0\) based on those estimators. Specifically, we consider the following class of estimators

\[
\hat{D}(\hat{w}_1, \hat{w}_2) = \{ \hat{w}_1 \bar{Y}_1 + (1 - \hat{w}_1) \bar{Y}_1^{\text{I}} \} - \{ \hat{w}_2 \bar{Y}_2 + (1 - \hat{w}_2) \bar{Y}_2^{\text{I}} \}
\]

(1)
where \( \hat{Y}_g^C, \hat{Y}_g^I \) are the means for complete pair data \( Y_g^C \) and incomplete pair data \( Y_g^I \), respectively, and \( \hat{w}_1 \) and \( \hat{w}_2 \) are random weights in \([0, 1]\). The first item on the right-hand side of equation (1) is a weighted sample mean of group 1, and the second item is a weighted sample mean of group 2. Therefore, \( \bar{D}(\hat{w}_1, \hat{w}_2) \) quantifies the difference in weighted sample means between two groups. As a special case, when \( \hat{w}_1 = n_0 / (n_0 + n_1) \) and \( \hat{w}_2 = n_0 / (n_0 + n_2) \), the \( \bar{D}(\hat{w}_1, \hat{w}_2) \) reduces to the difference in sample mean. As another special case, when \( \hat{w}_1 = \hat{w}_2 = 1 \), the \( \bar{D}(\hat{w}_1, \hat{w}_2) \) reduces to the difference in sample mean only using complete pairs.

Our proposed class of estimators provides a general weighting mechanism to take into account two specific characteristics of the data. First, for observations from the same group, those from complete pairs and from incomplete pairs are given different weights because they contain different amount of information. Second, observations from different groups are given different weights because different groups may have different missing proportions and different variances. In the following Theorem 1, we establish the asymptotic normality of \( \bar{D}(\hat{w}_1, \hat{w}_2) \).

**Theorem 1.** Suppose there exist deterministic weights \( 0 \leq w_1 \leq 1 \) and \( 0 \leq w_2 \leq 1 \) such that \( \sqrt{n}(\hat{w}_1 - w_1) \) and \( \sqrt{n}(\hat{w}_2 - w_2) \) are \( O_p(1) \). Under the MCAR assumption, that is, \((R_1, R_2)\) is independent of \((Y_1, Y_2)\), we have

\[
\sqrt{n}\left\{\bar{D}(\hat{w}_1, \hat{w}_2) - (\mu_1 - \mu_2)\right\} \rightarrow N\left(0, \sigma_D^2(w_1, w_2)\right).
\]

When there are missing values in both groups

\[
\sigma_D^2(w_1, w_2) = \left\{\frac{w_1^2}{p_{12}} + \frac{(1 - w_1)^2}{p_1 - p_{12}}\right\} \sigma_1^2 + \left\{\frac{w_2^2}{p_{12}} + \frac{(1 - w_2)^2}{p_2 - p_{12}}\right\} \sigma_2^2 - \frac{2w_1w_2\rho_1\sigma_1\sigma_2}{p_{12}},
\]

where \( p_g \) is the probability that \( Y_g \) is observed, \( p_{12} \) is the probability that both \( Y_1 \) and \( Y_2 \) are observed, \( \sigma_g^2 \) is the variance of \( Y_g \), and \( \rho \) is the Pearson’s correlation coefficient of \( Y_1 \) and \( Y_2 \). When missing values only present in group 1, the \( \sigma_D^2(w_1, w_2) \) is given by equation (2) with \( w_1 = 1, w_2 = 1, p_{12} = p_1 \), and 0/0 = 0. When missing values only present in group 2, the \( \sigma_D^2(w_1, w_2) \) is given by equation (2) with \( w_2 = 1, p_1 = 1, p_{12} = p_2 \), and 0/0 = 0. The proof of this theorem is provided in Appendix A.

Based on \( \bar{D}(\hat{w}_1, \hat{w}_2) \), we construct the following class of weighted mean difference (WMD) test statistics to assess the hypothesis \( H_0 \) versus \( H_1 \):

\[
T(\hat{w}_1, \hat{w}_2) = \frac{\sqrt{n}\bar{D}(\hat{w}_1, \hat{w}_2)}{\hat{\sigma}(\hat{w}_1, \hat{w}_2)},
\]

where \( \hat{\sigma}(\hat{w}_1, \hat{w}_2) \) is an estimator of \( \sigma_D(w_1, w_2) \) by plugging in \( \hat{w}_1, \hat{w}_2 \) as well as the following estimators into equation (2): \( \hat{p}_g = (n_g + n_g') / n \) is the non-missing proportion and \( \hat{\sigma}_g \) is the sample standard deviation for group \( g, g = 1, 2 \), \( \hat{p}_{12} = n_0 / n \) is the proportion of complete pairs, and \( \hat{\rho} \) is the sample Pearson’s correlation coefficient using complete pairs. Based on Theorem 1, \( T(\hat{w}_1, \hat{w}_2) \) converges to a standard normal distribution, \( N(0, 1) \), under \( H_0 \). Therefore, the critical region of WMD test statistics is \( \{T(\hat{w}_1, \hat{w}_2) > z_{1-a/2}\} \), where \( a \) is the significance level. Note that \( \hat{\sigma}(\hat{w}_1, \hat{w}_2) \) can underestimate the standard error of \( \bar{D}(\hat{w}_1, \hat{w}_2) \) when the sample size is small, which leads to inflated type I error rates of the WMD tests, although the type I error rates become closer to the nominal value of 0.05 when the sample size increases (see simulation studies in Supplemental Tables S1 and S2). Therefore, we recommend using a bootstrap-based estimator for the standard error of \( \bar{D}(\hat{w}_1, \hat{w}_2) \) in the calculation of WMD tests.

### 2.2 Power and optimal weight of WMD tests

According to Theorem 1, under \( H_1 \), \( T(\hat{w}_1, \hat{w}_2) = \sqrt{n}(\mu_1 - \mu_2) / \hat{\sigma}(w_1, w_2) \) converges to \( N(0, 1) \). Therefore, the power of WMD test statistics is \( 1 - \Phi[z_{1-a/2} - \sqrt{n}(\mu_1 - \mu_2) / \sigma_D(w_1, w_2)] + \Phi[-z_{1-a/2} - (\mu_1 - \mu_2) / \sigma_D(w_1, w_2)] \). Because the power is a decreasing function of \( \sigma_D(w_1, w_2) \), it is maximized when \( \sigma_D(w_1, w_2) \) is minimized. We present here the optimal weight that minimizes \( \sigma_D(w_1, w_2) \) with a detailed derivation provided in Appendix B.

When there are missing values in both groups, the optimal weight is

\[
(w_1^{opt}, w_2^{opt}) = \begin{cases} (w_1^*, w_2^*) & \text{if } 0 \leq w_1^* \leq 1, 0 \leq w_2^* \leq 1 \\ \text{argmin } \sigma_D^2(w_1, w_2) & \text{otherwise} \end{cases}
\]
with

$$w_1^* = \frac{p_2 p_{12} \sigma_1^2 + \rho \sigma_1 \sigma_2 (p_1 - p_{12})}{p_1 p_{12} \sigma_1^2 - \rho^2 \sigma_1^2 (p_1 - p_{12}) (p_2 - p_{12})}, \quad w_2^* = \frac{p_1 p_{12} \sigma_1^2 + \rho \sigma_1 \sigma_2 (p_2 - p_{12})}{p_1 p_{12} \sigma_1^2 - \rho^2 \sigma_1^2 (p_1 - p_{12}) (p_2 - p_{12})}$$

and \( A \) is a collection of four points \( A = \{(0, h_5(0)), (1, g(h_2(1))), (h_1(0), 0), (g(h_1(1), 1))\}, h_1(x) = (p_1 p_2 p_{12})^{-1} (p_1 p_2 p_{12} + x \rho \sigma_2 (p_1 - p_{12})), h_2(x) = (p_2 p_{12})^{-1} (p_1 p_2 p_{12} + x \rho \sigma_2 (p_2 - p_{12}))\), and \( g(z) \) equals 0 if \( z < 0 \); equals \( z \) if \( 0 \leq z \leq 1 \); and \( g(z) = 1 \) if \( z > 1 \).

When missing values only present in group 1, we have \( p_2 = 1, p_{12} = p_{1} \), and \( w_1 = 1 \). The optimal weight for \( w_2 \) is

$$w_2^{opt} = g(p_1 + \rho \sigma_2 p_{12}^2 (1 - p_1)).$$

When missing values only present in group 2, we have \( p_1 = 1, p_{12} = p_2 \), and \( w_2 = 1 \). The optimal weight for \( w_1 \) is

$$w_1^{opt} = g(p_2 + \rho \sigma_1 p_{12} (1 - p_2)).$$

The optimal weight is estimated by replacing \( p_1, p_2, p_{12}, \sigma_1, \sigma_2, \) and \( \rho \) with their estimators as described in Section 2.1.

The WMD test with this estimated optimal weight is referred to as WMD_{opt}.

### 2.3 Relationships to existing methods

Our proposed test is constructed by calculating the difference in weighted sample means between the two groups. When we set

$$\tilde{w}_1 = \frac{n_0}{n_0 + n_1}, \quad \tilde{w}_2 = \frac{n_0}{n_0 + n_2},$$

the weight for each observation in the first group becomes \( 1/(n_0 + n_1) \), and that for each observation in the second group becomes \( 1/(n_0 + n_2) \). The \( D \) reduces to

$$D = \tilde{Y}_1 - \tilde{Y}_2,$$

where \( \tilde{Y}_g \) is the sample mean of the \( g \)th group. The \( D \) is the simple mean difference between groups, which is referred to as WMD_{s}. The corresponding asymptotic variance formula reduces to

$$\frac{\sigma_1^2}{p_1} + \frac{\sigma_2^2}{p_2} - \frac{2 \rho \sigma_1 \sigma_2 p_{12}}{p_1 p_2}.$$

The WMD_{s} is asymptotically equivalent to the \( M_p \) test.\(^8\) Note that the \( M_p \) test assigns equal weight to each observation from the same group. However, the complete pair and incomplete pair may contain different amount of information. By introducing a weighted sample mean, our test allows different weights to complete and incomplete data, and thus could improve the power of the test. In addition, by introducing the weight, our test can also be considered as a weighted combination of paired and unpaired \( t \) statistics. Let

$$\tilde{w}_1 = \tilde{w}_2 = \frac{\lambda / \sqrt{n_0}}{\sqrt{\lambda / \sqrt{n_0}} + 1 / \sqrt{1/n_1 + 1/n_2}},$$

where \( \lambda \) is a constant between 0 and 1 in the test statistic formula proposed by Bhoj,\(^7\) our test statistic reduces to the test by Bhoj.\(^7\) Bhoj\(^7\) assumes that data are normally distributed and calculated \( P \)-values based on a \( t \)-distribution. As we do not require data to be normally distributed, our \( P \)-value is obtained using either normal approximation or bootstrap.

### 3 Simulation studies

#### 3.1 Continuous data

We first evaluated the performance of the proposed WMD test with the estimated optimal weight (WMD_{o}) for analyzing continuous data, and compared to the paired \( t \)-test on complete pairs (\( t_{cp} \)), paired \( t \)-test with imputation (\( t_{im} \)), linear mixed model (LMM), Wilcoxon signed-rank test on complete pairs (\( W_{cp} \)), Wilcoxon signed-rank test with imputation (\( W_{im} \)), \( P \)-value pooling method using Fisher’s method to combine \( P \)-values from Wilcoxon signed-rank and Wilcoxon rank-sum tests weighting by square root of geometric mean of sample sizes (\( F-W_{2} \)),\(^9\) SR-MW and MW-MW methods combining the
Wilcoxon signed-rank and Wilcoxon rank-sum tests, and WMD, which is a test based on the difference in sample means between groups and is asymptotically equivalent to the $M_0$ test. The bootstrap method with 1500 bootstrap samples was used for estimating the standard error of $\hat{D}(\hat{\omega}_1, \hat{\omega}_2)$ in the calculation of WMD and WMD. For $t_{\text{im}}$ and $W_{\text{im}}$, missing values were imputed by median of each group. For LMM, we used the `nlme` R-package with the default restricted maximum likelihood setting. For the SR-MW and MW-MW methods, we used the `robustrank` R-package. For the F-W method, we used the R codes provided by Qi et al. To mimic real data situation, we simulated data based on complete pairs of PSA data from the prostate cancer study described in Section 4.1. We randomly sampled $n$ subjects from the PSA data with replacement. For each subject, both the baseline and 6-month follow-up measurements were selected, which maintained the within-subject correlation in the real data. Because the baseline and 6-month measurements had large differences, all methods would yield powers close to one for detecting such difference. To better investigate the difference among methods, we reduced the difference between those two measurements by subtracting a value of 6 from the baseline measurement. We considered two scenarios of the correlation between the baseline and 6-month measurements. The first scenario directly adopted the correlation from the real data, which had a Spearman correlation coefficient $R = 0.60$. The second scenario was a weaker correlation with $R = 0.32$, which was generated by adding a random normal error term to the baseline measurement. We considered two missing scenarios: missing only in group 2 with 50% missing in group 2 and missing in both groups with 50% missing in both groups. We considered the sample size $n$ equal to 20, 30, 50, or 100. Under each sample size, we replicated the simulation 1000 times.

Our first set of simulations evaluated the type I error rate of our proposed method as well as other methods. To generate data under the null hypothesis, we randomly rearranged the labels of the paired baseline and 6-month measurements so that distributions of simulated data under the two time points were expected to be the same. Table 1 shows the type I error rate for each method at a given significant level $\alpha = 0.05$. WMD, WMD, $W_{\text{cp}}$, F-W, SR-MW, and MW-MW were able to control the type I error rate close to the nominal level of 0.05 under all sample sizes. Note that F-W was only applicable to the case when missing values only present in one group. $t_{\text{cp}}$ and LMMs were too conservative when $R = 0.60$ and the sample size was small. But their type I error rates became close to the nominal level as the sample size increased. $t_{\text{im}}$ had inflated type I error rate when $R = 0.32$, and $W_{\text{im}}$ had substantially inflated type I error rate under most scenarios, which were likely due to the reduced data variation after imputation.

Our next set of simulations evaluated and compared the power of different methods. Because imputation methods had inflated type I error rates, we excluded them from the power comparison. We generated data following a similar procedure as the simulations for type I error rate evaluation, except that the labels of the paired baseline and 6-month measurements from the PSA data were maintained. As shown in Table 2, WMD had the highest power among all methods across all scenarios. WMD had much higher power than $t_{\text{cp}}$ and $W_{\text{cp}}$, since those two methods only used the portion of complete pairs data. WMD also had much higher power than LMM for small sample sizes, likely due to LMM’s requirement of data normality that was not satisfied in the real data ($P<0.001$ based on Shapiro–Wilks test). In addition, WMD had much higher power than F-W, SR-MW, and MW-MW. Finally, the power of WMD was higher than that of WMD, especially for the scenario of a higher correlation ($R = 0.6$).

### 3.2 Discrete interval data

We assessed the performance of WMD for analyzing discrete interval data, and compared to several methods, including $W_{\text{cp}}$, $W_{\text{im}}$, cumulative link mixed model (CLMM), F-W, SR-MW, MW-MW, and WMD. Note that we did not include $t$-test and LMM in the comparison because those methods were designed for continuous data. Instead, we included CLMM in the comparison, which is more appropriate for discrete data. For CLMM, we used `clmm` function in the `ordinal` R-package. The `nAGQ` option in `clmm` function was set to 5 for best performance as suggested by the authors. Our data were simulated by random sampling with replacement from the complete pairs data of PAR4 from the breast cancer recurrence study described in Section 4.2. We considered two correlation scenarios of the paired data. The first scenario adopted the correlation from the real data, which had a Spearman correlation coefficient $R = 0.50$. The second scenario considered a higher correlation, where we increased the paired data correlation using the following steps. First, we calculated the 40 and 60 percentiles of group 1, denoted by $Q_1$ and $Q_2$, respectively. Second, for each observation $O$ in group 2, if $O < Q_1$, we decreased its value by one level; if $O \geq Q_2$, we increased its value by one level. The resulting data had a correlation coefficient $R = 0.75$. Other aspects of the simulation setting were the same as those for continuous data described in the previous subsection.

Table 3 shows the type I error rate for each method. WMD, F-W, and SR-MW had type I error rates close to the nominal value. $W_{\text{cp}}$ was very conservative when the sample size was small. WMD and CLMM had slightly inflated type I error rate for a few small sample size cases. MW-MW was conservative for some cases. $W_{\text{im}}$ had substantially inflated type I error rate in all scenarios. Table 4 shows the results for power comparison. We again excluded $W_{\text{im}}$ due to its substantially...
| Corr | Missing | n   | t _cp | t _im | LMM | W _cp | W _im | F-W _2 | SR-MW | MW-MW | WMD _4 | WMD _o |
|------|---------|-----|------|------|-----|-------|-------|--------|--------|--------|--------|--------|--------|
| R = 0.60 | (0%, 50%) | 20  | 0.014 | 0.043 | 0.017 | 0.028 | 0.082 | 0.049 | 0.036 | 0.053 | 0.072 | 0.072 |
|       |         | 30  | 0.021 | 0.043 | 0.036 | 0.048 | 0.075 | 0.060 | 0.040 | 0.043 | 0.064 | 0.067 |
|       |         | 50  | 0.050 | 0.053 | 0.050 | 0.049 | 0.074 | 0.064 | 0.035 | 0.043 | 0.066 | 0.067 |
|       |         | 100 | 0.050 | 0.058 | 0.049 | 0.047 | 0.060 | 0.069 | 0.055 | 0.043 | 0.062 | 0.059 |
| (50%, 50%) |       | 20  | 0.017 | 0.032 | 0.019 | 0.018 | 0.185 | –     | 0.035 | 0.057 | 0.040 | 0.071 |
|       |         | 30  | 0.008 | 0.032 | 0.012 | 0.031 | 0.148 | –     | 0.042 | 0.053 | 0.041 | 0.069 |
|       |         | 50  | 0.016 | 0.039 | 0.026 | 0.040 | 0.091 | –     | 0.040 | 0.033 | 0.045 | 0.051 |
|       |         | 100 | 0.035 | 0.050 | 0.043 | 0.043 | 0.040 | –     | 0.036 | 0.026 | 0.053 | 0.050 |
| R = 0.32 | (0%, 50%) | 20  | 0.033 | 0.067 | 0.045 | 0.042 | 0.121 | 0.070 | 0.049 | 0.073 | 0.064 | 0.067 |
|       |         | 30  | 0.037 | 0.060 | 0.038 | 0.043 | 0.118 | 0.055 | 0.041 | 0.048 | 0.060 | 0.061 |
|       |         | 50  | 0.040 | 0.074 | 0.052 | 0.049 | 0.116 | 0.057 | 0.042 | 0.044 | 0.063 | 0.066 |
|       |         | 100 | 0.066 | 0.079 | 0.068 | 0.067 | 0.147 | 0.044 | 0.037 | 0.039 | 0.067 | 0.065 |
| (50%, 50%) |       | 20  | 0.041 | 0.181 | 0.033 | 0.021 | 0.343 | –     | 0.035 | 0.069 | 0.063 | 0.073 |
|       |         | 30  | 0.026 | 0.187 | 0.035 | 0.031 | 0.388 | –     | 0.050 | 0.063 | 0.062 | 0.064 |
|       |         | 50  | 0.043 | 0.159 | 0.050 | 0.052 | 0.427 | –     | 0.051 | 0.055 | 0.060 | 0.065 |
|       |         | 100 | 0.046 | 0.150 | 0.055 | 0.052 | 0.477 | –     | 0.052 | 0.051 | 0.056 | 0.053 |

Each entry was based on 1000 simulation replicates. WMD _4_ and WMD _o_ were calculated based on 1500 bootstrap samples. Significant level α was set to 0.05. – indicates results unavailable because F-W _2_ is only applicable to cases when missing values only present in one group.
Table 2. Simulation results for power comparison based on continuous data.

| Corr   | Missing | n  | \( t_{cp} \) | LMM | \( W_{cp} \) | \( F-W_2 \) | SR-MW | MW-MW | WMD_a | WMD_o |
|--------|---------|----|--------------|-----|-------------|-------------|--------|--------|--------|--------|
|        | (0%, 50%) |    |              |     |             |             |        |        |        |        |
| R = 0.60 |         | 20 | 0.071        | 0.035 | 0.082       | 0.173       | 0.123  | 0.050  | 0.550  | 0.683  |
|        |         | 30 | 0.208        | 0.141 | 0.158       | 0.497       | 0.114  | 0.041  | 0.779  | 0.846  |
|        |         | 50 | 0.494        | 0.497 | 0.234       | 0.881       | 0.158  | 0.055  | 0.959  | 0.976  |
|        |         | 100| 0.939        | 0.988 | 0.446       | 1.000       | 0.291  | 0.045  | 1.000  | 1.000  |
|        | (50%, 50%) | 20 | 0.008        | 0.033 | 0.019       | –           | 0.074  | 0.095  | 0.200  | 0.442  |
|        |         | 30 | 0.048        | 0.158 | 0.068       | –           | 0.104  | 0.114  | 0.350  | 0.565  |
|        |         | 50 | 0.131        | 0.442 | 0.097       | –           | 0.096  | 0.088  | 0.565  | 0.725  |
|        |         | 100| 0.506        | 0.926 | 0.225       | –           | 0.121  | 0.088  | 0.917  | 0.960  |
| R = 0.32 |         | 20 | 0.162        | 0.169 | 0.179       | 0.272       | 0.240  | 0.274  | 0.408  | 0.410  |
|        |         | 30 | 0.296        | 0.316 | 0.290       | 0.424       | 0.320  | 0.343  | 0.562  | 0.559  |
|        |         | 50 | 0.462        | 0.529 | 0.391       | 0.573       | 0.416  | 0.423  | 0.681  | 0.679  |
|        |         | 100| 0.706        | 0.754 | 0.526       | 0.720       | 0.530  | 0.537  | 0.841  | 0.847  |
|        | (50%, 50%) | 20 | 0.085        | 0.111 | 0.062       | –           | 0.186  | 0.236  | 0.238  | 0.292  |
|        |         | 30 | 0.135        | 0.232 | 0.152       | –           | 0.193  | 0.230  | 0.322  | 0.364  |
|        |         | 50 | 0.246        | 0.446 | 0.249       | –           | 0.345  | 0.343  | 0.476  | 0.506  |
|        |         | 100| 0.462        | 0.731 | 0.382       | –           | 0.472  | 0.455  | 0.711  | 0.741  |

Each entry was based on 1000 replicates. \( \text{WMD}_a \) and \( \text{WMD}_o \) were calculated based on 1500 bootstrap samples. Significant level \( \alpha \) was set to 0.05. – indicates results unavailable because \( F-W_2 \) is only applicable to cases when missing values only present in one group.
Table 3. Simulation results for type I error rates based on discrete interval data.

| Corr  | Missing  | n   | \(W_{cp}\) | \(W_{im}\) | CLMM | F-\(W_2\) | SR-MW | MW-MW | WMD_1 | WMD_0 |
|-------|----------|-----|-------------|-------------|------|-----------|-------|-------|-------|-------|
| \(R = 0.75\) | (0%, 50%) | 20  | 0.020       | 0.139       | 0.044 | 0.055      | 0.025 | 0.022 | 0.061 | 0.051 |
|       |          | 30  | 0.028       | 0.138       | 0.052 | 0.046      | 0.036 | 0.028 | 0.051 | 0.047 |
|       |          | 50  | 0.042       | 0.163       | 0.065 | 0.037      | 0.033 | 0.023 | 0.053 | 0.049 |
|       |          | 100 | 0.059       | 0.211       | 0.079 | 0.024      | 0.054 | 0.017 | 0.050 | 0.059 |
|       | (50%, 50%) | 20  | 0.004       | 0.354       | 0.057 | –          | 0.026 | 0.023 | 0.114 | 0.068 |
|       |          | 30  | 0.007       | 0.403       | 0.043 | –          | 0.034 | 0.030 | 0.072 | 0.046 |
|       |          | 50  | 0.010       | 0.436       | 0.045 | –          | 0.028 | 0.028 | 0.063 | 0.033 |
|       |          | 100 | 0.041       | 0.392       | 0.065 | –          | 0.043 | 0.038 | 0.059 | 0.050 |
| \(R = 0.50\) | (0%, 50%) | 20  | 0.027       | 0.117       | 0.038 | 0.051      | 0.027 | 0.029 | 0.059 | 0.061 |
|       |          | 30  | 0.037       | 0.140       | 0.053 | 0.044      | 0.034 | 0.026 | 0.068 | 0.072 |
|       |          | 50  | 0.044       | 0.164       | 0.042 | 0.046      | 0.036 | 0.021 | 0.049 | 0.051 |
|       |          | 100 | 0.055       | 0.253       | 0.040 | 0.031      | 0.050 | 0.019 | 0.050 | 0.048 |
|       | (50%, 50%) | 20  | 0.007       | 0.314       | 0.051 | –          | 0.023 | 0.028 | 0.092 | 0.066 |
|       |          | 30  | 0.010       | 0.388       | 0.046 | –          | 0.032 | 0.036 | 0.089 | 0.065 |
|       |          | 50  | 0.027       | 0.430       | 0.032 | –          | 0.035 | 0.038 | 0.049 | 0.037 |
|       |          | 100 | 0.062       | 0.476       | 0.045 | –          | 0.034 | 0.028 | 0.040 | 0.061 |

Each entry was based on 1000 replicates. WMD_1 and WMD_0 were calculated based on 1500 bootstrap samples. Significant level \(\alpha\) was set to 0.05. – indicates results unavailable because F-\(W_2\) is only applicable to cases when missing values only present in one group.


| Corr | Missing | n | W_{cp} | CLMM | F-W_2 | SR-MW | MW-MW | WMD_s | WMD_o |
|------|---------|---|--------|------|-------|-------|-------|-------|-------|
| R = 0.75 | (0%, 50%) | 20 | 0.156 | 0.187 | 0.227 | 0.150 | 0.136 | 0.245 | 0.276 |
|       |         | 30 | 0.316 | 0.305 | 0.317 | 0.265 | 0.219 | 0.316 | 0.392 |
|       |         | 50 | 0.577 | 0.472 | 0.493 | 0.443 | 0.392 | 0.489 | 0.622 |
|       |         | 100| 0.921 | 0.827 | 0.719 | 0.765 | 0.740 | 0.770 | 0.939 |
|       | (50%, 50%) | 20 | 0.019 | 0.083 | –     | 0.069 | 0.100 | 0.206 | 0.214 |
|       |         | 30 | 0.080 | 0.141 | –     | 0.136 | 0.164 | 0.227 | 0.257 |
|       |         | 50 | 0.244 | 0.255 | –     | 0.232 | 0.280 | 0.241 | 0.402 |
|       |         | 100| 0.604 | 0.570 | –     | 0.472 | 0.574 | 0.412 | 0.741 |
| R = 0.50 | (0%, 50%) | 20 | 0.108 | 0.188 | 0.243 | 0.210 | 0.188 | 0.235 | 0.248 |
|       |         | 30 | 0.255 | 0.347 | 0.331 | 0.315 | 0.273 | 0.349 | 0.382 |
|       |         | 50 | 0.412 | 0.528 | 0.493 | 0.514 | 0.476 | 0.514 | 0.539 |
|       |         | 100| 0.711 | 0.829 | 0.753 | 0.811 | 0.803 | 0.784 | 0.828 |
|       | (50%, 50%) | 20 | 0.024 | 0.120 | –     | 0.075 | 0.112 | 0.214 | 0.220 |
|       |         | 30 | 0.096 | 0.199 | –     | 0.165 | 0.209 | 0.246 | 0.266 |
|       |         | 50 | 0.168 | 0.325 | –     | 0.303 | 0.339 | 0.346 | 0.356 |
|       |         | 100| 0.392 | 0.660 | –     | 0.568 | 0.653 | 0.538 | 0.676 |

Each entry was based on 1000 replicates. WMD_s and WMD_o were calculated based on 1500 bootstrap samples. Significant level \( \alpha \) was set to 0.05. – indicates results unavailable because F-W_2 is only applicable to cases when missing values only present in one group.

inflated type I error rates. WMD_o had the highest power among all methods across all cases. The difference in power between WMD_o and CLMM became larger as the correlation \( R \) increased from 0.50 to 0.75. The difference in power between WMD_o and the methods of F-W_2 and WMD_s became larger as the sample size increased.

4 Examples

4.1 A study on prostate cancer biomarkers

We considered a prostate cancer biomarker study. Data were obtained on 18 patients who underwent surgery for prostate cancer diagnosis from 2011 to 2015 at the University of Kentucky Medical Center and were also consented to provide longitudinal samples of the following biomarkers: PSA, CRP, and IL8. These samples were obtained at the baseline (prior to surgery), 3-month, and 6-month follow-ups, and the biomarker levels were measured as continuous variables. One important question of interest was to compare the pre- and post-surgery levels of those biomarkers. All patients had those biomarkers recorded at the baseline, 61% had missing values at 3-month and 50% had missing values at 6-month due to missed visits or loss of follow-up. Little’s test of MCAR\(^{15}\) using the mcar_test function in the naniar R-package\(^{16}\) has a \( P \)-value of 0.44, which suggests that the MCAR assumption is reasonable.

We applied our proposed method as well as existing methods to compare PSA, CRP, and IL8 levels at 3-month and 6-month follow-ups to the baseline. One patient was identified as an outlier based on the Dixon test\(^{17}\) for each of the three biomarkers, and was excluded from the analysis. Table 5 shows the \( P \)-values based on \( t_{cp}, \) LMM, \( W_{cp}, \) F-W_2, \( \) SR-MW, MW-MW, WMD_s, and WMD_o. The bootstrap method with 1500 bootstrap samples was used for estimating the standard error of \( \hat{D}(\hat{w}_1, \hat{w}_2) \) in the calculation of WMD_s and WMD_o. For the 3-month follow-up versus baseline comparison of PSA measurement, SR-MW and MW-MW were significant. For the 3-month follow-up versus baseline comparison of CRP and IL8 measurements, none of the method yielded a significant result. For the 6-month follow-up versus baseline comparison of PSA measurement, all methods were significant. For the 6-month follow-up versus baseline comparison of CRP measurement, none of the method yielded a significant result. For the 6-month follow-up versus baseline comparison of IL8 measurement, \( t_{cp}, \) LMM, \( W_{cp}, \) and WMD_o were significant while F-W_2, SR-MW, MW-MW, and WMD_s were not significant.

4.2 A study on biomarkers for breast cancer recurrence

He et al.\(^{1}\) evaluated expressions of five key cancer-related proteins, including PAR4, TWIST, pCrk/CrkL, SNAIL, and \( \beta \)-catenin, in paired primary and recurrent tumor tissue specimens from 22 patients that were diagnosed and treated at University of Kentucky Medical Center between 2000 and 2007. Two representative tissue cores were taken from each tumor and included in a tissue microarray construction. Based on immunohistochemistry, each tissue core was assigned an intensity score of 0 (negative), 1 (weak), 2 (moderate), or 3 (strong), and the scores of the two cores for each tumor
### Table 5. Data analysis results for the prostate cancer study.

| Biomarker | Comparison        | \( t_{cp} \) | LMM | \( W_{cp} \) | F-\( W_2 \) | SR-MW | MW-MW | WMD_1  | WMD_o  |
|-----------|-------------------|--------------|-----|--------------|-------------|-------|-------|--------|--------|
| PSA       | 3-month versus baseline | 0.549        | 0.351 | 0.402        | 0.104       | 0.047 | 0.017 | 0.538  | 0.487  |
|           | 6-month versus baseline  | 0.023        | 0.007 | 0.014        | <0.001      | <0.001| <0.001| <0.001 | <0.001 |
| CRP       | 3-month versus baseline | 0.413        | 0.261 | 0.834        | 0.188       | 0.195 | 0.056 | 0.051  | 0.120  |
|           | 6-month versus baseline  | 0.344        | 0.384 | 0.554        | 0.636       | 0.561 | 0.307 | 0.295  | 0.230  |
| IL8       | 3-month versus baseline | 0.615        | 0.739 | 0.800        | 0.953       | 0.849 | 0.970 | 0.731  | 0.778  |
|           | 6-month versus baseline  | 0.013        | 0.015 | 0.024        | 0.566       | 0.374 | 0.122 | 0.756  | 0.006  |

Numbers presented are \( P \)-values. \( P \)-values for WMD_1 and WMD_o were calculated based on 1500 bootstrap samples.
Table 6. Data analysis results for the breast cancer study.

| Protein biomarker | Missing       | W\text{\textsc{cp}} | CLMM | F-W\textsubscript{2} | SR-MW | MW-MW | WMD\textsubscript{s} | WMD\textsubscript{o} |
|-------------------|---------------|---------------------|------|----------------------|-------|-------|----------------------|----------------------|
| PAR4              | (5%, 0%)      | 0.112               | 0.061| 0.073                | 0.091 | 0.040 | 0.041                | 0.043                |
| TWIST             | (5%, 19%)     | 0.063               | 0.038| --                   | 0.058 | 0.030 | 0.008                | 0.016                |
| pCrk/CrkL         | (14%, 0%)     | 0.001               | <0.001| <0.001              | <0.001| <0.001| <0.001              | <0.001              |
| SNAIL             | (14%, 5%)     | 0.634               | 0.419| --                   | 0.538 | 0.348 | 0.355                | 0.484                |
| \(\beta\)-catenin| (5%, 5%)      | 0.255               | 0.317| --                   | 0.987 | 0.435 | 0.256                | 0.235                |

The missing proportions in parenthesis are for primary and recurrent tumor tissue samples, respectively. Other numbers presented are \(P\)-values. *: \(P\)-values based on the \(W\text{\textsc{cp}}\) test have been previously reported by He et al.\cite{1} \(P\)-values for WMD\textsubscript{s} and WMD\textsubscript{o} were calculated based on 1500 bootstrap samples. – indicates results unavailable because F-W\textsubscript{2} is only applicable to cases when missing values only present in one group.

5 Discussion

We have developed a class of test statistics based on the difference in weighted sample means between two groups for partially paired data. This class of tests is very general because it allows different weights for observations from different groups as well as from complete and incomplete pairs within the same group. We have further derived the optimal weight for this class of tests. Simulation results demonstrate that our test with the optimal weight is more powerful than existing methods.

Our tests are free of any model or distributional assumptions. The calculation of tests is direct and does not require any modeling of the correlation between the paired outcomes. In addition, the \(P\)-value is calculated based on the asymptotic normality of the test statistics which does not require any assumption on the distribution of the data. As they do not impose any distributional assumption, our tests provide a unified approach to deal with both continuous and discrete interval data.

There are two strategies to develop tests specifically for partially paired data. The first strategy considers weighted linear combinations of test statistics comparing paired data and test statistics comparing unpaired data.\cite{3,4,5,6,7} The second strategy first constructs estimators of the population mean difference between the two groups, and then derives test statistics based on the estimators.\cite{8} An advantage of the second strategy is that in addition to the \(P\)-value, it provides estimates for the population mean difference between groups, quantifying the magnitude of the difference in a meaningful way. We take the second strategy in this article and develop a general class of estimators for the population mean difference. We further obtain the efficient estimator by deriving the optimal weight that minimizes variances of estimators within the class.

Like existing methods,\cite{3,4,5,6,7,9,10} our tests require the MCAR assumption, which is a reasonable assumption in practice as shown in the two cancer biomarker studies in Section 4. In addition, our tests can be extended to the situation of missing at random given covariates by weighting each observation by the inverse of the estimated propensity score, that is, the probability of the corresponding outcome is observed given the covariate values of the corresponding individual. The asymptotic variance and the optimal weight of the extended tests will need to be derived.

Acknowledgements

We thank Drs Tom Tucker, Rina Plattner, Vivek Rangnekar, Binhua Zhou, and Chunming Liu for sharing their data on protein biomarkers for breast cancer recurrence. We also thank Drs Lili Tian and Qianya Qi for providing R codes for their F-W\textsubscript{2} method.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by the National Institutes of Health (1R03CA179661-01A1, 5P20GM103436-15) and the Biostatistics and Bioinformatics Shared Resource Facility of the University of Kentucky Markey Cancer Center (P30CA177558). The prostate cancer biomarker study was supported by the National Cancer Institute (5R01CA143428).

ORCID iD
Li Chen https://orcid.org/0000-0001-9719-7652

Supplemental material
Supplemental materials for this article are available online.

References
1. He C, Plattner R, Rangnekar V, et al. Potential protein markers for breast cancer recurrence: a retrospective cohort study. Cancer Causes Control 2019; 30: 41–51.
2. Little RJ and Rubin DB. Statistical analysis with missing data. 793. Hoboken, NJ: John Wiley & Sons, 2019.
3. Dubnicka SR, Blair RC and Hettmansperger TP. Rank-based procedures for mixed paired and two-sample designs. J Mod Appl Stat Methods 2002; 1: 6.
4. Brunner E, Domhof S and Langer F. Nonparametric analysis of longitudinal data in factorial experiments. New Y ork: Wiley, 2002.
5. Koniedtschke F, Harrar SW, Lange K, et al. Ranking procedures for matched pairs with missing data-asymptotic theory and a small sample approximation. Comput Stat Data Anal 2012; 56: 1090–1102.
6. Fong Y, Huang Y, Lemos MP, et al. Rank-based two-sample tests for paired data with missing values. Biostatistics 2018; 19: 281–294.
7. Bhoj DS. Testing equality of means of correlated variates with missing observations on both responses. Biometrika 1978; 65: 225–228.
8. Martinez-Camblor P, Corral N and Maria de la Hera J. Hypothesis test for paired samples in the presence of missing data. J Appl Stat 2013; 40: 76–87.
9. Kuan PF and Huang B. A simple and robust method for partially matched samples using the p-values pooling approach. Stat Med 2013; 32: 3247–3259.
10. Qi Q, Yan L and Tian L. Testing equality of means in partially paired data with incompleteness in single response. Stat Methods Med Res 2019; 28: 1508–1522.
11. Qi Q, Yan L and Tian L. Analyzing partially paired data: when can the unpaired portion(s) be safely ignored? J Appl Stat 2022; 49: 1402–1420.
12. Pinheiro J, Bates D, DebRoy S, et al. nlme: Linear and nonlinear mixed effects models. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://CRAN.R-project.org/package=nlme. R package version 3.1-153.
13. Fong Y. robustranks: robust rank-based tests. R Foundation for Statistical Computing, Vienna, Austria, 2020. https://cran.r-project.org/web/packages/robustranks/index.html. R package version 2019.9-10.
14. Christensen RHB. A tutorial on fitting cumulative link mixed models with clmm2 from the ordinal package, 2019. https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2/tutorial.pdf.
15. Little RJ. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988; 83: 1198–1202.
16. Tierney N, Cook D, McBain M, et al. Data structures, summaries, and visualisations for missing data. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://CRAN.R-project.org/package=nanir. R package version 0.6.1.
17. Komsta L. outliers: Tests for outliers. R Foundation for Statistical Computing, Vienna, Austria, 2015. https://CRAN.R-project.org/package=nanir. R package version 0.6.1.
18. Van der Vaart AW and Wellner J. Weak convergence and empirical processes: with applications to statistics. New Y ork: Springer Science & Business Media, 1996.
19. Van der Vaart AW. Asymptotic statistics. 3. New Y ork: Cambridge University Press, 1998.

Appendix
Appendix A. Proof of Theorem 1
In this section, we show the asymptotic normality of the estimator \( \hat{D}(\hat{w}_1, \hat{w}_2) \) using modern empirical process theory. Let \( P_n \) and \( P \) denote the empirical measure and the distribution under the true model, respectively. For a measurable function \( f \) and measure \( Q \), the integral \( \int fdQ \) is abbreviated as \( Qf \). We consider the following three situations.

Situation 1: Missing values present in both groups.
First,

\[
\sqrt{n} \left\{ \hat{D}(\hat{\omega}_1, \hat{\omega}_2) - (\mu_1 - \mu_2) \right\} = \frac{P_{n} f_1}{P_{g_1}} - \frac{P_{f_1}}{P_{g_1}} + \frac{P_{f_2}}{P_{g_1}} - \frac{P_{f_3}}{P_{g_1}} + \sqrt{n} \left\{ \frac{P_{n} f_2}{P_{g_2}} - \frac{P_{f_2}}{P_{g_2}} + \frac{P_{f_3}}{P_{g_2}} - \frac{P_{f_4}}{P_{g_2}} \right\}
\]

Under MCAR assumption, the first term of (4) can be written as

\[
\frac{P_{n} f_1}{P_{g_1}} - \frac{P_{f_1}}{P_{g_1}} = \frac{P_{n} f_2}{P_{g_2}} - \frac{P_{f_2}}{P_{g_2}} = \frac{P_{n} f_3}{P_{g_2}} - \frac{P_{f_3}}{P_{g_2}} = \frac{P_{n} f_4}{P_{g_2}} - \frac{P_{f_4}}{P_{g_2}}
\]

where

\[
\begin{align*}
&f_1 = w_1 Y_1 R_1 R_2, \quad g_1 = R_1 R_2 \\
&f_2 = (1 - w_1) Y_1 R_1 (1 - R_2), \quad g_2 = R_1 (1 - R_2) \\
&f_3 = w_2 Y_2 R_1 R_2, \quad g_3 = g_1 = R_1 R_2 \\
&f_4 = (1 - w_2) Y_2 R_2 (1 - R_1), \quad g_4 = R_2 (1 - R_1)
\end{align*}
\]

Since \( P_{n} f_i / P_{g_i} \rightarrow P f_i / P g_i, P f_i < \infty \) and \( g_i \) are bounded monotonic functions, there exists a \( P \)-Donsker class \( F \) such that \( P^n (g_{n, f_i} / P_{g_i} \in F) \rightarrow 1 \) based on Example 2.10.27 of Van Der Vaart and Wellner. In addition, it is easy to verify that \( f_i, g_i \) satisfy the other two conditions in Lemma 1. Therefore, based on Lemma 1, we obtain that the first term of (4) is equal to \( \sqrt{n} (P_n - P) \xi + o_p(1) \), where

\[
\xi = \left\{ \frac{f_1}{P_{g_1}} - \frac{g_1 P f_1}{(P_{g_1})^2} + \frac{f_2}{P_{g_2}} - \frac{g_2 P f_2}{(P_{g_2})^2} - \frac{f_3}{P_{g_3}} - \frac{g_3 P f_3}{(P_{g_3})^2} + \frac{f_4}{P_{g_4}} - \frac{g_4 P f_4}{(P_{g_4})^2} \right\}
\]

The second term of (4) can be written as \( \sqrt{n} (\hat{\omega}_1 - w_1) (\tilde{Y}_1 - \hat{Y}_1) - \sqrt{n} (\hat{\omega}_2 - w_2) (\tilde{Y}_2 - \hat{Y}_2) \), and thus is \( o_p(1) \) by the MCAR assumption and the condition that \( \sqrt{n} (\hat{\omega}_1 - w_1) \) and \( \sqrt{n} (\hat{\omega}_2 - w_2) \) are \( O_p(1) \). Therefore, \( \sqrt{n} (\hat{D}(\hat{\omega}_1, \hat{\omega}_2) - (\mu_1 - \mu_2)) \) is asymptotically normal with mean 0 and variance

\[
\begin{align*}
\sigma^2_j(w_1, w_2) &= \text{Var}(\xi) \\
&= \text{Var} \left\{ \frac{w_1 R_1 R_2}{p_{12}} + \frac{(1 - w_1) R_1 (1 - R_2)}{p_1 - p_{12}} \right\} (Y_1 - \mu_1) \\
&- \text{Var} \left\{ \frac{w_2 R_1 R_2}{p_{12}} + \frac{(1 - w_2) R_2 (1 - R_1)}{p_2 - p_{12}} \right\} (Y_2 - \mu_2) \\
&= E \left\{ \frac{w_1 R_1 R_2}{p_{12}} + \frac{(1 - w_1) R_1 (1 - R_2)}{p_1 - p_{12}} \right\}^2 \text{Var}(Y_1) \\
&+ E \left\{ \frac{w_2 R_1 R_2}{p_{12}} + \frac{(1 - w_2) R_2 (1 - R_1)}{p_2 - p_{12}} \right\}^2 \text{Var}(Y_2)
\end{align*}
\]
Lemma 1.\[\] When missing values present only in Group 2, we have \(w_1 = 1, p_2 = 1, p_{12} = p_1,\) and \(f_2 = 0.\) Then the first term of (4) can be written as
\[
-2E \left( \frac{w_1 R_1 R_2}{p_{12}} + \frac{(1 - w_1) R_1 (1 - R_2)}{p_1 - p_{12}} \right) \left( \frac{w_2 R_1 R_2}{p_{12}} + \frac{(1 - w_2) R_2 (1 - R_1)}{p_2 - p_{12}} \right) \right \} \text{Cov}(Y_1, Y_2)
\]
\[
= \left\{ \frac{w_2^2}{p_{12}} + \frac{(1 - w_2)^2}{p_1 - p_{12}} \right \} \sigma_2^2 + \left\{ \frac{w_2^2}{p_{12}} + \frac{(1 - w_2)^2}{p_2 - p_{12}} \right \} \sigma_2^2 = \frac{2w_1 w_2 \rho_1 \sigma_2}{p_{12}}
\]

Situation 2: Missing values present only in group 1.
When missing values present only in Group 1, we have \(w_1 = 1, p_2 = 1, p_{12} = p_1,\) and \(f_2 = 0.\) Then the first term of (4) can be written as
\[
\sqrt{n} \left\{ \frac{p_{n f}}{p_{n g}} \right\} - \sqrt{n} \left\{ \frac{p_{n f}}{p_{n g}} \right\} - \sqrt{n} \left\{ \frac{p_{n f}}{p_{n g}} \right\} - \sqrt{n} \left\{ \frac{p_{n f}}{p_{n g}} \right\}
\]

Similar to Situation 1, we obtain \(\sqrt{n}(\hat{D}(\hat{w}_1, \hat{w}_2) - (\mu_1 - \mu_2))\) is asymptotically normal with mean 0 and variance
\[
\sigma_2^2(w_1, w_2) = \text{Var} \left\{ \frac{f_1}{p_{0 g}} - \frac{g_1 p_{f1}}{(p_{0 g})^2} \right\} - \text{Var} \left\{ \frac{f_3}{p_{0 g}} - \frac{g_3 p_{f3}}{(p_{0 g})^2} \right\} + \text{Var} \left\{ \frac{f_4}{p_{0 g}} - \frac{g_4 p_{f4}}{(p_{0 g})^2} \right\}
\]

\[
= \text{Var} \left\{ \frac{R_1 R_2}{p_1} (Y_1 - \mu_1) - \frac{R_1 R_2}{p_1} (Y_2 - \mu_2) \right\}
\]

\[
= \frac{1}{p_1} + \left\{ \frac{w_2^2}{p_{12}} + \frac{(1 - w_2)^2}{1 - p_1} \right \} \sigma_2^2 = \frac{2w_1 \rho_1 \sigma_2}{p_{12}}
\]

Situation 3: Missing values present only in group 2.
When missing values present only in Group 2, we have \(w_2 = 1, p_1 = 1, p_{12} = p_2,\) and \(f_4 = 0.\) Similarly, we obtain \(\sqrt{n}(\hat{D}(\hat{w}_1, \hat{w}_2) - (\mu_1 - \mu_2))\) is asymptotically normal with mean 0 and variance
\[
\sigma_2^2(w_1, w_2) = \left\{ \frac{w_2^2}{p_2} + \frac{(1 - w_2)^2}{1 - p_2} \right \} \sigma_2^2 + \frac{2w_1 \rho_1 \sigma_2}{p_2}
\]

Lemma 1. For any measurable functions \(f\) and \(g,\) if there exists a \(P\)-Donsker class \(F\) such that \(P^n \left\{ \frac{g^n f}{(P_{g})^2} \in F \right\} \rightarrow 1,\)
\[
P \left\{ \frac{g^n f}{(P_{g})^2} \right\} \rightarrow 0, and \frac{g^n f}{(P_{g})^2} \in L_2(P),\) then
\[
\sqrt{n} \left\{ \frac{P_{n f}}{P_{n g}} \right\} = \sqrt{n}(P_{n} - P) \frac{f}{P_{g}} - \frac{g^n f}{(P_{g})^2} + o_p(1)
\]

Proof.
\[
\sqrt{n} \left\{ \frac{P_{n f}}{P_{n g}} \right\} - \frac{P_{f}}{P_{g}} \right\}
\]

\[
= \sqrt{n}(P_{n} - P) \frac{f}{P_{g}} - \frac{g^n f}{(P_{g})^2}
\]

\[
= \sqrt{n}(P_{n} - P) \left\{ \frac{f}{P_{g}} - \frac{g^n f}{(P_{g})^2} \right\} - \sqrt{n}(P_{n} - P) \left\{ \frac{g^n f}{(P_{g})^2} - \frac{g^n f}{(P_{g})^2} \right\}
\]

By the conditions and Lemma 19.24 of Van der Vaart,\(^{19}\) the second term of 5 is \(o_p(1).\)
Appendix B. Derivation of the optimal weight

Situation 1: Missing values present in both groups.

The optimal weight is the \((w_1, w_2)\) values that minimizes the asymptotic variance \(\sigma_D^2(w_1, w_2)\). To find the minimization point of \(\sigma_D^2(w_1, w_2)\), we solve the first derivative equations \(\partial \sigma_D^2(w_1, w_2)/\partial w_1 = 0\) and \(\partial \sigma_D^2(w_1, w_2)/\partial w_2 = 0\) and obtain the root as

\[
\begin{align*}
    w_1^* &= \frac{p_{12} \sigma_1^2 + \rho \sigma_1 \sigma_2 p_{12}(p_1 - p_{12})}{p_{12} \sigma_1^2 - \rho^2 \sigma_1^2(p_1 - p_{12})}, \\
    w_2^* &= \frac{p_{12} \sigma_2^2 + \rho \sigma_1 \sigma_2 p_{12}(p_2 - p_{12})}{p_{12} \sigma_2^2 - \rho^2 \sigma_2^2(p_1 - p_{12})}.
\end{align*}
\]

Note that our weight is defined under the constraints that \(0 \leq w_1 \leq 1\) and \(0 \leq w_2 \leq 1\). If \((w_1^*, w_2^*)\) falls in this squared region, it is the optimal weight. Otherwise, the minimization of the variance will be reached on the boundary of this squared region. Since the boundary consists of four lines, we consider the following set \(A\) as a collection of the minimization point on each of the four lines,

\[
A = \left\{ \left( \begin{array}{c} 0 \\ \frac{p_{12}}{p_2} \end{array} \right), \left( \begin{array}{c} p_{12} \\ \frac{0}{p_1} \end{array} \right) \right\}
\]

where \(g(z) = \begin{cases} 0 & \text{if } z < 0 \\ z & \text{if } 0 \leq z \leq 1 \\ 1 & \text{if } z > 1 \end{cases} \)

Then optimal weight is chosen as the point in \(A\) that minimizes the variance. In summary, the optimal weight is

\[
(w_1^{opt}, w_2^{opt}) = \begin{cases} (w_1^*, w_2^*) & \text{if } 0 \leq w_1^* \leq 1, 0 \leq w_2^* \leq 1 \\ \text{argmin } \sigma_D^2(w_1, w_2) & \text{otherwise} \end{cases}
\]

Situation 2: Missing values present only in group 1.

When missing values only present in group 1, we have \(p_2 = 1, p_{12} = p_1,\) and \(w_1 = 1\). The optimal weight of \(w_2\) is the point that minimizes \(\sigma_D^2(1, w_2)\). We solve \(d\sigma_D^2(1, w_2)/dw_2 = 0\) and obtain the root as \(w_2^* = \frac{p_1 \sigma_1 + \rho \sigma_1 (1-p_1)}{\sigma_2}\). Note that \(0 \leq w_2 \leq 1\). If \(0 \leq w_2^* \leq 1\), the optimal weight is \(w_2^*\). If \(w_2^* < 0\), the optimal weight is 0. If \(w_2^* > 1\), the optimal weight is 1.

In summary, the optimum weight of \(w_2\) is \(w_2^{opt} = g(w_2^*)\). Similarly, when missing values only present in group 2, we have \(p_1 = 1, p_{12} = p_2,\) and \(w_2 = 1\). The optimum weight of \(w_1\) is \(w_1^{opt} = g(w_1^*)\), where \(w_1^* = \frac{p_2 \sigma_2 + \rho \sigma_2 (1-p_2)}{\sigma_1}\).

As shown in Situation 1, \(\sigma_D^2(w_1, w_2)\) is minimized at \(w_1^* = \frac{p_1 \sigma_1 + \rho \sigma_1 (1-p_2)}{\sigma_1}\). Under the constraint that \(0 \leq w_2 \leq 1\), the optimum weight is \(w_1^{opt} = g(w_1^*)\).

When missing values present only in group 1, we have

\[
\begin{align*}
\sigma_D^2(w_1, w_2) &= \frac{1}{p_{12}} \sigma_1^2, \\
&= \frac{w_2^2}{p_1(1-p_1)} - 2w_2 \frac{p_1 \sigma_1^2}{p_1(1-p_1)} + C_4.
\end{align*}
\]
where $C_4$ is a constant. Then $\sigma_D^2(w_1, w_2)$ is minimized at

$$w_2^* = \left\{ \frac{p_1 \sigma_2^2}{p_1 (1 - p_1)} + \frac{\rho \sigma_1 \sigma_2}{p_1} \right\} \bigg/ \left\{ \frac{\sigma_2^2}{p_1 (1 - p_1)} \right\}$$

$$w_2^* = \frac{p_1 \sigma_2 + \rho \sigma_1 (1 - p_1)}{\sigma_2}$$

**Situation 3: Missing values present only in group 2.**

Similarly, when missing values present only in group 2, we obtain the optimal weight as

$$w_1^* = \frac{p_2 \sigma_1 + \rho \sigma_2 (1 - p_2)}{\sigma_1}$$