I’ve lost my ID(s)! Can I still test for a difference in means? A comparison of methods for partially matched data.

Raymond Pomponioa\*, Ryan Petersona and A. N. Author(s)b

aDepartment of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado-Denver Anschutz Medical Campus, Aurora, CO; bDepartment, University, City, Country

\*Raymond Pomponio, Raymond.Pomponio@cuanschutz.edu

Provide short biographical notes on all contributors here if the journal requires them.

I’ve lost my ID(s)! Can I still test for a difference in means? A comparison of methods for partially matched data.

Type or paste your abstract here as prescribed by the journal’s instructions for authors.

Keywords: t-test, unmatched data, missing identifiers, record matching, paired data

# Introduction

It is common in biostatistics to test the equality of two means. In study designs with two independent, normally distributed samples, Student’s *t*-test is appropriate. In study designs with paired samples, at least two approaches are available. First, the one-sample *t*-test can be applied to the unit-level differences between the first timepoint and the second timepoint, i.e., the change scores between timepoints. Second, although less popular, the two-sample *t*-test modified for correlated data can be used, assuming the correlation between timepoints is constant and estimable; see [1]. This work adapts the latter version of the *t*-test to address the challenge of paired samples with missing unit-level identifiers, a type of data we refer to as paired but unmatched.

Several methods have been developed to test the equality of means in study designs that include a hybrid of independent samples and paired samples; in biostatistics such data often arise from dropout between timepoints. These methods involve tests based on modified maximum likelihood, multiple imputation procedures, or pooling statistics across paired and independent subsets of data. See, for example [2] [3] [4]. This type of data has been referred to as partially paired, or partially correlated.

However, aside from dropout, a different mechanism of missingness can lead to unmatched data, that is, when unit-level identifiers have been lost or withheld for anonymity. For example, an employer may survey its employees both before and after an intervention, but for confidentiality reasons, the employer may make the collection of identifiers optional. More generally, anonymity (or pseudo-anonymity, where participants optionally volunteer identifying information) may be advised or required for legal or ethical reasons. However, unmatched data in this setting does not confer independence, in fact we will often expect high degrees of unmeasurable correlation. This type of study design, with paired but unmatched samples, has received relatively less attention. In this work, we aim to provide guidance on testing the equality of means in such datasets.

Unmatched data present a challenge for testing the equality of means. The one-sample *t*-test cannot be used, due to the inability to match samples and calculate change scores. Using Student’s two-sample ­*t*-test assumes samples are uncorrelated, which is an unrealistic assumption of repeated measures in biostatistics. One available alternative involves calculating the minimum possible correlation coefficient given the observed data; this ‘worst-case’ value can be used as the estimated correlation in the two-sample *t*-test modified for correlated data. However, this approach yields a maximally conservative test and is poorly powered.

In cases where even a small number of samples are matched, an opportunity exists to use the matched pairs to inform an estimate of the correlation between all pairs. We focus on the case where a subset of samples can be matched, and we refer to this type of data as partially matched. Future work may devote attention to entirely unmatched data, though we note the difficulty in dealing with that type of data lies in estimating a correlation without any paired samples.

[Figure 1 here]

Our study is motivated by our application, a dataset in which survey responses were collected from 149 physicians both before and after an educational intervention. Of those, 69 (46%) physicians were assigned to a ‘control’ group. The collection of identifiers was made optional through survey fields in which one could enter the first letter of one’s paternal last name, the first letter of one’s maternal last name, the last two digits of one’s phone number, and the last two numbers of one’s DNI. Within the intervention group, 9 (11%) survey responses were matched on the optional identifier. In the control group, 10 (14%) survey responses were matched. The application provided an example of partially matched data. We sought to identify a well powered test for the equality of means between the pre- and post- intervention timepoints for the intervention and control groups separately, while controlling Type I error rate at a nominal level.

# Methods

## Notation / Definitions

Let denote the true, unobserved correlation between pre- and post-intervention scores, denoted by **X** and **Y**, respectively. We draw *n* paired samples , each of which represents a unit-level observation. Because data are paired but partially matched, we only observe matched pairs for *m* samples, and the remaining *n – m* paired samples are unmatched. For unmatched samples we cannot identify which observation in corresponds to its pair in . Our aim is to evaluate the difference in means, , between **X** and **Y**.

Assuming **X** and **Y** are normally distributed, and the variance of **X** and **Y** are equal, we have the following formula for the two-sample *t*-test modified for correlated data, from [1]:

(1)

Note that the above statistic is equivalent to when *t* is the statistic from Student’s *t*-test with equally sized samples. Smaller values of result in higher standard error for the difference in means, and thus a test statistic that is more conservative. Conversely, larger values of result in lower standard error for the difference in means, and thus a test statistic that is less conservative. Also note that applying Student’s *t*-test in unmatched data is equivalent to setting in the above equation and thus, assuming independence of **X** and **Y**.

In applied scenarios this test requires an estimate of , which we will denote *r*. This additional requirement may explain the relative unpopularity of the test compared to the one-sample *t*-test. However, simulations have demonstrated that even with this requirement, the modified *t*-test can improve power in modestly sized datasets (e.g., *n* = 25), compared to the one-sample *t*-test, while controlling the Type I error rate near the nominal level of 0.05; see [1]. The improvement in power is due to the greater degrees of freedom in the modified *t*-test statistic, compared to the one-sample *t*-test statistic (2*n* – 2 versus *n* – 1 degrees of freedom, respectively).

In partially matched data, the challenge of applying the modified *t*-test lies in finding a suitable estimator of the correlation, *r*, despite observing only a subset of matched samples in the dataset. In the next section, we considered several candidate estimators for correlation. Each is denoted by , and the modified *t*-test corresponding to each estimator will be denoted by .

## Estimators of correlation

### Maximally conservative estimator

Arguably, the maximally conservative estimator for is -1, which yields the maximal standard error of the mean difference estimate. For real data sets, this correlation is only possible if every sample at the first timepoint has a single identical value at the second timepoint, which is rare. One can instead calculate the minimum possible correlation coefficient for any paired dataset by determining the correlation of *sorted* outcomes within each time.

Let *j* denote the index of the samples sorted in ascending order. Then the following formula defines the maximally conservative estimate of the correlation:

(2)

Note that calculation of the numerator in the above equation requires sorting observed samples from **X** in ascending order while sorting observed samples from **Y** in descending order. Although expected to have considerable bias towards underestimating the true correlation, the above estimator is calculable even when data are completely unmatched. This is one advantage of . The remaining estimators of correlation require at least one matched sample to be calculated.

### Pearson correlation of the matched samples

An intuitive correlation estimate in the presence of partially matched data uses the matched samples and ignores the matched samples. The following formula gives the Pearson correlation coefficient for matched samples:

(3)

The above estimator requires at least two matched samples to be calculable. When exactly two matched samples are available, the estimator will return a value of either 1 or -1. With three or more matched samples, the estimator will be substantially less biased than the maximally conservative estimator, .

### Quantile estimator of the matched samples

Recognizing that conservatism in estimating might be desirable for hypothesis testing, yet the degree of conservativism in might be too extreme, we hypothesized that a quantile lower-bound estimate of might yield additional power while maintaining appropriate Type I error rate control.

The Fisher transformation of a Pearson correlation can be used to obtain an approximately normally distributed quantity and thus, a confidence interval for . Our estimator is constructed from the lower bound of an 80% confidence interval for the one-sided hypothesis test of a correlation between matched samples. Let be the standard error of the Fisher-transformed correlation coefficient. Then the following formula gives the quantile estimator of the correlation of *m* matched samples:

(4)

Where is the critical value chosen based on a desired confidence level.

Our rationale behind the quantile estimator was to account for the uncertainty in when computing the two-sample *t*-test modified for correlated data. This uncertainty exists even when data are entirely matched. The modified *t*-test using the sample correlation was shown to achieve the same nominal significance level that the paired-samples *t*-test achieved in sample sizes of 100 or greater; see [1]. In smaller samples, an estimator of correlation that is biased toward underestimating the true correlation will result in a more conservative test. In theory, there exists some for which the Type I error is controlled at any desired level. In practice, we selected the 20th quantile as a semi-conservative estimator that we expected to generalize to a variety of datasets, so the critical value of was 0.842. We refer to this estimator as the 20th quantile estimator of the correlation of the matched samples, or . The quantile estimator requires at least four matched samples to be calculable, due to its basis on the confidence interval calculation, in which the denominator of the standard error is .

Note that, as with other estimators, . However, for any finite sample size, . Therefore, will be more conservative than other approaches yet it possesses the same asymptotic properties of unbiasedness.

### Bayesian estimator of the matched samples

One undesirable property of the quantile estimator is its bias toward underestimating the true correlation, even when sample sizes may be large enough to warrant less conservatism. A Bayesian approach permits some conservatism in the form of shrinkage towards a prior mean of , while also accompanying evidence from observed data.

Several Bayesian estimators of correlation are available for bivariate normal data with known variances and small sample sizes; see [5]. We adapted the estimator based on the posterior mean assuming an arcsine prior by first standardizing the data to comply with the assumptions of zero means and unit variances. The arc-sine prior is equivalent to a generalized beta (2, 1, 0.5, 0.5) prior.

Let and denote standardized samples. Let , , and denote quantities obtained from the *m* matched samples. Then the following formula gives the posterior mean correlation assuming an arc-sine prior:

(5)

Note that standardizing the data is achieved using both matched and unmatched samples. However, to have a non-zero estimate, the above estimator requires at least one matched sample.

### EM Algorithm estimator of correlation

The EM algorithm offers a general approach to obtaining maximum likelihood estimates under incomplete data scenarios; see [6]. In the case of partially matched data, the quantity is incompletely observed due to the inability to match all samples and calculate the cross product. This quantity is a sufficient statistic for the bivariate normal distribution.

The missing quantity can be partitioned into an observed quantity, , and an unobserved quantity assuming the first *m* samples are matched. The latter quantity has a defined expectation when the parameters of the bivariate normal distribution are known (e.g., ):

(4)

Note that maximum likelihood estimates of can be obtained regardless of whether data are matched or unmatched. Our implementation of the EM algorithm iteratively updates the expectation of and the maximum likelihood estimate of until convergence. We provide further details in Appendix A1. At least one matched sample is required for valid estimates of the correlation.

## Simulation study

Data were simulated from bivariate normal distributions and from bivariate ordinal distributions. The latter were obtained by ‘binning’ values of the normal distribution to derive a right-skewed ordinal variable. Bins were selected to yield a seven-level ordinal distribution to mimic the empirical distributions of the outcomes in our application.

We specified the following values of true correlation to simulate data: -0.9, -0.5, -0.25, 0, 0.25, 0.5, and 0.9. For the ordinal distribution, the true correlation was not preserved due to the transformation from a continuous space to an ordinal space. However, we estimated the following effective correlations for ordinal data, based on simulation: -0.58, -0.36, -0.19, 0, 0.20, 0.43, and 0.84, respectively for each successive value of correlation. We provide further details in Appendix A2.

Samples sizes of 10, 20, 50, 100, and 200 were simulated. The proportion of matched samples varied from 0 to 1, to mimic varying conditions of partially matched data. True mean differences of 0, 0.25, and 0.5 standardized units were simulated. We fixed the variances of both variables to one, although in ordinal data we noted this resulted in effective variances of 2.2.

We simulated 10,000 datasets for every combination of the above simulation parameters (resulting in 16.8 million datasets). For each of the five estimators of correlation mentioned previously, we monitored bias and mean squared error compared to the true correlation (or the effective correlation, in ordinal datasets). We computed standard errors of the difference in means. We derived the two-sample *t*-test modified for correlated data using each of the estimators and monitored Type I error rates (when the true mean difference was zero) and power (when the true mean difference was 0.25 or 0.5 standardized units); in all scenarios we used a nominal Type I error rate of . We compared estimators to one another based on resulting metrics of bias, mean squared error, Type I error rates, and power. For reference, we also included Student’s *t-*test which assumes independence, as well as an ‘oracle’ approach in which the true correlation was known.

## Application

Using the above estimators, we applied modified *t*-tests to evaluate the difference in means between timepoints for the intervention and control groups separately. We computed standard errors, 95% confidence intervals for the difference in means, and p-values for each of the candidate methods. Since some physicians were lost to follow-up, we modified the maximally conservative test and the test based on the EM algorithm estimator to address imbalances in the number of responses between pre- and post- intervention. We provide further details in Appendix A4.

# Results

## Simulation study

When the number of matched samples was small (i.e., less than ten), certain estimators yielded invalid correlations in some simulated datasets. We considered these to be estimation failures, and such failures occurred most in ordinal datasets, where the potential for a subset of matched samples to lack any variance was higher than in continuous datasets. The failure rates of estimators were typically below 5%; failure rates across simulation settings are shown in Table A5. For example, when the generating distribution was ordinal, the overall sample size was ten, and the number of matched samples was two, the Bayesian estimator failed to yield a valid correlation in 3.6% of datasets, while the EM algorithm estimator and the maximally conservative estimator failed at rates of 0.01% and 0.1%, respectively. The Pearson estimator and the 20th Quantile estimator could not estimate a correlation between with only two matched samples; this was previously mentioned as limitations to those estimators. However, when the generating distribution was ordinal, the overall sample size was twenty, and the number of matched samples was four, both and failed at a common rate of 16%.

[Figure 2 here]

The Pearson estimator, the EM algorithm estimator, and the Bayesian estimator all exhibited bias toward overestimating the correlation when negative and underestimating the true correlation when positive; the three estimators were unbiased when the true correlation was zero. Of all five estimators, was generally least biased across values of correlation. The EM algorithm was slightly more biased than , though when the data were 100% matched the two estimators yielded identical correlations. Lastly, the 20th Quantile estimator was biased toward underestimating the true correlation, though the bias was most severe at true correlation values of 0.25 and 0.5. At a true correlation of 0.9, bias in was substantially reduced. The use of the ordinal generating distribution altered these results somewhat, with exhibiting more bias than when the true correlation was negative, though was still the least biased.

[Figure 3 here]

Among the Pearson estimator, the EM algorithm estimator, and the Bayesian estimator, there was generally an inverse relationship between bias and mean squared error (MSE). Across all values of correlation, had the highest MSE among these three estimators. For values of correlation between -0.5 and 0.5, had the lowest MSE, while exhibited MSE between that of and . The 20th Quantile estimator exhibited asymmetric MSE across values of correlation, with lower MSE when the true correlation was negative, and higher MSE when the true correlation was positive. The maximally conservative estimator demonstrated substantially high MSE for any correlation greater than -0.25; this makes sense since will generally underestimate the true correlation and will be a poor estimator when the true correlation is positive. The use of the ordinal generating distribution did not substantially alter these trends.

[Figure 4 here]

Standard errors generally decreased as the true correlation increased (this was expected based on the formula for the denominator of the modified *t*-test), except for the maximally conservative estimator, which yielded consistently high standard errors across all values of correlation. The 20th Quantile estimator yielded standard errors that were generally between those of and the remaining estimators. We noted that the Bayesian estimator yielded higher standard errors and was more conservative than and when the true correlation was positive, especially when the number of matched samples was small. As the number of matched samples increased beyond ten, there was little difference in the standard errors between , , and . The use of the ordinal generating distribution did not substantially alter these trends.

[Figure 5 here]

In normally distributed datasets with four matched samples and a sample size of 20, the modified *t*-test based on the 20th Quantile estimator, , achieved nearly nominal Type I error rate control across all values of correlation. The test based on the Pearson estimator, , demonstrated increasingly worse Type I error rates as the true correlation increased. The test based on the EM algorithm estimator, , demonstrated generally consistent, but inflated Type I error rates across all values of correlation, though it was not as pronounced as Pearson’s estimator. The test based on the Bayesian estimator, , demonstrated mixed results, with more Type I error when the true correlation was negative and nearly nominal Type I error control when the true correlation was positive. The test based on the maximally conservative estimator, , was expectedly conservative across all values of correlation and particularly conservative when the true correlation was positive. We noted that Student’s two-sample *t*-test demonstrated optimal Type I error control only when the true correlation was zero, as expected; Student’s *t*-test was increasingly conservative as the true correlation increased.

In ordinal datasets with four matched samples and a sample size of 20, all approaches except the maximally conservative test demonstrated inflated Type I error rates, compared to the identical scenario with normally distributed datasets. Among the candidates, demonstrated the most consistent Type I error control across values of correlation and only modest Type I error rate inflation, with an error rate between 0.069 and 0.101. This consistency in Type I error across correlation values was maintained even as the number of matched samples was reduced to two.

[Figure 6 here]

We observed a general tradeoff between Type I error control and power. That is, the most conservative tests in terms of Type I error were generally the least powered to detect true mean differences. All tests based on our candidate estimators demonstrated greater power as the true correlation increased, except for the maximally conservative test which displayed less power as the true correlation increased. Similarly, Student’s two-sample *t*-test exhibited less power with increasing correlation, except when the difference was large (i.e., 0.5 standardized units) and the sample size was large (i.e., 50 or greater). Thus, the tests with the greatest power with four matched samples and a sample size of 20 were , , , and , respectively (in descending order). We noted that the power afforded by an ‘oracle’ test, in which the true correlation was known, was generally not as conservative as , and was similar in power to and when the true correlation was positive. The use of the ordinal generating distribution did not substantially alter these trends.

In larger samples (i.e., 100 or greater) with 20-50% of samples matched, the results of , , and generally mimicked one another, with nearly optimal Type I error control and similar power. In such samples, was relatively more conservative than the preceding tests. This suggests the choice of the 20th Quantile may be more suitable for small samples. We provide additional results of the simulation study in Appendix A6.

## Application

We applied all candidate methods for estimating correlation to our partially matched dataset for participants in the ‘Intervention’ and ‘Control’ groups, separately. The mean difference between the pre- and post- intervention outcomes was 0.556 in the intervention group, and -0.536 in the control group.

[Table 1 here]

Using the Bayesian estimator of correlation, we would infer a 95% CI of (0.118, 0.997) for the difference in means in the intervention group. We report a p-value of 0.014, thus we would conclude there was a change in outcomes following intervention at a significance level of 0.05. With Student’s two-sample *t*-test, however, we would infer a 95% CI of (-0.023, 2.238; p = 0.062). With the 20th quantile approach, we have 95% CI: (…; p = 0.050). With the maximally conservative approach, we would infer a 95% CI of (-0.243, 1.358; p = 0.17). The discrepancy in hypothesis testing conclusions between the former and latter approaches highlights the difference in power between our proposed method and currently available methods. Translating to the standard error estimate of the mean difference, compared to the Bayesian correlation estimate (SE=0.224), r\_q20 yielded 25% higher SE, Student’s two-sample t-test had 32% higher SE, and the maximally conservative approach had 83% higher SE.

[Table 2 here]

Using the Bayesian estimator of correlation, we would infer a 95% CI of (-0.974, -0.098) and conclude a significant (p=0.018) change between timepoints in the control group. We would reach the same hypothesis testing conclusion with all the candidate methods except the maximally conservative approach, which yielded a 95% CI of (-1.164, 0.091; p=0.097). Student’s *t*-test yielded a 95% CI of (-0.988, -0.085) and the same conclusion as the test based on the Bayesian estimator of correlation, albeit with less precision (i.e., a wider confidence interval).

# Discussion

In this work, we described the challenge of testing the equality of means in unmatched data and identified currently available methods for doing so, namely Student’s *t*-test and a maximally conservative test. We sought to improve upon these methods, provided the existence of a small number of matched samples.

In small (i.e., n=20) datasets with only four matched samples, our simulation study demonstrated that the modified *t*-test based on the 20th quantile correlation estimate offered consistent Type I error control, while affording more power than the maximally conservative approach, across all values of correlation. Provided the true correlation was greater than 0.5, this test also afforded more power than Student’s *t*-test. These results suggest that the quantile estimator test is an improvement upon existing methods when dealing with small samples of partially matched data.

In large (i.e., n=200) datasets with 10% or more matched samples, our simulation study demonstrated that the modified *t*-test based on either the Bayesian estimator or the EM algorithm estimator of correlation offered consistently closest-to-nominal Type I error control across all values of correlation, although the 20th quantile estimator yielded consistently *under* nominal Type I error control, while being much more powerful than the maximally conservative estimator). The Bayesian estimator tended to be slightly more conservative than the EM algorithm estimator in terms of Type I error. Both approaches were well powered to detect medium and large differences in means in simulation. These results suggest two available tests that offer improvements over existing methods when dealing with large samples of partially matched data.

Our simulation study also demonstrated that the Pearson correlation of matched samples yields a modified t-test with suboptimal performance, since the Type I error rate was increasingly inflated as the true correlation increased. This inflation in Type I error was persistent in datasets where the number of matched samples was less than 25. These results suggest instead using either the quantile estimator or the Bayesian estimator when dealing with partially matched data. This is particularly relevant, since the first choice of an estimator for partially matched data may be the Pearson correlation, absent other intuition.

Although methods exist for finding the maximum likelihood estimates of the bivariate normal distribution with missing data, the focus of those methods has primarily been on cases where data are missing due to dropout; see [7]. Our case does not involve the same mechanism of missingness, since we do not assume dropout, but we assume missing identifiers. Therefore the methods in this study address a different, but related scenario to what has been addressed in [2] [3] [4], and [7].

We acknowledge that a fully Bayesian approach might be desirable, for example one positing prior distributions for all five parameters of the bivariate normal distribution. Such an approach has been successful in developing a Bayesian alternative to the *t­*-test; see [8]. However, partially matched data make the computation of the likelihood intractable, since at least some of the paired samples cannot be matched and the cross product is incompletely observed. Instead, we have chosen to incorporate Bayesian methodology using the estimator , which posits a prior for the correlation only.

Based on the results of our simulation study, we make the following recommendations for testing the equality of means in partially matched data. First, when controlling the Type I error rate is of primary concern and between four and ten matched samples are available, the modified *t*-test based on the 20th quantile estimator offers a reasonable level of conservatism and is the most appropriate among our candidate methods. Second, when power is of primary concern and at least two matched samples are available, the modified *t*-test based on the EM algorithm estimator offers a well-powered test with generally consistent Type I error control (albeit slightly inflated). The modified *t*-test based on the Bayesian estimator offers a balance between the two preceding methods, that is, Type I error and power, and is generally consistent when the number of matched samples is at least four. Both the EM algorithm estimator and the Bayesian estimator will converge towards unbiasedness as the number of matched samples grows large. Lastly, when applying these methods to ordinal outcomes data, we expect a greater degree of Type I error rate inflation, although the three previously mentioned tests stabilize to reasonable Type I error rate once the number of matched samples is larger than ten. Practically, and especially in ordinal data settings, we suggest using a method for getting the 20th quantile estimate using bootstrapping rather than the analytical formulation which assumes normality.

In this work, we have considered only the scenario of equal, but unknown variances in simulated and actual datasets. We leave the possibility of extending this framework to unequal variances for future work. It also may be pertinent to extend some of these methods into multiple repeated measures.

# References

|  |  |
| --- | --- |
| [1] | D. W. Zimmerman, “Correcting Two-Sample z and t Tests for Correlation: An Alternative to One-Sample Tests on Difference Scores,” *Psicológica,* vol. 33, pp. 391-418, 2012. |
| [2] | B. Guo and Y. Yuan, “A comparative review of methods for comparing means using partially paired data,” *Statistical Methods in Medical Research,* vol. 26, no. 3, p. 1323–1340, 2017. |
| [3] | P.-e. Lin and L. E. Stivers, “On difference of means with incomplete data,” *Biometrika,* vol. 61, no. 2, pp. 325-334, 1974. |
| [4] | H. M. Samawi and R. Vogel, “Notes on two sample tests for partially correlated (paired) data,” *Journal of Applied Statistics,* vol. 41, no. 1, pp. 109-117, 2014. |
| [5] | B. K. Fosdick and A. E. Raftery, “Estimating the Correlation in Bivariate Normal Data With Known Variances and Small Sample Sizes,” *The American Statistician,* vol. 66, no. 1, pp. 34-41, 2012. |
| [6] | A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society Series B,* vol. 39, no. 1, pp. 1-38, 1977. |
| [7] | R. C. Dahiya and R. M. Korwar, “Maximum Likelihood Estimates for a Bivariate Normal Distribution with missing data,” *The Annals of Statistics,* vol. 8, no. 3, pp. 687-692, 1980. |
| [8] | J. K. Kruschke, “Bayesian Estimation Supersedes the t Test,” *Journal of Experimental Psychology: General,* vol. 142, no. 2, pp. 573-603, 2013. |

Figure 1: Illustration of datasets that are paired versus paired but unmatched.

|  |  |
| --- | --- |
| A | Table  Description automatically generated |
| B | **Chart, funnel chart  Description automatically generated** |

Caption: The dataset in panel (A) is paired and matched. The dataset in panel (B) is paired but unmatched since we cannot identify which observation in X corresponds to its pair in Y.

Figure 2: Simulation results evaluating the bias in estimators of interest.

|  |  |
| --- | --- |
| A | Chart, line chart  Description automatically generated |
| B | **Chart, line chart  Description automatically generated** |

Caption: Datasets in panel (A) were simulated with bivariate normal distributions. Datasets in panel (B) were simulated with bivariate ordinal distributions. Note that the maximally conservative estimator was withheld from these plots to aid visualization.

Figure 3: Simulation results evaluating the mean squared error (MSE) in estimators of interest.

|  |  |
| --- | --- |
| A | Chart, line chart  Description automatically generated |
| B | **Chart, line chart  Description automatically generated** |

Caption: Datasets in panel (A) were simulated with bivariate normal distributions. Datasets in panel (B) were simulated with bivariate ordinal distributions. Note that the maximally conservative estimator was withheld from these plots to aid visualization.

Figure 4: Simulation results evaluating the standard errors (SE) using estimators of interest.

|  |  |
| --- | --- |
| A | Chart, line chart  Description automatically generated |
| B | **Chart, line chart  Description automatically generated** |

Caption: Datasets in panel (A) were simulated with bivariate normal distributions. Datasets in panel (B) were simulated with bivariate ordinal distributions.

Figure 5: Simulation results evaluating the Type I error for candidate methods of interest.

|  |  |
| --- | --- |
| A | Chart, line chart  Description automatically generated |
| B | **Chart, line chart  Description automatically generated** |

Caption: Datasets in panel (A) were simulated with bivariate normal distributions. Datasets in panel (B) were simulated with bivariate ordinal distributions. Note that the true difference in means was zero. The line labelled ‘Independent’ corresponds to Student’s t-test.

Figure 6: Simulation results evaluating power for candidate methods of interest.

|  |  |
| --- | --- |
| A | Graphical user interface  Description automatically generated |
| B | **Graphical user interface, diagram  Description automatically generated** |

Caption: Datasets in panel (A) were simulated with bivariate normal distributions. Datasets in panel (B) were simulated with bivariate ordinal distributions. On the left of both panels, the true difference in means was 0.25 standardized units; on the right, the difference was 0.5 standardized unites. Note that the line labelled ‘Independent’ corresponds to Student’s t-test.

Table 1: Results of tests for equality of means in the Intervention Group.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Correlation** | **T Statistic** | **Std. Error** | **95% CI** | | **p-value** |
| Max. Conservative | -0.906 | 1.364 | 0.409 | -0.243 | 1.358 | 0.175 |
| Pearson | 0.418 | 2.470 | 0.226 | 0.115 | 1.000 | **0.015** |
| 20th Quantile | 0.102 | 1.987 | 0.281 | 0.008 | 1.107 | 0.050 |
| Bayesian | 0.427 | 2.487 | 0.224 | 0.118 | 0.997 | **0.014** |
| EM Algorithm |  |  |  |  |  |  |
| Independence (Student’s *t*-test) | 0 | 1.883 | 0.296 | -0.023 | 1.138 | 0.062 |

Caption: The mean difference between the pre- and post- intervention outcomes was 0.556 (pre- minus post-), suggesting average decline in outcomes following the intervention. Using the Bayesian estimator of correlation, we would infer a 95% CI of (0.118, 0.997) and conclude a significant change following intervention. With Student’s t-test, i.e., assuming no correlation, we would infer a 95% CI of (-0.023, 2.238) and conclude no significant change.

Table 2: Results of tests for equality of means in the Control Group.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Correlation** | **T Statistic** | **Std. Error** | **95% CI** | | **p-value** |
| Max. Conservative | -0.933 | -1.674 | 0.320 | -1.164 | 0.091 | 0.097 |
| Pearson | 0.059 | -2.400 | 0.223 | -0.974 | -0.098 | **0.018** |
| 20th Quantile | -0.253 | -2.079 | 0.258 | -1.042 | -0.031 | **0.040** |
| Bayesian | 0.058 | -2.399 | 0.224 | -0.974 | -0.098 | **0.018** |
| EM Algorithm |  |  |  |  |  |  |
| Independence (Student’s *t*-test) | 0 | -2.328 | 0.230 | -0.988 | -0.085 | **0.022** |

Caption: The mean difference between the pre- and post- intervention outcomes was -0.536 (pre- minus post-), suggesting average improvement in outcomes following the intervention. Using the Bayesian estimator of correlation, we would infer a 95% CI of (-0.974, -0.098) and conclude a significant change following intervention. With Student’s t-test, i.e., assuming no correlation, we would infer a 95% CI of (-0.988, -0.085) and reach the same conclusion.

Appendix

A1. Description of EM algorithm implementation

A2. Figure showing effective correlation versus true correlation, for ordinal datasets

A3. Figure showing histograms of simulated and actual datasets with ordinal outcomes

A4. Explanation on how proposed estimators were modified for the application datasets

A5. Additional figures presenting results of the simulation study