**Help! I’ve lost my ID(s), can I still test for a difference in means?**

**Abstract**

Indicate the abstract paragraph with a heading or by reducing the font size. 200 word limit.

A typical research article should be no more than 20 manuscript pages.

Introduction:

Methodology or Approach: Present your novel method to solve the challenge in this or more sections and demonstrate that the proposed method has answered questions that were not properly or optimally addressed by relevant existing methods. Defer lengthy mathematical proofs to supplemental material whenever applicable.

Application of your Methodology: Present application of your novel method to solve the data challenge presented in your Introduction in this or more sections. Also note: Simulation studies can be presented as a sub-section in this section or in a new standalone section.

Conclusion and Discussion: In this section, present the conclusion and discussion.

**Keywords**

Keywords help readers find your article, so are vital for discoverability. If the journal instructions for authors don’t give a set number of keywords to provide, aim for five or six.

**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 a constant; see [1]. This work adapts the latter version of the *t*-test to addresses 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 procedure, 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. Study designs with paired but unmatched samples have received relatively less attention, and 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 ­*t*-test is equivalent to assuming samples are uncorrelated, which is often an unrealistic assumption of repeated measures in biostatistics. One available alternative involves calculating the minimum possible correlation coefficient given the observed data; this 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 a survey field in which one could enter the last four digits of one’s phone number. Within the control group, 10 (14%) survey responses were matched on the optional identifier. In the intervention group, 9 (11%) of survey responses were matched. Thus, the application is 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 control and intervention groups separately, while controlling Type I error at a nominal level.

**Materials and Methods**

***Notation / Definitions***

Let **X** and **Y** denote random variables for which we wish to compare means. Let denote the true, unobserved correlation between **X** and **Y**. 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]:

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:*

One can calculate the minimum possible correlation coefficient for any paired dataset, even if entirely unmatched, which yields a maximally conservative test for the equality of means (i.e., it would result in the fewest number of rejected null hypotheses).

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

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 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 *m* matched samples and ignores the *n - m* matched samples. The following formula gives the Pearson correlation coefficient for matched samples:

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, yet the degree of conservativism in might be too extreme, we hypothesized that a quantile estimate of might yield additional power while maintaining appropriate Type I error 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:

Where is the critical value chosen based on a desired confidence level. Since our desired confidence level was 80%, 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 .

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.

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 .

*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 expectation 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:

Note that standardizing the data is achieved using both matched and unmatched samples. However, the above estimator requires at least one matched sample to be calculable.

*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., ).

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 the Appendix. 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 the Appendix.

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). We compared estimators to one another based on resulting metrics of bias, mean squared error, Type I error and power. For reference, we also included a naïve approach of assuming independence for all simulated datasets, and an ‘oracle’ approach in which the true correlation was known.

***Application***

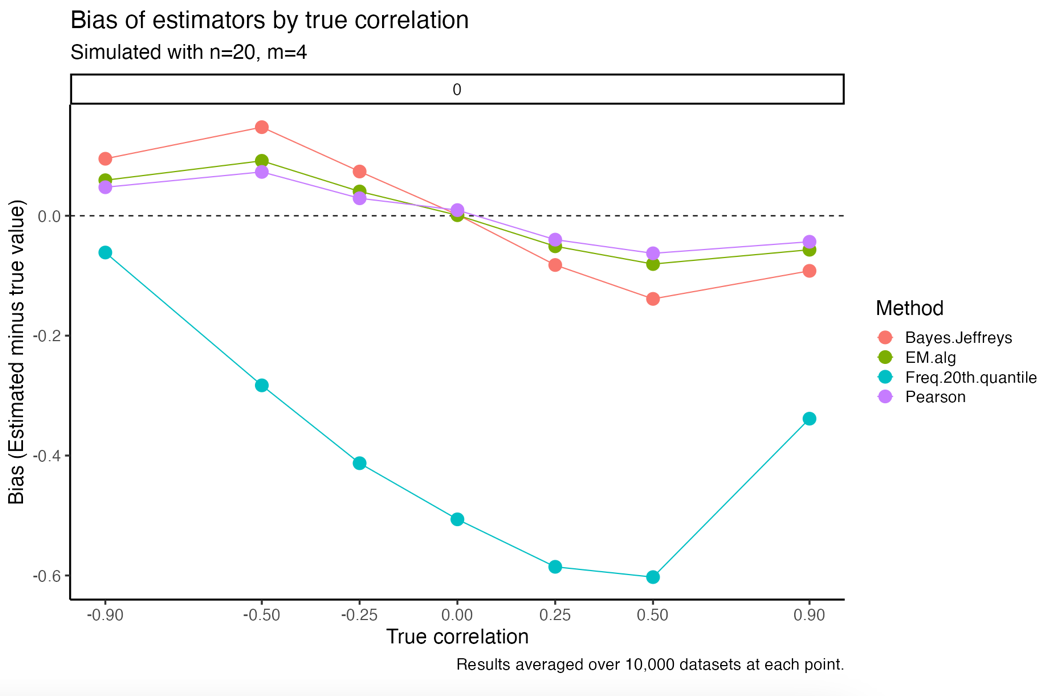
Using the above estimators, we applied modified *t*-tests to evaluate the difference in means between timepoints for the control and intervention groups separately. We computed standard errors, 95% confidence intervals for the difference in means, and p-values for each of the candidate methods. Since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

**Results**

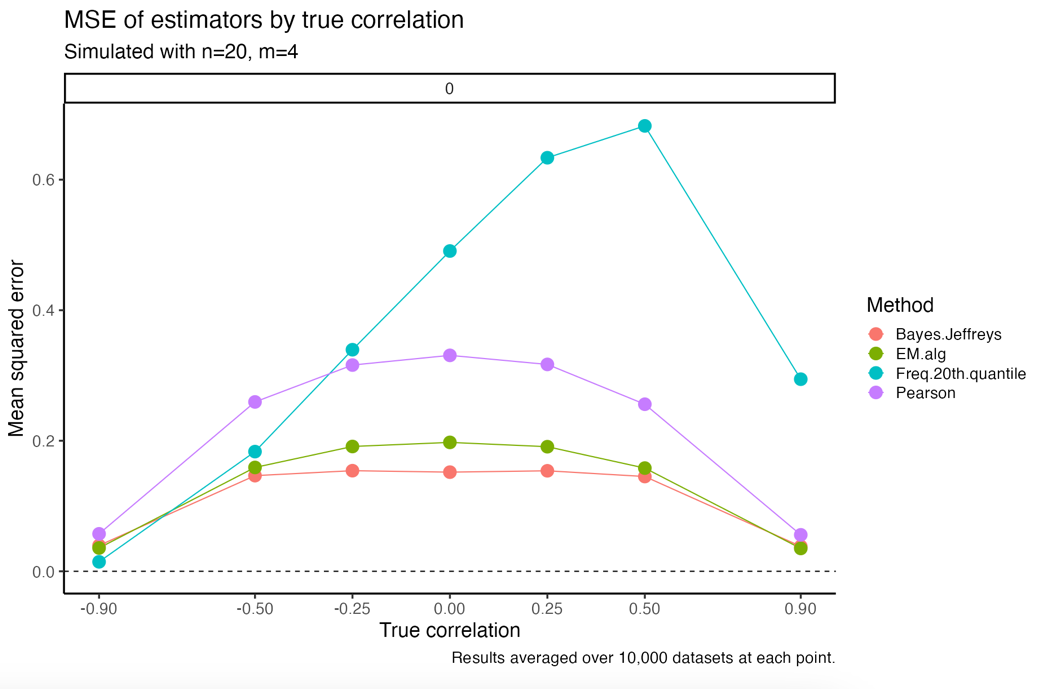
***Simulation study***

Not all estimators yielded a valid correlation, particularly when the number of matched samples was small (i.e., less than four) and the generating distribution was ordinal.

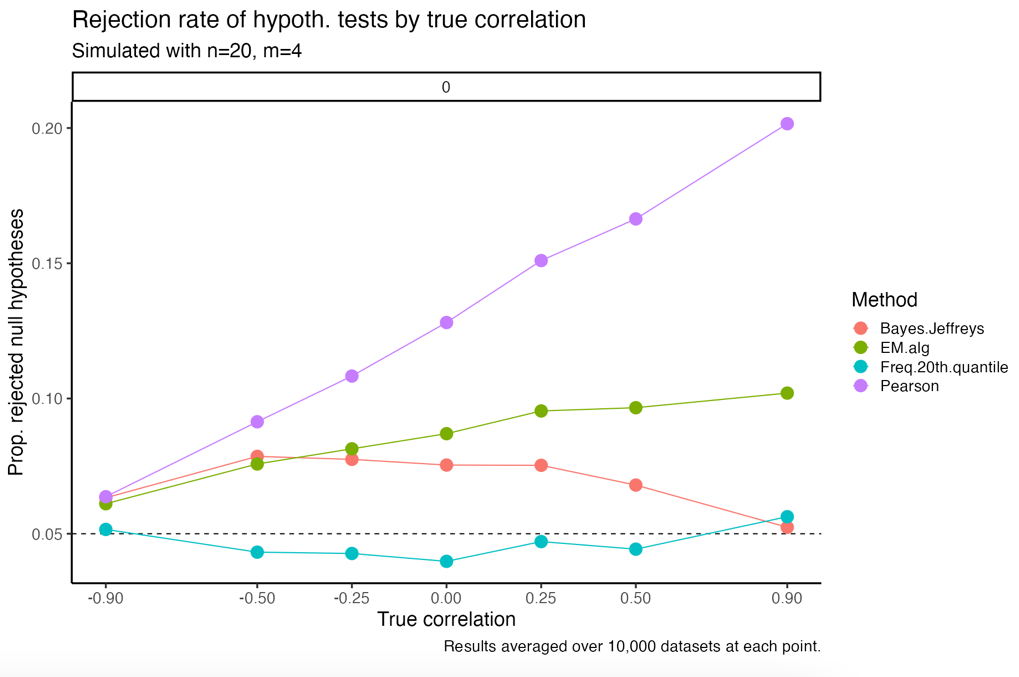
For normally distributed data with a sample size of 20 and four matched samples, the Pearson correlation of the matched samples (rmatched) was the least biased estimator across all values of correlation. The EM algorithm estimator (rEM) was slightly more biased than rmatched, followed by the Bayesian estimator (rBayes). All three of these estimators were biased toward overestimating the correlation when negative and underestimating the correlation when positive. Additionally, the 20th quantile estimator (rq20) was consistently biased toward underestimating the correlation and was least biased when the true correlation was 0.9. When the number of matched samples was reduced from four to two, there was considerably less separation among rmatched, rEM, and rBayes in terms of bias (rq20 was not calculable due to insufficient matched samples).



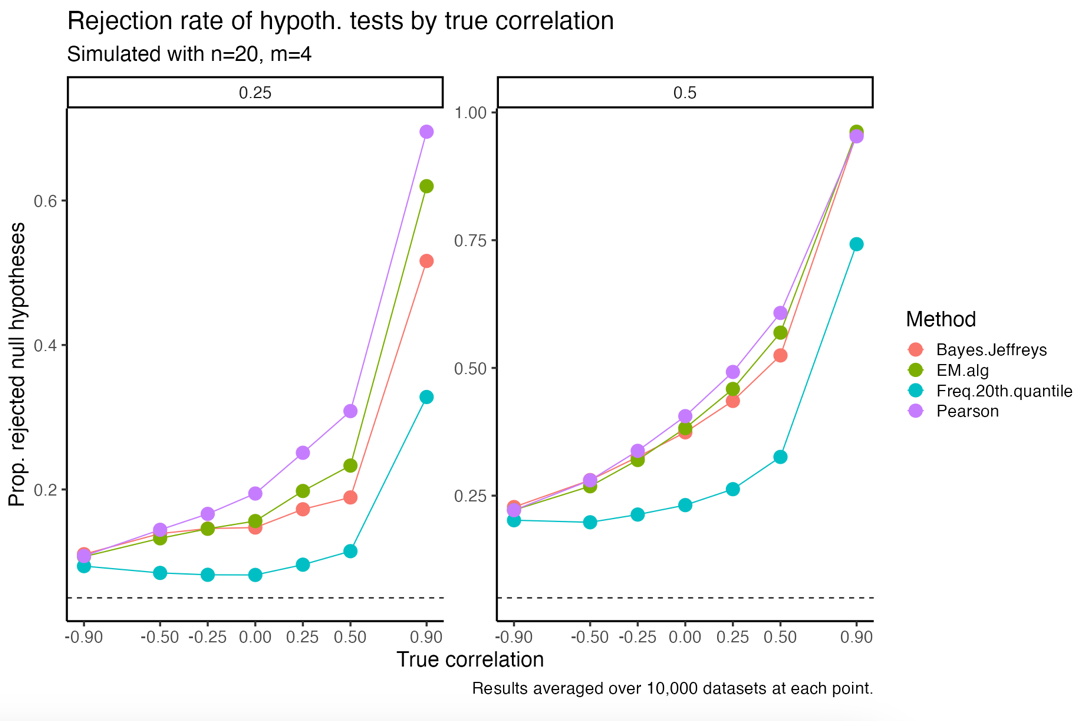
There was generally an inverse relationship between bias and mean squared error (MSE). For normally distributed data with a sample size of 20 and four matched samples, the Bayesian estimator (rBayes) was overall lowest in MSE except at extreme values of correlation. The EM algorithm estimator (rEM) trailed behind rBayes with slightly greater MSE at all but the most extreme values of correlation. The Pearson estimator (rmatched) trailed behind both former estimators. These differences persisted even as the number of matched samples was reduced from four to two. The 20th quantile estimator (rq20) was asymmetrically more variable when the true correlation was positive, and less variable when the true correlation was negative.

****

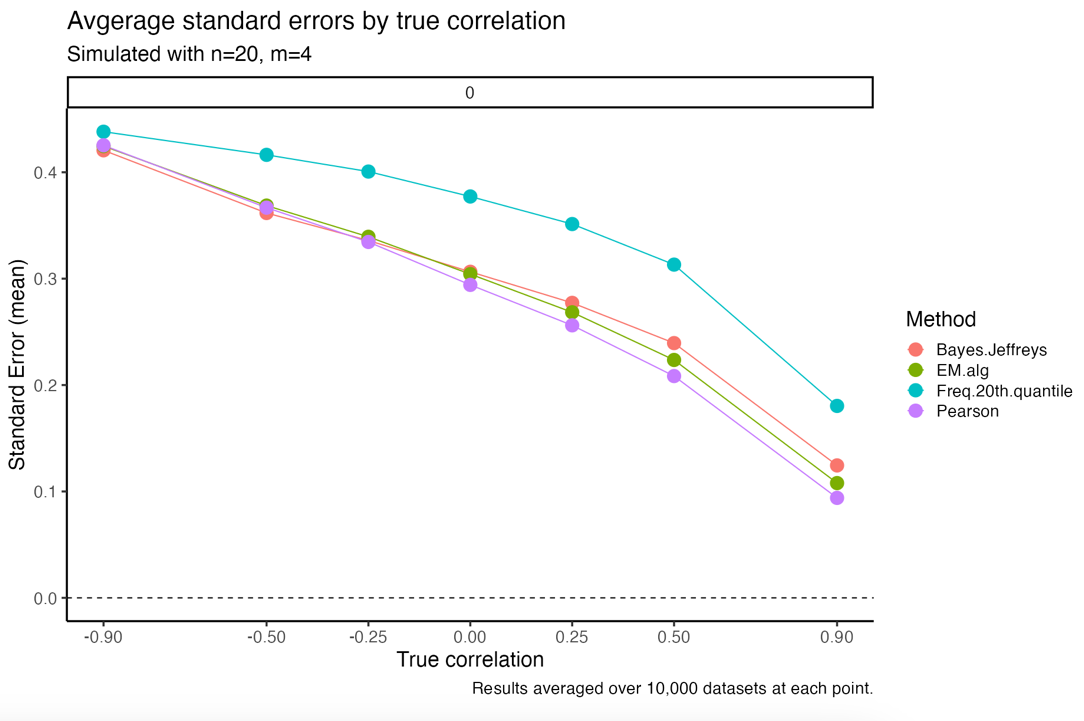
When comparing the Type I error rate among tests computed with each of the estimators of correlation, the 20th quantile estimator (rq20) was more conservative. With normally distributed data, a sample size of 20, and four matched samples, rq20 controlled Type I error near the nominal level of 0.05 across all values of correlation. The Bayesian estimator (rBayes) yielded a test that was generally less conservative than the nominal level, but far more conservative than the Pearson estimator (rmatched) for all values of correlation greater than -0.25. The same was true of the EM algorithm estimator (rEM), although rEM was slightly less conservative than rBayes. There was a direct relationship between the magnitude of the true correlation and the Type I error of the Pearson correlation estimator (rmatched). When the true correlation was positive, this test rejected the null hypothesis at least 10% of the time, which was twice the nominal rate.



There was generally a trade-off between Type I error control and power. With the same simulation settings mentioned in the previous paragraph (bivariate normal, n=20, m=4), the Pearson correlation (rmatched) afforded the most power, although the improvement in power was smaller when the true mean difference was 0.5 standardized units. The 20th quantile estimator (rq20) demonstrated the least power among the four estimators (although the maximally conservative estimator was underpowered compared to all estimators).



The relative conservatism exhibited in the test with the 20th quantile estimator can be attributed to reduced precision. The following plot presents standard errors (averaged over all simulations) for the various estimators. We observed that rq20 yielded the least precision among the four estimators.



***Application dataset***

**Discussion**

In this work, we described the challenge of testing the equality of means in unmatched data and identified available methods for doing so, namely the two-sample *t*-test and a maximally conservative approach. We sought to improve upon these methods, provided that a small number of matched samples existed.

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 the two-sample *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 consistent Type-I error control across all values of correlation. 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 test with suboptimal performance, since the Type-I error 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 generally avoiding the Pearson correlation estimator, and 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 [7]. Our case does not involve the same mechanism of missingness, since we do not assume dropout, but we assume missing identifiers.

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 [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 incorporated Bayesian philosophy 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 Type-I error is of primary concern and at least four 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 controls, that is, Type-I error and power, and is generally consistent when the number of matched samples is at least four. Lastly, when applying these methods to ordinal outcomes data, we expect a greater degree of Type-I error inflation, although the three previously mentioned tests stabilize once the number of matched samples is about ten.

**Acknowledgements**

**References**

**Tables**

table(s) with caption(s) (on individual pages)

Show clearly in your article text where the tables and figures should appear, for example, by writing [Table 1 near here].

**Figures**

figures; figure captions (as a list)

**Appendix**

Description of EM algorithm implementation

Figure of effective correlation versus true correlation

Details on how estimators were modified for application datasets