JOURNAL OF APPLIED STATISTICS
2025, VOL. 52, NO. 6, 1161–1176
https://doi.org/10.1080/02664763.2024.2418473

Taylor & Francis
Taylor & Francis Group

Evaluating the median *p*-value method for assessing the statistical significance of tests when using multiple imputation

Peter C. Austin[a,b,c], Iris Eekhout[d] and Stef van Buuren[d,e]

[a]ICES, Toronto, Canada; [b]Institute of Health Policy, Management and Evaluation, University of Toronto, Canada; [c]Sunnybrook Research Institute, Toronto, Canada; [d]Department of Child Health, Netherlands Organization for Applied Scientific Research TNO, Leiden, The Netherlands; [e]Department of Methodology and Statistics, University of Utrecht, Utrecht, The Netherlands

ABSTRACT

Rubin's Rules are commonly used to pool the results of statistical analyses across imputed samples when using multiple imputation. Rubin's Rules cannot be used when the result of an analysis in an imputed dataset is not a statistic and its associated standard error, but a test statistic (e.g. Student's t-test). While complex methods have been proposed for pooling test statistics across imputed samples, these methods have not been implemented in many popular statistical software packages. The median *p*-value method has been proposed for pooling test statistics. The statistical significance level of the pooled test statistic is the median of the associated *p*-values across the imputed samples. We evaluated the performance of this method with nine statistical tests: Student's t-test, Wilcoxon Rank Sum test, Analysis of Variance, Kruskal-Wallis test, the test of significance for Pearson's and Spearman's correlation coefficient, the Chi-squared test, the test of significance for a regression coefficient from a linear regression and from a logistic regression. For each test, the empirical type I error rate was higher than the advertised rate. The magnitude of inflation increased as the prevalence of missing data increased. The median *p*-value method should not be used to assess statistical significance across imputed datasets.

ARTICLE HISTORY
Received 19 November 2023
Accepted 8 October 2024

KEYWORDS
Missing data; multiple imputation; Rubin's Rules; hypothesis testing

1. Introduction

Missing data are common in applied research. Missing data occur when the value of a variable is recorded for some subjects, but not for all subjects in the dataset. Ignoring missing data and conducting statistical analyses in those subjects with complete data can result in biased estimates of statistics and decreased precision. Multiple imputation (MI), which was initially proposed by Rubin, is a statistical method to address missing data that entails using a missing data model to impute or fill in multiple plausible values for missing data [5]. This results in the creation of M complete samples (M $>$ 1), each with the same size

CONTACT Peter C. Austin peter.austin@ices.on.ca Institute for Clinical Evaluative Sciences, G106, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada

as the original sample. However, in each of these M samples, all missing data have been replaced with plausible values. The analyst then conducts a statistical analysis in each of these M complete samples. The results of the statistical analyses are then pooled across the M complete samples. MI is being used with increasing frequency (see Figure 2.1 in van Buuren [8]).

The most common method for pooling results of statistical analyses across complete samples is Rubin's Rules [5]. When estimating a parameter for which there is an associated standard error (e.g. a regression coefficient and its associated standard error), Rubin's Rules provides a method for pooling the statistic and its standard error across the imputed samples. One can then assess the statistical significance of the pooled parameters using conventional statistical methods. However, when the result of an analysis in a complete sample is not a parameter estimate and associated standard error, but a test statistic (e.g. Student's t-test for testing the equality of the mean of a continuous variable in two independent populations), then Rubin's Rules cannot be applied to pool the test statistics across the complete samples. Different methods have been proposed for pooling test statistics across imputed samples [8]. However, unlike Rubin's Rules, these methods have not been implemented in commonly used statistical software packages (e.g. SAS PROC MIANALYZE allows for pooling of parameter estimates, but not of test statistics; the R function mice::pool() expects a standard error, which is not available for objects of class htest), and can be difficult for analysts to implement on their own.

Eekhout and colleagues proposed the median *p*-value method to assess the statistical significance of test statistics for categorical variables in logistic regression, that are pooled across imputed samples [2]. The statistical test is applied in each of the M complete samples and the *p*-value associated with the test is noted in each of the M complete samples. The median of these M *p*-values is the median *p*-value and is used to formally test the given hypothesis across the imputed datasets. Thus, one would reject the null hypothesis if the median *p*-value were less than or equal to 0.05. Given the simplicity of this method and its ease of implementation, it would be useful to explore the validity of the method across a range of statistical tests.

The objective of the current paper was to evaluate the type I error rate of the median *p*-value method for determining the significance levels of test statistics pooled across imputed samples. The paper is structured as follows: in Section 2, we provide a brief description of the median p-value method. In Section 3, we describe a series of Monte Carlo simulations that were used to address this question. In Section 4, we report the results of these simulations. Finally, in Section 5, we summarize our findings and place them in the context of the existing literature.

2. The median *P*-value method

The median *p*-value test is a method that has been proposed for use when the result of the statistical analysis in each of the M complete datasets is a *p*-value (i.e. from a statistical test) and not a set of regression coefficients and their associated standard errors [2]. After analyses in the M complete datasets, there will be M *p*-values, one from each of the M complete datasets: $p_1, p_2, \ldots, p_M$. The median *p*-value is the median of these M *p*-values. If the median *p*-value is less than the specified significance level (e.g. 0.05), then the analyst would reject the null hypothesis.

3. Monte Carlo simulation methods

We used Monte Carlo simulations to evaluate the type I error rate of the median *p*-value method when applied to the following 9 statistical tests: Student's t-test, Wilcoxon Rank Sum test, Analysis of Variance (ANOVA), Kruskal-Wallis test, the test of significance for Pearson's correlation coefficient, the test of significance for Spearman's correlation coefficient, the Chi-squared test, the test of significance for a regression coefficient from a linear regression model estimated using ordinary least squares (OLS), and the test of significance for a regression coefficient estimated from a logistic regression model.

We describe the simulations for each of these tests in the following sub-sections. In each set of simulations, we simulated data under the null hypothesis, induced missing data under a missing at random (MAR) missing data mechanism, created complete datasets using the MICE algorithm, and then applied the median *p*-value method. Since none of our statistical tests involved longitudinal data with a temporal ordering of the variables, we only considered non-monotone (or general) patterns of missing data and did not consider monotone patterns of missing data.

For a given statistical test, we considered 10 scenarios defined by the proportion of missing data. This factor ranged from 0 (no missing data) to 0.9 in increments of 0.1. In each scenario, we simulated 1,000 datasets. For example, in the first scenario, the prevalence of missing data was 0, while in the second scenario, the presence of missing of missing data was 0.1, while in the tenth scenario the prevalence of missing data was 0.9. For a given scenario (i.e. a given prevalence of missing data), the proportion of simulated datasets in which we rejected the null hypothesis is an estimate of the empirical type I error rate of the median *p*-value method.

When estimating the empirical type I error rate in settings with no missing data, the following approach was used for all nine statistical tests: in each simulated dataset, prior to inducing missing data, we applied the statistical test (e.g. Student's t-test) and noted the statistical significance of the test. This was categorized as statistically significant ($P \leq 0.05$) or as not statistically significant ($P > 0.05$) (i.e. we rejected or accepted the null hypothesis of no difference in means). The empirical type I error rate in the absence of missing data was the proportion of simulated datasets in which we rejected the null hypothesis across the 1,000 simulation replicates.

3.1. Student's t-test and Wilcoxon rank sum test

We consider the two sample Student's t-test that does not assume equal variances in the two groups (and thus the variance is estimated separately in each group and the Welch approximation to the degrees of freedom is used) and the Wilcoxon rank sum test (also known as the Mann Whitney U test). The former is a parametric test that tests the hypothesis that the mean of a given variable is equal in two independent populations. The latter is a non-parametric test that tests whether the distribution of the variable is the same in two independent populations.

For a given value of the proportion of missing data (as noted above, ranging from 0 to 0.9 in increments of 0.1), we simulated samples of size 1,000. For each of the 1,000 subjects we simulated two variables: (i) a binary group variable G (taking the values A vs. B) using a Bernoulli distribution with parameter 0.5 (i.e. 50% of subjects were labeled as 'A', while

50% were labeled as 'B'); (ii) a continuous variable, X, such that each subject's value of X was drawn from a standard normal distribution. Since X was generated from the same distribution for those in the two levels of the group variable, we were generating data under the null hypothesis: the population mean of X was the same in the two groups (or, for the Wilcoxon rank sum test, the distribution of X was the same in the two groups).

We then induced missing data in the random sample using the mice::ampute() function [6]. We induced missing data such that there were two missing data patterns: (i) the binary group variable G was missing and the continuous variable X was observed; (ii) the binary group variable G was observed and the continuous variable X was missing. We used a missing at random (MARRIGHT) missing data mechanism that created more missing data for the higher values [8] (pages 70–73). Thus, the likelihood of missing data in the group variable was positively associated with X and the likelihood of missing data in X was positively associated with the group variable. Except for the proportion of incomplete rows (called 'prop'), all arguments in the mice::ampute() function were left at default, resulting in equal prevalences for the two missing data patterns. In short, the process to generate missing values works as follows. The user specifies the allowed missing data patterns (patterns), the relative frequency of each pattern (freq), and the proportion of incomplete cases (prop). For MAR and MNAR mechanisms, we can optionally specify the weights of predictors to create a linear combination per pattern (weights) and the location on the linear combination where missing values should be assigned (type). We set the weights to all be equal to 1. Thus, the magnitude of the association of X with missingness in G was the same as the magnitude of the association of G with missingness in X. For each row in the complete data, the algorithm randomly draws a missing data pattern using a combination of prop and freq, constructs the sum scores, and creates missing values and the specified locations. For the exact details, we refer to Schouten and colleagues [6,7]. The decision to set the prevalences of the two missing data patterns to be equal was made to simplify the simulations, rather than allowing there to be multiple scenarios defined the relative frequency of the two missing data patterns. We specified that the prevalence of missing data be the same for the two variables because it seemed to be a balanced approach, rather than allowing the missingness in one variable to dominate the analyses. Multiple imputation using the multivariate imputation using chained equations (MICE) algorithm was used to impute missing values [8–10]. We created M complete datasets, where M was set equal to the percentage of subjects for whom there was missing data [11]. Thus, for example, when data were missing for 50% of the subjects, we created 50 complete datasets. In each of the M complete datasets we used a t-test to compare the mean of X between the two groups. We then computed the median p-value across the M complete datasets.

We repeated the above process 1,000 times and determined the proportion of simulated datasets in which we rejected the null hypothesis. This is the empirical type I error rate of the median p-value method. Due to our use of 1,000 simulation replicates, empirical type I error rates less than 0.0365 or greater than 0.0635 are statistically different from the advertised rate of 0.05 using a standard normal-theory test. We also obtained a non-parametric estimate of the density function of the median p-values across the 1,000 simulation replicates. If the median p-value test was behaving as advertised, the distribution would be uniform U(0,1). The above process was repeated for each of the different prevalences of missing data. We also obtained a non-parametric estimate of the density function of the p-value for the t-test across the M x 1,000 imputed datasets.

The two-sample t-test allows for the comparison of the mean of a continuous variable between two groups. An alternative way to test this hypothesis is to use a univariate linear regression model in which the continuous variable is regressed on a binary indicator variable denoting group membership. A rationale for including the use of linear regression for comparing group means is that one can compare this method, which involves the application of Rubin's Rules, with a method that pools *p*-values. We compared the performance of linear regression with that of the median *p*-value test used with the two-sample t-test. To do so, we repeated the above simulations using linear regression (estimated using OLS) to regress the continuous variable X on a binary indicator variable denoting group membership (thus, the analysis model was a linear regression model in which X was regressed on the binary indicator variable denoting group membership). The analysis model was fit in each of the complete datasets and the estimated regression coefficients were pooled using Rubin's Rules. The statistical significance of the estimated regression coefficient for group membership was assessed. We then repeated the above simulations using the Wilcoxon rank sum test.

3.2. ANOVA and Kruskal-Wallis test

ANOVA is a parametric statistical test that tests the equality of means across a set of independent populations. The Kruskal-Wallis test is a non-parametric test of whether the distribution of a continuous variable is the same in different independent populations.

This set of simulations was similar to those described in Section 3.1, with minor modifications. First, as in Section 3.1, we generated two variables for each subject: (i) a 3-level group variable G, such that the prevalence of each of the three levels was 1/3 (i.e. the three groups were approximately of equal size in the simulated samples) (this is the primary difference from the simulations described in Section 3.1 which involved simulating a binary group variable); (ii) as in Section 3.1, a continuous variable X from a standard normal distribution. Therefore, the mean of X was the same across the three levels of the group variable. We then used ANOVA to test the null hypothesis that the mean of X was the same across the three levels of the group variable. Apart from these differences, these simulations were similar to those described in Section 3.1. As above, the weights that were used to calculate the weighted sum scores when inducing missing data were all set equal to 1.

We then repeated the above simulations using the Kruskal-Wallis rank sum test to test whether the distribution of X was the same across the three groups.

3.3. Tests of Pearson's and Spearman's correlation coefficient

Pearson's correlation coefficient is a metric for quantifying the linear correlation between two continuous variables. Spearman's correlation coefficient is equal to Pearson's correlation coefficient when applied to the ranks of the two variables.

We simulated samples of size 1,000. For each subject, we simulated two continuous variables, X and Y, from *independent* standard normal distributions. Thus, X and Y were independent of one another and had a true correlation of zero (i.e. the null hypothesis of a zero correlation was true).

We then induced missing data in the random sample using a MAR missing data mechanism. There were two missing data patterns: (i) X was observed and Y was missing; (ii) X

was missing and Y was observed (thus, either X could be missing or that Y could be missing, but both variables could not be missing for the same subject). The prevalences of the two missing data patterns were set equal to one another. As above, the weights that were used to calculate the weighted sum scores when inducing missing data were all set equal to 1. In each of the M complete datasets we estimated Pearson's correlation coefficient between X and Y and tested whether it was different from zero.

We repeated the above simulations using Spearman' correlation coefficient.

3.4. The Chi-squared test

The Chi-squared test tests for an association between two categorical variables. It tests whether the distribution of one categorical variable differs across the levels of the other categorical variable. For each of 1,000 subjects we simulated two 3-level categorical variables: X and Y. For each of X and Y, the prevalence of each of the three levels was 1/3. Furthermore, X and Y were simulated to be independent of one another (i.e. the null hypothesis was true). We induced missing data using a MAR missing data mechanism so that there were two patterns of missing data: (i) X was observed and Y missing; (ii) X was missing and Y was observed, with the prevalences of the two patterns of missing data being equal. As above, the weights that were used to calculate the weighted sum scores when inducing missing data were all set equal to 1. Apart from this modification, these simulations were similar to those described above.

3.5. OLS regression

For each of 1,000 subjects we simulated three predictor variables from independent standard normal distributions: $x_{ij} \sim N(0,1)$, for $j = 1, 2, 3$ and $i = 1, \ldots, 1000$. For each subject we generated a continuous outcome using the following model: $y_i = 0x_{1i} + x_{2i} + x_{3i} + \varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2 = (2/0.25) - 2)$, so that variation in the predictor variables explained 25% of the variation in the continuous outcome (i.e. the model R^2 was 0.25). Note that in this data-generating process, the outcome is conditionally independent of X_1 (i.e. the regression coefficient for the first covariate is equal to 0). Thus, we are simulating data under the null hypothesis that the regression parameter for X_1 was equal to zero.

We then induced missing data in the random sample. We induced missing data such that there were four missing data patterns: (i) X_1 missing with the other three variables being observed; (ii) X_2 missing with the other three variables being observed; (iii) X_3 missing with the other three variables being observed; and (iv) Y missing with the other three variables being observed. We set the prevalences of the four missing data patterns equal to one another (thus, when the prevalence of missing data was 40%, the prevalence of missing data for each of the four variables was 10%). We used a missing at random (MAR) missing data mechanism. Thus, the likelihood of missing data for a given variable was related to the values of the three other variables, but not to that variable itself. As above, the weights that were used to calculate the weighted sum scores when inducing missing data were all set equal to 1. Multiple imputation using the mice algorithm was used to impute missing value. We created M complete datasets, where M was set equal to the percentage of subjects for whom there was missing data. In each of the M complete datasets we regressed the continuous outcome on the three predictor variables using OLS regression (thus, the analysis

model consisted of a linear regression model in which the continuous outcome variable, Y, was regressed on X_1, X_2, and X_3). In each of the M complete datasets we noted the p-value associated with testing the null hypothesis that the regression coefficient for the first predictor variable was equal to zero. We then computed the median p-value across the M complete datasets. We also used Rubin's Rules to pool the estimated regression coefficients across the M imputed datasets and used the estimated regression coefficient and its associated standard error to test whether the estimated regression coefficient was statistically significantly different from 0.

We obtained a non-parametric estimate of the density function of the median p-values across the 1,000 simulation replicates. We also obtained a non-parametric estimate of the density function of the p-value for testing the statistical significance of the regression coefficient for X_1 across the M x 1,000 imputed samples. The above process was repeated for each of the prevalences of missing data.

3.6. Logistic regression

These simulations were similar to those described in Section 3.5, with minor modifications. For each subject, the probability of the occurrence of the binary outcome was defined as $p_i = \Pr(Y_i = 1) = \frac{\exp(0x_{1i}+x_{2i}+x_{3i})}{1+\exp(0x_{1i}+x_{2i}+x_{3i})}$. We then generated binary outcomes from a Bernoulli distribution with subject-specific parameter p_i. Thus, the odds of the outcome is conditionally independent of the first covariate. Missing data were induced as described above, with the missing data models being of the same as for OLS regression in the previous section. As above, the weights that were used to calculate the weighted sum scores when inducing missing data were all set equal to 1. The analysis model was a logistic regression model in which the binary outcome was regressed on the three continuous predictor variables.

All simulations were conducted using the R statistical programming language (version 3.6.3) [4]. Missing data were induced using the 'ampute' function in the mice package for R (version 3.13.0). Missing data were imputed using the mice function in the mice package. In the MICE algorithm, continuous variables were imputed using Bayesian linear regression, binary variables were imputed using logistic regression, and categorical variables were imputed using a multinomial logistic model.

4. Monte Carlo simulation results

We summarize the results of the Monte Carlo simulations for each of the different tests separately. Note that the empirical type I error rates for the nine tests are reported in the same figure (Figure 1) to facilitate a comparison of the empirical type I error rate across tests. On the figure we have superimposed a horizontal line denoting the advertised type I error rate of 0.05.

4.1. Student's t-test and Wilcoxon rank sum test

The relationship between the prevalence of missing data and the empirical type I error rate of the median p-value method are reported in Figure 1. The median p-value method, when used with either Student's t-test or the Wilcoxon rank sum test resulted an empirical

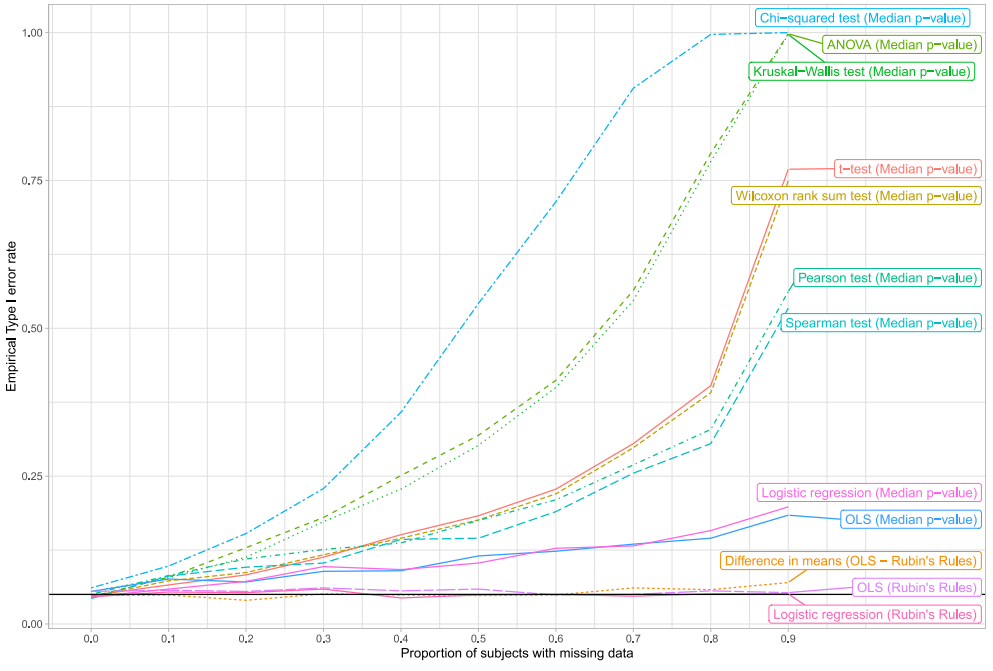

Figure 1. Empirical Type I error rates of the different tests.

type I error that exceeded the target rate of 0.05. The inflation in the empirical type I error rate increased as the prevalence of missing data increased and attained approximately 0.75 (or 75%) when the prevalence of missing data was very high. In contrast, analyses in the absence of missing data (i.e. when the prevalence of missing data was 0) had empirical type I error rates that did not differ from the target rate of 0.05. The use Rubin's Rules with OLS regression to test the equality of means between the two groups resulted in empirical type I error rates that did not differ meaningfully from the advertised rate.

The non-parametric estimates of the densities of median *p*-values across the 1,000 simulation replicates are reported in Figure 2. There is one panel for each prevalence of missing data. In each of the nine scenarios, the empirical distribution of the median *p*-values was non-uniform. There tended to be fewer large *p*-values than would be anticipated under a standard uniform distribution. The non-parametric estimates of the density functions for the *p*-values from the t-tests and Wilcoxon rank sum tests in the M x 1,000 imputed datasets are also reported in each panel of the figure. As with the median *P*-value, there tended to be fewer large *p*-values than would be anticipated under a standard uniform distribution. Thus, a separate analysis per imputed dataset does not produce a uniform distribution for the *p*-value under the null hypothesis. The t-test has n-1 as the degrees of freedom, which is too high for imputed data. As a result, we observe a shift to the left in the *p*-value distributions. The magnitude of the shift grows with the proportion of missing data.

4.2. ANOVA and Kruskal-Wallis test

The relationship between the prevalence of missing data and the empirical type I error rate of the median *p*-value method are reported in Figure 1. The empirical type I error rate

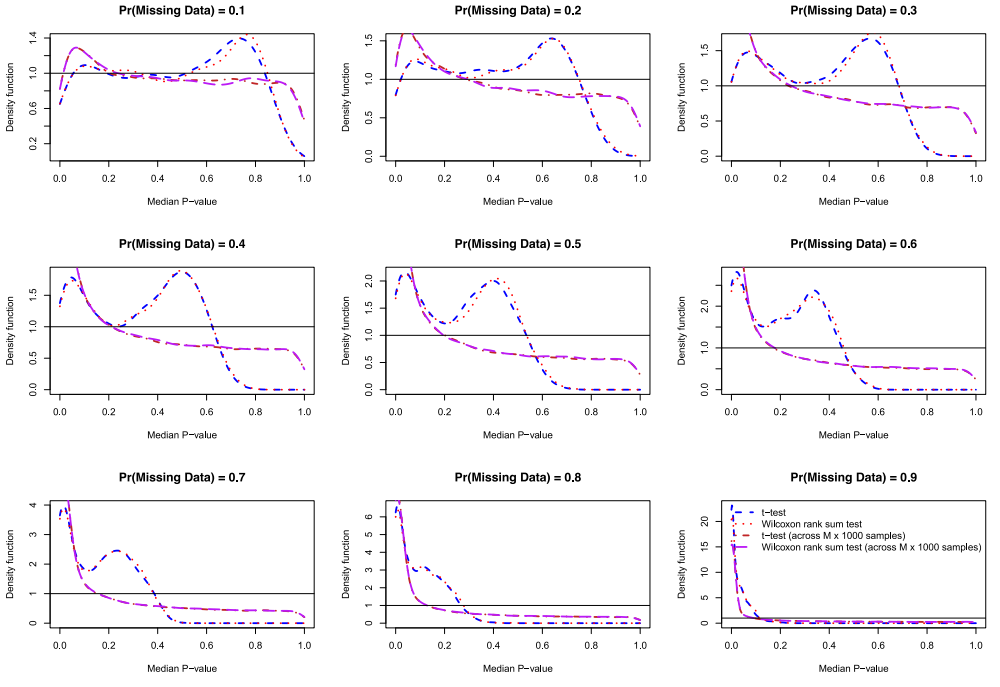

Figure 2. Empirical distribution of median *p*—values across simulations replicates: *t*—test and Wilcoxon rank sum test.

for the median *p*-value method when used the ANOVA and the Kruskal-Wallis test was higher than the advertised rate of 0.05. The empirical type I error rate was approximately 1 (or 100%) when the prevalence of missing data was very high. In contrast, the empirical type I error rates of the two tests in the absence of missing data had empirical type I error rates that were not different from the advertised rate of 0.05.

The non-parametric estimates of the densities of median *p*-values across the 1,000 simulation replicates are reported in Figure 3. The distribution of median *p*-values was not standard uniform in each of the nine scenarios. As above, there tended to be fewer large *p*-values than one would expect under a standard uniform distribution. The non-parametric estimates of the density functions for the *p*-values from the ANOVA and Kruskal-Wallis tests in the M x 1,000 imputed datasets are also reported in each panel of the figure. The shape of the deviations from uniformity was similar to Figure 2.

4.3. Pearson's correlation coefficient and Spearman's correlation coefficient

The relationship between the prevalence of missing data and the empirical type I error rate of the median *p*-value method are reported in Figure 1. For both correlation coefficients, the empirical type I error rate of the median *p*-value method was higher than the advertised rate of 0.05 across all scenarios. The empirical type I error rate, which increased with increasing prevalence of missing data, was high when the prevalence of missing data exceeded 0.6. The empirical type I error rates of the tests conducted in the complete data (prior to inducing missing data) were no different from the advertised rate of 0.05.

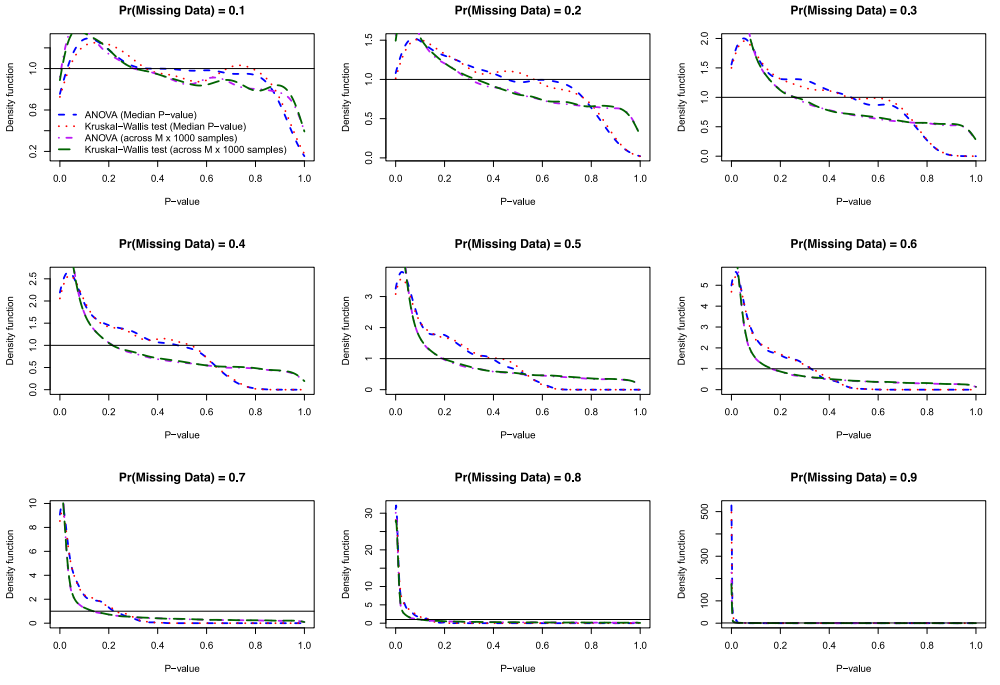

Figure 3. Empirical distribution of median *p*—values across simulations replicates: ANOVA and Kruskal—Wallis test.

The non-parametric estimates of the densities of median *p*-values across the 1,000 simulation replicates are reported in Figure 4. As with the previous tests, the empirical distribution of median *p*-values and M x 1,000 *p*-values were not standard uniform. The shape of the deviations from uniformity was similar to Figure 2.

4.4. The Chi-squared test

The relationship between the prevalence of missing data and the empirical type I error rate of the median *p*-value method are reported in Figure 1. The empirical type I error rate of the median *p*-value method when applied to the Chi-squared test was higher than the advertised rate. Furthermore, the empirical type I error rate increased as the prevalence of missing data increased. The empirical type I error rate was approximately 1 (100%) when the prevalence of missing data was 0.80. The empirical type I error rate of the Chi-squared test applied to the complete data (prior to inducing missingness) was not meaningfully different from the advertised rate.

The non-parametric estimates of the densities of median *p*-values across the 1,000 simulation replicates are reported in Figure 5. As with the previous tests, the empirical distribution of the median *p*-values was not standard uniform. There tended to be fewer large *p*-values than would be expected under a standard uniform distribution. The non-parametric estimates of the density functions for the *p*-values from the chi-squared tests in the M x 1,000 imputed datasets are also reported in each panel of the figure. As with the median *P*-value, there tended to be fewer large *p*-values than would be anticipated under

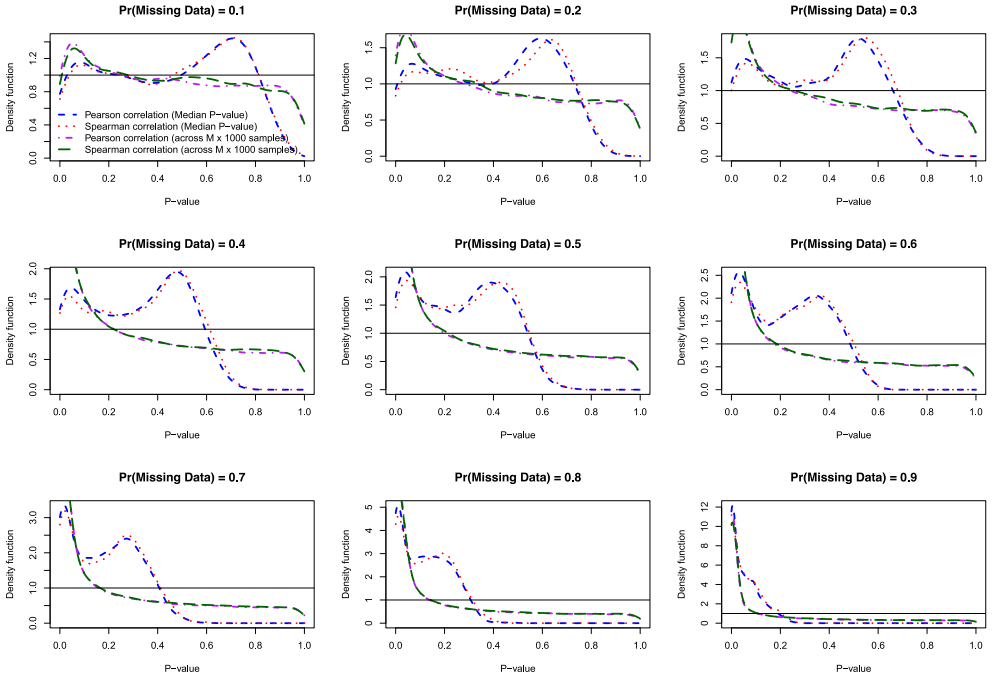

Figure 4. Empirical distribution of median *p*—values across simulations replicates: Pearson and Spearman correlation coefficient.

a standard uniform distribution. However, the magnitude of the discrepancy was less than was observed for the median *P*-values.

4.5. OLS regression

The relationship between the prevalence of missing data and the empirical type I error rate of the median *p*-value method are reported in Figure 1. The empirical type I error rates for the median *p*-value method applied to OLS regression were higher than the advertised rate of 0.05. While the empirical type I error rates increased with increasing prevalence of missing data, the amplification of the type I error rate was substantially less than was observed above. The use of Rubin's Rules across the M imputed dataset and the use of OLS regression in the complete data (before inducing missing data) resulted in empirical type I error rates that were no different from the advertised rate.

The non-parametric estimates of the densities of median *p*-values across the 1,000 simulation replicates are reported in Figure 6. The empirical distribution of the median *p*-values did not follow a standard uniform distribution, and shifted to the left for higher proportions of missing data. While the use of Rubin's Rules resulted in an empirical distribution of *p*-values that was closer to standard uniform, it did result in fewer very small *p*-values and very large *p*-values than would be expected under a standard uniform distribution. Similarly, the distribution of *p*-values across the M x 1,000 imputed samples had fewer large *p*-values than would be expected under a standard uniform distribution. However, the magnitude of the discrepancy was smaller than was observed for the median *p*-values.

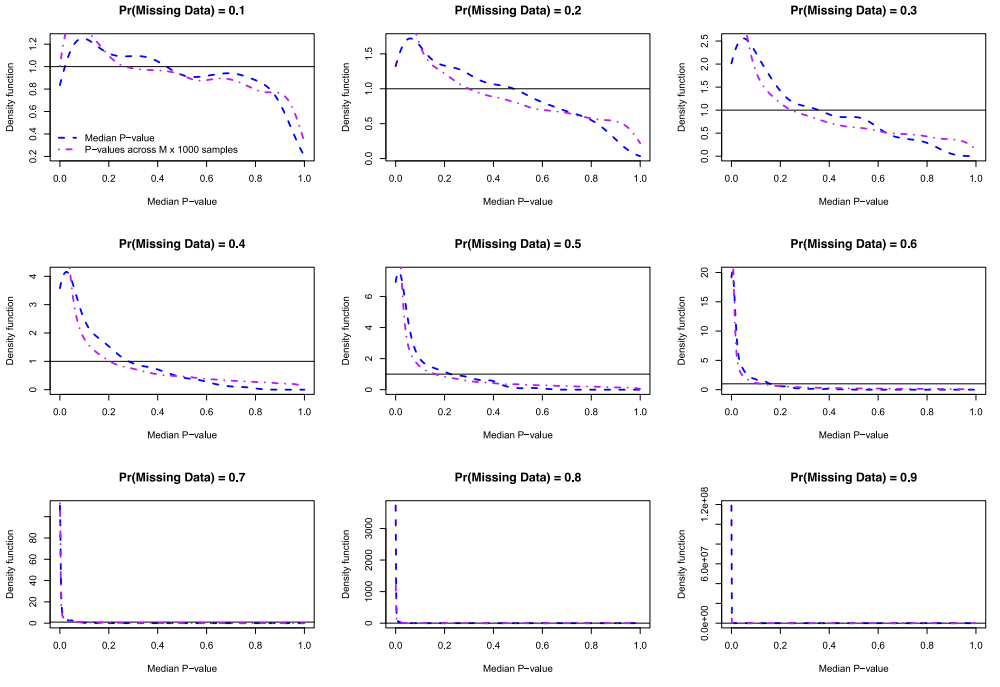

Figure 5. Empirical distribution of median *p*—values across simulations replicates: Chi—squared test.

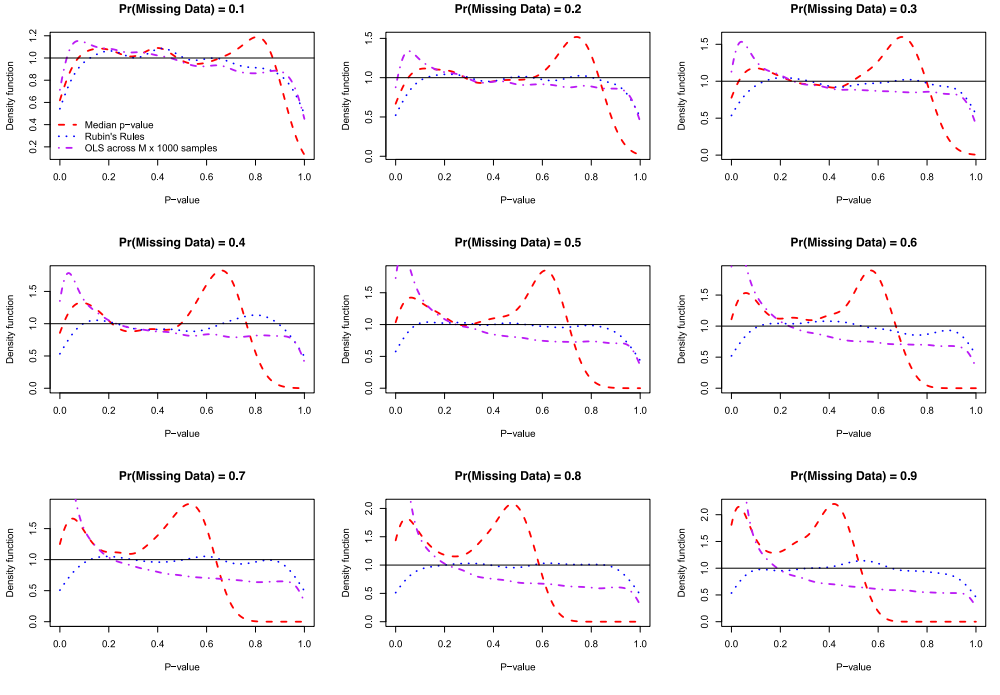

Figure 6. Empirical distribution of median *p*—values across simulations replicates: OLS regression.

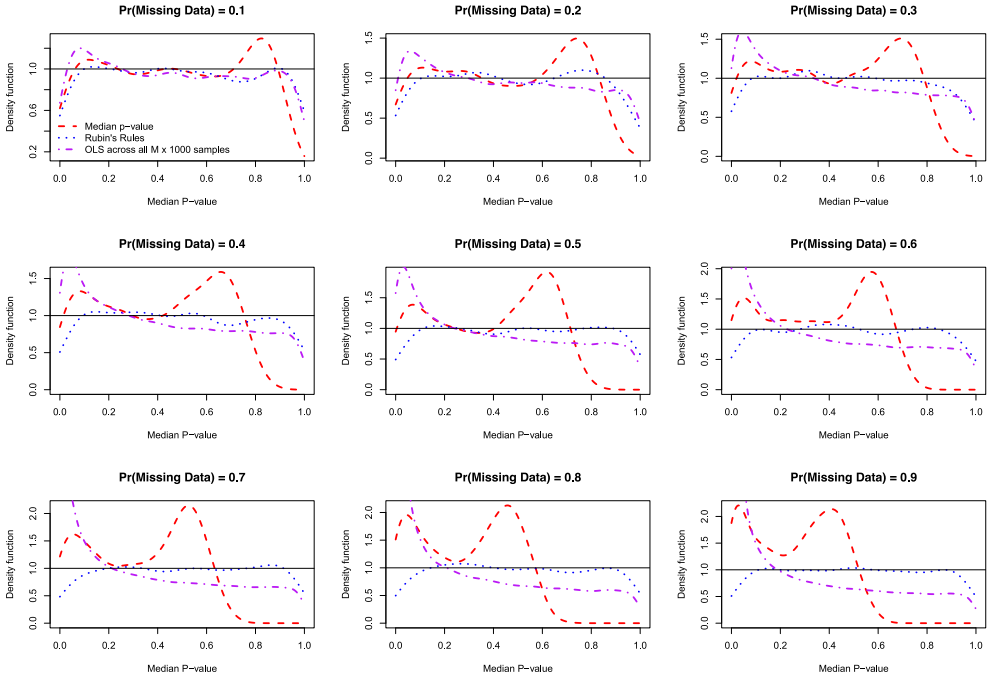

Figure 7. Empirical distribution of median p—values across simulations replicates: Logistic regression.

4.6. Logistic regression

The relationship between the prevalence of missing data and the empirical type I error rate of the median p-value method are reported in Figure 1. Results were very similar to those observed for OLS regression.

The non-parametric estimates of the densities of median p-values across the 1,000 simulation replicates are reported in Figure 7. Results were very similar to those observed for OLS regression.

5. Discussion

We evaluated the performance of the median p-value method across nine different statistical tests: Student's two sample t-test, the Wilcoxon Rank Sum test, Analysis of Variance (ANOVA), the Kruskal-Wallis test, the test of significance for Pearson's correlation coefficient, the test of significance for Spearman's correlation coefficient, the Chi-squared test, the test of significance for a regression coefficient from a linear regression model estimated using ordinary least squares (OLS), and the test of significance for a regression coefficient estimated from a logistic regression. Across all nine tests, we found that the median p-value method resulted in inflated type I error rates that exceeded the advertised rate.

To the best of our knowledge, only three prior studies have assessed the validity of the median p-value method. Eekhout and colleagues, who coined the term median p-value method, used simulations to assess the performance of the median p-value method for assessing the statistical significance of a categorical predictor variable in a logistic regression model [2]. In scenarios in which 25% of subjects had missing data, the empirical type

I error rate for testing the significance of a categorical variable, for which there truly was no effect, ranged from 0.065 to 0.077. When 40% of subjects had missing data, the empirical type I error rates ranged from 0.083 to 0.098. These estimates are qualitatively similar to those observed in the current study when examining logistic regression with continuous covariates. We hypothesize that Eekhout and colleagues would have observed larger type I error rates had they examined scenarios with higher prevalences of missing data. The empirical type I error rates observed by Eekhout and colleagues that we described above are when the outcome variable was included in the imputation model, as has been suggested elsewhere [11]. Eekhout *et al.* noted that, when the outcome variable was excluded from the imputation model, then the empirical type I error rate for the median *p*-value method was lower than the advertised rate (note that this is not an ideal solution as it entails double the work, with separate imputations for pooling test statistics (without the outcome in the imputation model) and separate imputations for pooling parameter estimates (with the outcome in the imputation model)). Bolt and colleagues used simulations to compare the performance of the median *p*-value method with that of competitor methods for pooling generalized additive models (GAMs) across imputed datasets [1]. When focusing on the single covariate for which the null hypothesis was true, the use of the median *p*-value method to pool the results of GAMs resulted in mildly inflated empirical type I error rates (0.06–0.08) (see Figures 2 and 3). When using predictive mean matching for MI, the D2 pooling method and a variant of D2 tended to result in empirical type I error rates that were closer to the advertised rate of 0.05. The authors simulated data such that data were missing for approximately 35% of subjects. Based on our findings, we hypothesize that the empirical type I error rate would increase as the prevalence of missing data increased. Panken and Heymans compared the median *p*-value method with three methods for variable selection for logistic regression models when using MI [3]. While the median *p*-value method was less complex and easier to implement than the competitor methods, it was shown to have performance that was at least as good as those of the competitor methods. The methods were compared using three metrics: (i) the selection frequency of variables; (ii) the agreement between the *p*-values of the selected variables to those obtained when the variable selection process was conducted in the original complete sample (prior to data being set to missing); (iii) the stability of the selected regression model. We note that this does not entail a formal evaluation of the type I error rate of the median *p*-value method. Rather, it examines the performance of the method when used for selecting variables for a logistic regression model.

Our study is subject to certain limitations. The primary limitation relates to our use of Monte Carlo simulations. Due to the computational complexity of simulations involving MI, we were only able to examine a limited number of scenarios for each statistical test. However, these limited numbers of scenarios were adequate to illustrate that the empirical type I error rate was higher than advertised in most scenarios and that the inflation increased as the prevalence of missing data increased. A secondary limitation was that we restricted our study to nine statistical tests. Due to space constraints and the computational complexity of our simulations, we were unable to include additional statistical tests. However, while the magnitude of the inflation in the type I error rate varied across statistical tests, our primary finding was consistent across tests: an inflation of the type I error rate and that this inflation increased as the prevalence of missing data increased.

In summary, across nine statistical statistics, the median *p*-value method resulted in empirical type I error rates that exceeded the advertised rate. For each of the nine tests, the magnitude of the inflation increased as the prevalence of missing data increased. For several tests, the empirical type I error rate was very high when the prevalence of missing data was high. We would suggest that, while it may be applicable in specific settings, the median *p*-value method not be used to assess statistical significance across imputed datasets. Rather than using the median *p*-value method, we suggest that authors use formal methods that have been proposed for pooling test statistics across imputed samples [8].

Authors' contributions

PCA conceived the study, conducted the simulations, wrote the manuscript, and approved the final manuscript. SvB and IE provided input on the design of the simulations, revised the manuscript, and approved the final manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethics approval and consent to participate

The study used Monte Carlo simulations using randomly generating data. No person-level data were used in this study. There was no direct or indirect human participation involved in this study.

Funding

ICES is an independent, non-profit research institute funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). As a prescribed entity under Ontario's privacy legislation, ICES is authorized to collect and use health care data for the purposes of health system analysis, evaluation and decision support. Secure access to these data is governed by policies and procedures that are approved by the Information and Privacy Commissioner of Ontario. The use of the data in this project is authorized under section 45 of Ontario's Personal Health Information Protection Act (PHIPA) and does not require review by a Research Ethics Board. This document used data adapted from the Statistics Canada Postal CodeOM Conversion File, which is based on data licensed from Canada Post Corporation, and/or data adapted from the Ontario Ministry of Health Postal Code Conversion File, which contains data copied under license from ©Canada Post Corporation and Statistics Canada. Parts of this material are based on data and/or information compiled and provided by CIHI and the Ontario Ministry of Health. The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOH or MLTC is intended or should be inferred. This research was supported by operating grant from the Canadian Institutes of Health Research (CIHR) (PJT 166161).

Data availability statement

No data were used in the current study. The current study consisted entirely of Monte Carlo simulations.

References

[1] M.A. Bolt, S. MaWhinney, J.W. Pattee, K.M. Erlandson, D.B. Badesch, and R.A. Peterson, *Inference following multiple imputation for generalized additive models: an investigation of the median p-value rule with applications to the Pulmonary Hypertension Association Registry and Colorado COVID-19 hospitalization data*. BMC Med. Res. Methodol. 22 (2022), pp. 148.

[2] I. Eekhout, M.A. van de Wiel, and M.W. Heymans, *Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: power and applicability analysis*. BMC Med. Res. Methodol. 17 (2017), pp. 129.

[3] A.M. Panken, and M.W. Heymans, *A simple pooling method for variable selection in multiply imputed datasets outperformed complex methods*. BMC Med. Res. Methodol. 22 (2022), pp. 214.

[4] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2005.

[5] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, 1987.

[6] R. Schouten, P. Lugtig, J. Brand, and G. Vink, *Generate Missing Values with Ampute*, 2022 [cited 2024 June 6, 2024]. Available at: https://rianneschouten.github.io/mice_ampute/vignette/ampute.html.

[7] R.M. Schouten, P. Lugtig, and G. Vink, *Generating missing values for simulation purposes: a multivariate amputation procedure*. J. Stat. Comput. Simul. 88 (2018), pp. 2909–2930.

[8] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed., CRC Press, Boca Raton, FL, 2018.

[9] S. van Buuren, *Multiple imputation of discrete and continuous data by fully conditional specification*. Stat. Methods Med. Res. 16 (2007), pp. 219–242.

[10] S. van Buuren, and K. Groothuis-Oudshoorn, *mice: multivariate imputation by chained equations in R*. J. Stat. Softw. 45 (2011).

[11] I.R. White, P. Royston, and A.M. Wood, *Multiple imputation using chained equations: issues and guidance for practice*. Stat. Med. 30 (2011), pp. 377–399.