Statistica Sinica Preprint No: SS-2026-0062	
Title	Combining Rules for Chi-square Statistics from Multiply-imputed Data
Manuscript ID	SS-2026-0062
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0062
Complete List of Authors	Ashok K Chaurasia and Mohsen Ebadi
Corresponding Authors	Ashok K Chaurasia
E-mails	achaurasia.uconn@gmail.com
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

COMBINING RULES FOR CHI-SQUARE STATISTICS FROM MULTIPLY-IMPUTED DATA

Ashok Chaurasia* and Mohsen Ebadi

University of Waterloo

Abstract: Multiple Imputation (MI) is a popular method for dealing with missing data in applications. A critical step in MI involves using Rubin's rules to combine statistics from the analysis of imputed data. However, Rubin's rules require a sufficient condition that the estimator for the parameter of interest $\boldsymbol{\theta} \in \mathbb{R}^k$ must be (approximately) multivariate normal. This condition is problematic for non-normal statistics. We propose new rules for combining Chi-square statistics (obtained from analysis of each imputed data) by first transforming them to normal, and then combining them using Rubin's univariate combining rules. Simulation results demonstrate that our proposed methodology controls Type I error rates and provides superior statistical power in small-to-moderate sample sizes ($n > k$, but not $n \gg k$). The proposed method reduces the computational burden from $O(k^3)$ to $O(1)$ for inference about $\boldsymbol{\theta}$, and eliminates reliance on the assumption of *proportional and equal loss of information*, which is assumed in Rubin's traditional multivariate combining rules to obtain a stable, non-singular variance-covariance matrix. These advantages make the proposed method particularly effective for missing data applications with small-to-moderate sample sizes.

Key words and phrases: Chi-square statistics, Combining rules, Missing at random, Missing data, Multiple imputation, Normal transformation, Proportional and equal loss of information.

*Corresponding Author, Email: a4chaura[at]uwaterloo[dot]ca
ORCID: https://orcid.org/0000-0001-6745-0493

Statistica Sinica: Newly accepted Paper

1. Introduction

Missing data are a common problem in data analysis. A variable's value for an individual or experimental unit is considered missing if its value is not observed. Missing data result in loss of information for the parameter of interest (say, $\boldsymbol{\theta} \in \mathbb{R}^k$), which complicates subsequent analysis. The goal of handling missing data is to mitigate this loss of information (about $\boldsymbol{\theta}$). Traditional approaches methods like complete case analysis (CCA) use only complete cases, thereby mishandling the missing data problem by ignoring it. Such approaches typically introduce greater bias and higher variance, reducing both the accuracy and precision of parameter estimates. For CCA, the bias and imprecision worsen as the missingness rate increases. Multiple Imputation (MI; Rubin, 1987), is a widely used technique for handling missing values. As a Monte Carlo method, MI replaces each missing value with a simulated value to create a set of complete datasets that facilitate standard analysis. The MI procedure consists of three phases. In the first phase — called the imputation phase — each missing value is replaced with a value from a plausible set of values under a set of imputation models. This process generates several complete-data versions (multiply-imputed datasets) that jointly serve as a surrogate for the original incomplete data. These multiply-imputed data reflect the additional uncertainty/variability introduced in the original incomplete data due to missingness. In the second phase — the analysis phase — individual imputed data are analyzed separately for the common model of interest, and their respective parameter point and variance estimates are collected. In the third and final phase — the combining

stage — the collected point and variance estimates are combined using Rubin's rules to yield statistics for overall inference for the parameter of interest.

This framework of MI relies on certain assumptions about the nature of missingness (i.e., its pattern and mechanism) and the parameter of interest for Rubin's rules to apply and yield reliable statistical inference. Extensions of Rubin's rules include pooling methods for multivariate parameters (Reiter, 2007; Barnard and Rubin, 1999; Li et al., 1991; Rubin, 1987), likelihood ratio (LR) tests (Chan and Meng, 2022; Meng and Rubin, 1992), marginal or partial Pearson's correlation coefficient (Schafer, 1997), ANOVA results (Grund et al., 2016; Raghunathan and Dong, 2011), Beta- and F-statistics (Chaurasia, 2023), coefficient of determination (van Ginkel, 2019).

A sufficient condition for applying Rubin's rules is that the estimate for the parameter of interest is (approximately) normally distributed. Another factor that plays a critical role in applying these combining rules is that missingness is not to the extent that it creates an unstable (i.e., singular) variance-covariance matrix. The normality condition is problematic because many statistics of interest have non-normal distributions. Furthermore, instability of the total variance-covariance matrix in MI becomes particularly problematic as the rate of missingness increases, and the sample size of incomplete data n is comparable to k such that $n > k$, but not $n \gg k$.

One way this instability is addressed in MI is by assuming a *proportional and equal loss of information* (p.e.l.i) across the k-components of $\boldsymbol{\theta}$. However, violating one or more of these conditions and assumptions can lead to invalid and unreliable statistical inferences.

For example, Chan and Meng (2022) highlighted that the existing method by Meng and Rubin (1992) for combining LR test statistics — which should follow an F-distribution — can yield negative statistic values, failing to ensure monotonic power.

In this article, we propose a simple method to address the aforementioned statistical and computational issues by applying transformation functions before using Rubin's univariate combining rules. The normal transformation functions satisfy the sufficient condition for applying Rubin's rules, while the univariate rules eliminate reliance on *p.e.l.i.* by avoiding computations involving matrix multiplication and inversion.

To demonstrate the proposed method, we focused on rules for combining statistics from multiply-imputed data that follow an approximate chi-squared distribution. While any normal approximation of a chi-square random variable can be used in principle, we used four specific transformations that have demonstrated superior performance and compared their inferential reliability. A key practical advantage of our proposed methodology is that it only requires access to the chi-square test statistics after each imputed dataset has been analyzed, thereby alleviating the computational burden for the overall inference in MI. This methodology is significantly simpler than existing approaches, such as those by Chan and Meng (2022); Meng and Rubin (1992), which require access to both parameter estimates and imputed data for multiple passes to compute the final LR test statistic. Since LR test statistics are asymptotically chi-squared, our methodology can be easily adopted to combine these results from multiply-imputed data.

The remainder of the paper is organized as follows. Section 2 provides an overview of

MI combining rules for univariate and multivariate parameters. Section 3 outlines the necessary preliminaries for transforming chi-square random variables to an approximate normal distribution, and Section 4 provides the motivation for the proposed transformations. The proposed rules for combining chi-square statistics are introduced in Section 5. Simulation studies evaluating and comparing the performance of these methods are presented in Section 6. Section 7 discusses various areas of application where the proposed methodology can easily be adopted for multiply-imputed data. A data application of the proposed method is presented in Section 8 in the context of the Mantel-Haenszel test to assess the association between infant survival and prenatal care level while controlling for clinic. Finally, the article concludes with closing remarks in Section 9.

2. Background and Notation for Multiple Imputation

Suppose $\boldsymbol{\theta} \in \mathbb{R}^k$ is the parameter of interest for incomplete data $\mathcal{D}_{inc} = (\mathcal{D}_{obs}, \mathcal{D}_{mis})$, where $\mathcal{D}_{obs}$ and $\mathcal{D}_{mis}$ are the observed and missing parts of $\mathcal{D}_{inc}$. In the traditional application of MI in `R` via package `mice`, the imputation stage involves specification of imputation models for each incomplete variable in the data (Van Buuren and Groothuis-Oudshoorn, 2011, see Fully Conditional Specification (FCS)). The imputation phase results in m imputed datasets denoted by the set $\{D_j : j = 1, \ldots, m\}$ that serve as a surrogate for $\mathcal{D}_{inc}$. The second phase of MI involves analysis of each imputed data D_j for the common model of interest and collecting their respective point and variance estimates for $\boldsymbol{\theta}$, indexed by imputation number $j = 1, \ldots, m$. Suppose these point and variance estimates are denoted by sets

$\{Q_j : j = 1, \ldots, m\}$, and $\{U_j : j = 1, \ldots, m\}$, respectively. The last phase of MI involves combining point and variance estimates $\{(Q_j, U_j) : j = 1, \ldots, m\}$ using Rubin's combining rules for overall inference for $\boldsymbol{\theta}$. These combining rules are discussed next.

For parameter $\boldsymbol{\theta}$, the multivariate combining rules presented here use degrees of freedom estimates proposed by Barnard and Rubin (1999) for inference about $\boldsymbol{\theta}$. For these rules it is presumed that the incomplete data were imputed m times under the ignorable missingness mechanism (Little and Rubin, 2002), which assumes that the data are Missing at Random (MAR) along with *parameter distinctness* (i.e., independence between $\boldsymbol{\theta}$ and the missingness parameter denoted by ϕ).

For the hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, if $(Q_j - \boldsymbol{\theta}_0) \sim \mathcal{N}_k(\mathbf{0}, U_j)$, then Rubin's combining rules (Rubin, 1987) for overall point and variance estimates (for $\boldsymbol{\theta}$) from MI are given as follows.

$$\text{Point estimate: } \bar{Q} = m^{-1}\sum_{j=1}^{m} Q_j; \tag{2.1}$$

$$\text{Within-data variance-covariance: } \bar{U} = m^{-1}\sum_{j=1}^{m} U_j; \tag{2.2}$$

$$\text{Between-data variance-covariance: } B = (m-1)^{-1}\sum_{j=1}^{m} (Q_j - \bar{Q})(Q_j - \bar{Q})'; \tag{2.3}$$

$$\text{Total variance: } T = \bar{U} + (1 + m^{-1})B; \tag{2.4}$$

$$\text{(Average) Relative increase in variance: } r = k^{-1}(1 + m^{-1})\ \mathrm{tr}(B\bar{U}^{-1}); \tag{2.5}$$

$$\text{(Stable) Total variance-covariance: } \bar{U}_* = (1 + r)\bar{U}. \tag{2.6}$$

Rubin (1987) proposed $\bar{U}_*$ as a stable total variance-covariance estimate since T can get unstable (i.e., singular due to a dominant B that is, by definition, singular) for modest

proportions of missingness and m. The definition of $\bar{U}_*$ assumes a *proportional and equal loss of information* (*p.e.l.i.*) across the k-components of $\boldsymbol{\theta}$. MI inference for the hypothesis of interest is given by the Wald statistic

$$D_m = k^{-1} \left(\bar{Q} - \boldsymbol{\theta}_0\right)' \bar{U}_*^{-1} \left(\bar{Q} - \boldsymbol{\theta}_0\right), \tag{2.7}$$

which is referenced to the F-distribution with k numerator degrees of freedom and v_{BR}^* denominator degrees of freedom by Barnard and Rubin (1999). The denominator degrees of freedom need to be estimated since the original data is incomplete, and is given by $v_{\text{BR}}^* = (v_m^{-1} + v_{\text{obs}}^{-1})^{-1}$, where $v_m = (m-1)\gamma^{-2}$, $v_{\text{obs}} = (1-\gamma)\ v_{\text{com}}\ (v_{\text{com}}+1)\,/\,(v_{\text{com}}+3)$, $\gamma = k^{-1}(1+m^{-1})\ \text{tr}(BT^{-1})$, and v_{com} is the denominator degrees of freedom that would have been used if the original data had no missing values.

Lastly, Rubin's univariate combining rules stem from using $k = 1$ in the multivariate combining rules with the key results that $T = \bar{U}_*$ and that T is always non-singular.

3. Normal transformation functions for Chi-square random variables

This section covers the normal transformations for chi-squared random variables and serves as a preliminary for Sections 4 and 5. For $D \sim \chi_\nu^2$ with ν degrees of freedom, we considered the following normal approximations, where $F_D(\cdot)$ and $\Phi(\cdot)$ respectively denote the cumulative distribution functions of the chi-squared distribution (with ν degrees of freedom) and the standard normal distribution.

For $\nu > \frac{2}{3}$, Wilson and Hilferty (1931, WH) proposed the normal approximation

$$Z = \frac{X - \mu}{\sigma}, \tag{3.8}$$

where $X = \sqrt[3]{\frac{D}{\nu}}$, $\mu = 1 - \frac{2}{9\nu}$, and $\sigma = \sqrt{1-\mu}$ such that $F_D(d) \approx \Phi(z)$.

For $\nu > 1$, Peizer and Pratt (1968, PZ) proposed the standard normal approximation

$$Z = \begin{cases} Z_1 & \text{if } D = \nu - 1 \\ Z_2 & \text{if } D \neq \nu - 1 \end{cases}, \tag{3.9}$$

where $Z_1 = \frac{-\left(\frac{1}{3} + \frac{0.08}{\nu}\right)}{\sqrt{2\nu - 2}}$, $Z_2 = \left(\frac{D - \nu + \frac{2}{3} - \frac{0.08}{\nu}}{|D - \nu + 1|}\right)\sqrt{(\nu-1)\log\left(\frac{\nu-1}{D}\right) + D - (\nu - 1)}$ such that $F_D(d) \approx \Phi(z)$.

For $\nu > \frac{\sqrt{\frac{13}{6}}}{6} - \frac{1}{18}$, Canal (2005, LC) proposed the normal approximation

$$Z = \frac{X - \mu}{\sigma}, \tag{3.10}$$

where $X = \left(\frac{D}{\nu}\right)^{\frac{1}{6}} - \frac{1}{2}\left(\frac{D}{\nu}\right)^{\frac{1}{3}} + \frac{1}{3}\left(\frac{D}{\nu}\right)^{\frac{1}{2}}$, $\mu = \frac{5}{6} - \frac{1}{9\nu} - \frac{7}{648\nu^2} + \frac{25}{2187\nu^3}$, and $\sigma = \sqrt{\frac{1}{18\nu} + \frac{1}{162\nu^2} - \frac{37}{11664\nu^3}}$ such that $F_D(d) \approx \Phi(z)$.

For $\nu > 0$, Goria (1992, MG) proposed the normal approximation

$$Z = \frac{X - \mu}{\sigma}, \tag{3.11}$$

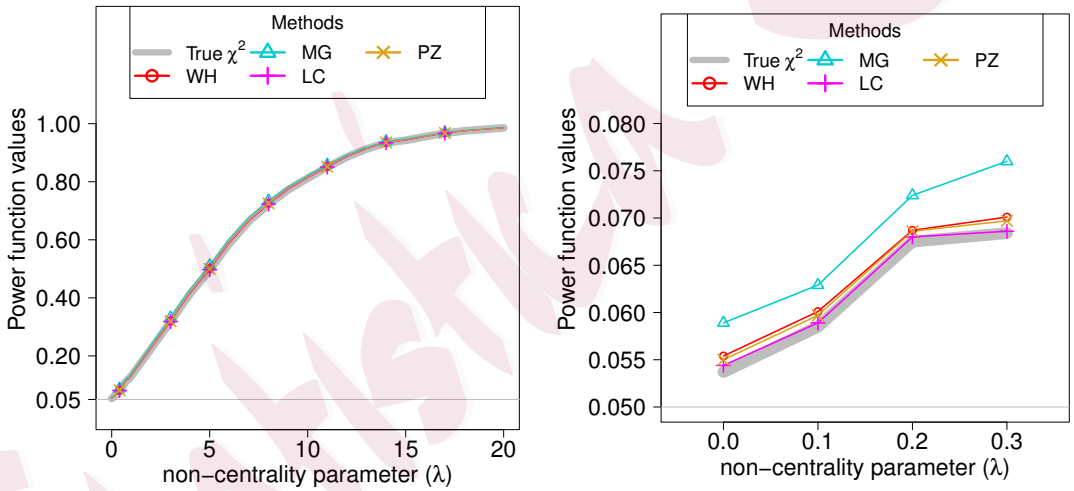
where $X = \left(\frac{D}{\nu}\right)^{\frac{1}{2}} + 4\left(\frac{D}{\nu}\right)^{\frac{1}{4}}$, $\mu = 5 - \nu^{-1} - \frac{3}{128\nu^2} + \frac{311}{20048\nu^3}$, and $\sigma^2 = \frac{9}{2\nu} + \frac{1}{8\nu^2} - \frac{207}{256\nu^3}$ such that $F_D(d) \approx \Phi(z)$.

4. Motivation

To assess the potential and precision of the transformation functions covered in Section 3, we conducted a simple simulation study in `R` to compare them via their power curves. This simulation generated random values from a chi-square distribution with increasing values for the non-centrality parameter λ. Specifically, a matrix X with $N = 10,000$ rows and $\ell = 30$ columns was generated from a chi-squared distribution with fixed degrees of freedom ν. Rows of X represent individual simulation runs, while the columns correspond to non-centrality values: $\lambda =$ (0 to 0.9 by increments of 0.1, and 1 to 20 by increments of 1). For the significance level of $\alpha = 0.05$, the decision matrix for X, denoted by D_X, was obtained under the null hypothesis of a central chi-square distribution (with ν degrees of freedom) such that 1=$\{P(\chi^2_\nu > X) < \alpha\}$, and 0=otherwise. The power curve for X was obtained by applying the `colMeans` function to D_X. To evaluate the power curves for transformations from Section 3, X was first transformed via each transformation function as $f(X)$, and then their corresponding decisions (denoted as D_f) were recorded under a standard normal distribution.

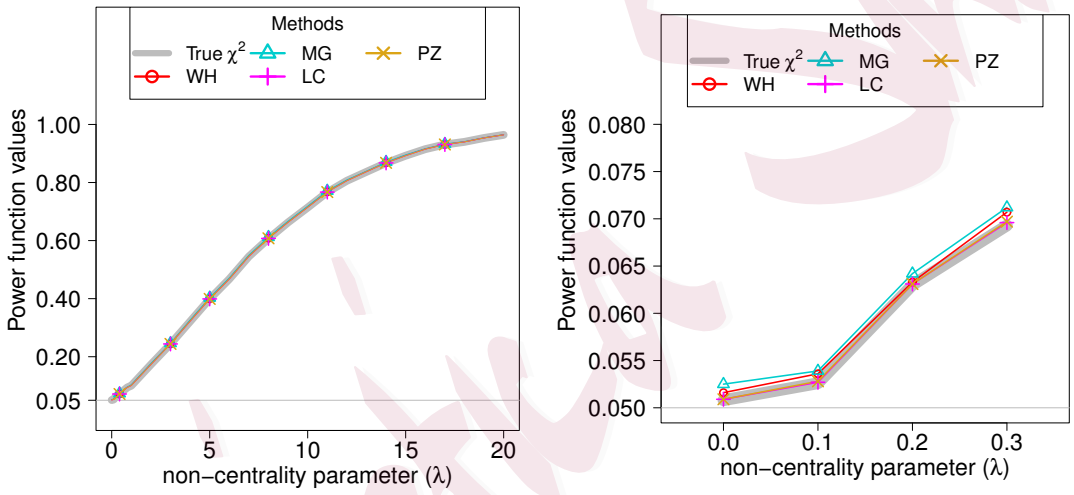
Results from this small simulation study are shown in plots of Figure 1 for $\nu = 2$, and Figure 2 for $\nu = 4$. In these figures, panel (b) presents a magnified view of panel

(a), focusing specifically on the X-axis where $\lambda = 0,\ 0.1,\ 0.2,\ 0.3$. These figures suggest that all transformations from Section 3 qualify as α-level tests and are hence comparable, with preference order (from highest to lowest power) for transformations MG, WH, PZ, and then LC with respect to the true power curve (shown in grey) of the Chi-square distributed random variable (with ν degrees of freedom) for all λ values. It is also noted that as ν increases, the separation between the power curves of transformations diminishes with preference order preserved, thereby suggesting that the proposed transformation functions are reliable in applications where ν is large.



(a) (b)

Figure 1: Power function curves for $\chi^2(\nu, \lambda)$ using transformations WH, PZ, MG, LC for degrees of freedom $\nu = 2$ and non-centrality parameter (a) $\lambda \in [0, 20]$ and (b) $\lambda \in [0, 0.3]$.



(a) (b)

Figure 2: Power function curves for $\chi^2(\nu, \lambda)$ using transformations WH, PZ, MG, LC for degrees of freedom $\nu = 4$ and non-centrality parameter (a) $\lambda \in [0, 20]$ and (b) $\lambda \in [0, 0.3]$.

5. Combining rules for Chi-squared statistics from Multiply-Imputed data

5.1 Notation and Computation

Similar to Section 2, suppose $\boldsymbol{\theta} \in \mathbb{R}^k$, $\mathcal{D}_{inc} = (\mathcal{D}_{obs}, \mathcal{D}_{mis})$, and $\mathcal{D}_{inc}$ is imputed m times. Now, let the point and variance estimates for $\boldsymbol{\theta}$ from m imputed datasets be represented by sets $\{Q_j : j = 1, \ldots, m\}$ and $\{U_j \equiv V(Q_j) : j = 1, \ldots, m\}$, respectively. Next, let Wald test-statistics for the hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ be represented by the set $\{W_j \equiv (Q_j - \boldsymbol{\theta}_0)' U_j^{-1} (Q_j - \boldsymbol{\theta}_0) : j = 1, \ldots, m\}$, where $W_j \sim \chi^2_k$ (asymptotically).

The proposed method first requires transforming W_j (for all j) to a normally distributed random variable and then applying Rubin's univariate combining rules to pool the transformed random variables while using the denominator degrees of freedom estimate by Barnard and Rubin (1999) for inference about $\boldsymbol{\theta}$. Specifically, we have the normal transformation $f(\cdot)$ of W_j using one of the functions in Section 3 to yield the set $\{\psi_j \equiv f(W_j) : j = 1, \ldots, m\}$ on which we apply Rubin's univariate combining rules to yield the statistic:

$$D_m^{(\psi)} = \sigma_f^{-2} \left(\bar{\psi} - \mu_f\right)^2, \tag{5.12}$$

where $\bar{\psi} = m^{-1} \sum_{j=1}^{m} \psi_j$, with μ_f and σ_f determined under transformation f from Section 3. Under the null hypothesis, $D_m^{(\psi)}$ is referenced to the F-distribution with 1 and v_{BR} degrees of freedom, where v_{BR} is derived from the set of transformed statistics $\{\psi_j\}_{j=1}^m$. Note that under the proposed method, each ψ_j is exactly normal such that it yields approximately the same tail probabilities as the original W_j, which is chi-squared distributed.

Statistica Sinica: Newly accepted Paper

5.2 Theoretical Justification

Inference for $\boldsymbol{\theta}$ using traditional multivariate combining rules, with v_{BR}^* as the denominator degrees of freedom, is performed via the test statistic $D_m := k^{-1}\left(\bar{Q}-\boldsymbol{\theta}\right)' \bar{U}_*^{-1}\left(\bar{Q}-\boldsymbol{\theta}\right)$ where $\bar{Q} = m^{-1}\sum_{j=1}^{m} Q_j$, $\bar{U} = m^{-1}\sum_{j=1}^{m} U_j$ is the within-imputation variance estimate, $\bar{U}_* = (1+r_m)\bar{U}$ is the *stable* non-singular total variance-covariance (matrix) estimate, and r_m is the relative increase in variance (scalar) estimate. Barnard and Rubin (1999) showed that D_m achieves nominal coverage rates, denoted as ζ, under the (approximate) reference F distribution such that for moderate m,

$$\lim_{n\to\infty} \Pr_{\boldsymbol{\theta}}\left(F_{k,v_{\text{BR}}^*} > D_m\right) = \zeta. \tag{5.13}$$

Inference for $\boldsymbol{\theta}$ can be obtained via the test statistic $D_m^{(\psi)}$ with v_{BR} as the denominator degrees of freedom using $\{\psi_j\}_{j=1}^m$. From Barnard and Rubin (1999) we have that $D_m^{(\psi)}$ achieves nominal coverage of ζ under the (approximate) reference F distribution such that for moderate m,

$$\lim_{n\to\infty} \Pr_{\boldsymbol{\theta}}\left(F_{1,v_{\text{BR}}} > D_m^{(\psi)}\right) = \zeta. \tag{5.14}$$

From Eqs. (5.13) and (5.14), we have for moderate m and $n\to\infty$ that

$$\left|\Pr_{\boldsymbol{\theta}}\left(F_{1,v_{\text{BR}}} > D_m^{(\psi)}\right) - \Pr_{\boldsymbol{\theta}}\left(F_{k,v_{\text{BR}}^*} > D_m\right)\right| = o(1). \tag{5.15}$$

Eq. (5.15) implies that inferential decision rates about $\boldsymbol{\theta}$ when using traditional D_m and proposed $D_m^{(\psi)}$, with their respective reference distributions, are asymptotically the same for large n and moderate to large m. In Eqs. (5.13)–(5.15) a large m yields a stable total variance estimate under MAR missingness mechanism. Additionally, for Eq. (5.13) there is the *p.e.l.i.*

Statistica Sinica: Newly accepted Paper

assumption across components of $\boldsymbol{\theta}$ for $\bar{U}_*$ to serve as a *stable* non-singular surrogate matrix for the total variance-covariance of $\hat{\boldsymbol{\theta}}$, which is not needed under the proposed method with $D_m^{(\psi)}$ where ψ_j are univariate and normally distributed by definition. In Eqs. (5.13)–(5.15), a large n yields the property of statistical consistency that allows for asymptotic normality of the related estimator.

5.3 Some conceptual, theoretical, and practical comparisons to existing methods

Our proposed method can be compared to methods of Chan and Meng (2022, CM) and Meng and Rubin (1992, MR). In these comparisons, our proposed method has the advantage of being simpler in terms of computation and not being reliant on additional or stronger assumptions. For example, consider the D_L statistic of MR, which is computationally intensive because it requires two passes of parameter estimates through the LR function for each individual imputed data to compute their r_L as a difference of two LR estimates. Additionally, D_L is imprecisely referenced to an F-distribution with the cause of imprecision being that D_L is not strictly positive because of r_L being computed as a difference (Chan and Meng, 2022). In contrast, our proposed statistic $D_m^{(\psi)}$ only requires a single pass through multiply-imputed data to get access to Wald statistics W_j, and is guaranteed to be positive as per its definition in Eq. (5.12). In MR, the authors showed that $D_L = D_m + o(1)$ almost surely, which means statistics D_L and D_m are asymptotically the same. In contrast, with D_m and our proposed method $D_m^{(\psi)}$, the focus is on their respective decision rates such that

$\Pr_{\boldsymbol{\theta}}\left(F_{1,v_{\text{BR}}} > D_m^{(\psi)}\right) = \Pr_{\boldsymbol{\theta}}\left(F_{k,v_{\text{BR}}^*} > D_m\right) + o(1)$ absolutely, as in (5.15).

Recognising that r_L of MR is not strictly positive, Chan and Meng (2022) proposed replacing r_L with $r_L^+ = \max\{0, r_L\}$ when computing D_L; let this be denoted as D_L^+. However, this modification forces r_L to be positive, which has the effect of imposing $\Pr(r_L < 0) = 0$ and consequently costing a loss in its power. The authors also proposed another alternative method, denoted as $\hat{D}_L$, for combining LR statistics from multiply-imputed data that guarantees $\hat{D}_L \geq 0$ (Chan and Meng, 2022, see Eqs. 2.7 and 2.8). However, this method remains computationally intensive (due to parameter estimation from optimization over stacked imputed datasets, and two passes through the LR function similar to MR), with its corresponding $\hat{r}_L$ (still) not guaranteed to be non-negative for a finite number of imputations and requires stronger assumptions (Chan and Meng, 2022, pp. 1498-1500).

6. Simulation Study: Likelihood Ratio Test for Logistic Regression

This simulation study used the framework of modelling a binary outcome using a logistic regression model to assess the reliability of our proposed method. The data generation process is based on the model

$$\underset{n\times 1}{\boldsymbol{\eta}} = \underset{n\times k}{X}\ \underset{k\times 1}{\boldsymbol{\beta}} \tag{6.16}$$

where

$\eta_i = log\left(\dfrac{\Pr(y_i = 1)}{\Pr(y_i = 0)}\right)$ for $i = 1, 2, \ldots, n$;

y_i is a binary random variable such that $y_i \sim Bernoulli(p_i)$ with $p_i = (1 + \exp\{-\eta_i\})^{-1}$;

X is the design matrix for n observations and k predictors with each predictor generated from Uniform(0,1) and independent of other predictors; and

$\boldsymbol{\beta}$ is the fixed vector of regression coefficients for the design matrix X, which changes its values to reflect the data configurations of (i) no association between $\boldsymbol{y}$ and X (labelled as "zero-association"), and (ii) non-zero association between $\boldsymbol{y}$ and X (labelled as "non-zero-association"). Values of $\boldsymbol{\beta}$ by configuration are as follows:

$$\boldsymbol{\beta} = \begin{cases} (0,0,0)' & \text{for } k=3 \text{ and zero-association} \\ (0,0,0,0,0)' & \text{for } k=5 \text{ and zero-association} \\ (-1.2,-1.2,2)' & \text{for } k=3 \text{ and non-zero-association} \\ (-1,-1,2.2,-1,-1)' & \text{for } k=5 \text{ and non-zero-association} \end{cases} \tag{6.17}$$

The objective of interest is to test the following hypotheses

$$H_0 : \begin{cases} \boldsymbol{\beta}_{\{1,3\}|\{2\}} = \mathbf{0} & \text{for } k=3 \\ \boldsymbol{\beta}_{\{1,3,4,5\}|\{2\}} = \mathbf{0} & \text{for } k=5 \end{cases} \text{ vs. } H_a : \begin{cases} \boldsymbol{\beta}_{\{1,3\}|\{2\}} \neq \mathbf{0} & \text{for } k=3 \\ \boldsymbol{\beta}_{\{1,3,4,5\}|\{2\}} \neq \mathbf{0} & \text{for } k=5 \end{cases}, \tag{6.18}$$

where the notation $\boldsymbol{\beta}_{A|B}$ denotes that the components of $\boldsymbol{\beta}$ within set A are zero conditional on the components of $\boldsymbol{\beta}$ within B being non-zero with $A \cap B = \{\}$. In other words, $\boldsymbol{\beta}_{A|B}$ implies the full model design matrix contains columns (i.e., predictor variables) corresponding to regression coefficients of set $A \cup B$, while the reduced model design matrix only contains columns corresponding to regression coefficients of set B.

We considered various data configurations based on combinations of associations between $\boldsymbol{y}$ and X (zero vs. non-zero association), sample size $n = \{50, 100\}$, number of predictors $k = \{3, 5\}$, and percentage of missingness $\delta = \{5, 10, 15, 20, 25\}$. For each combination of (hypothesis,n,k,δ), data were generated and LR test was performed using (i) the fully observed data, (ii) complete cases when missingness was induced, and (iii) m-imputed data with imputation size $m = 100$.

For creating missing values in the simulated data, we employed a Missing At Random (MAR) mechanism as follows. For a certain percentage of missingness $\delta \in (0, 1)$, the $\delta \times 100^{th}$ greatest integer percentile of $\boldsymbol{x}_1$ was recorded; denote this as $\boldsymbol{x}_1(\delta)$. All cases with their x_1-value less than or equal to $\boldsymbol{x}_1(\delta)$ are recorded as cases where missingness will occur; let these cases be denoted by the set $\mathcal{I}$. Next, among cases in $\mathcal{I}$, (approximately) the first half of these cases are made to have their Y-value set to be missing, while (approximately) the second half of the cases among $\mathcal{I}$ are made to have their x_2-value set to be missing. For example, suppose $\delta = 15\%$ and sample size is $n = 50$, then $\boldsymbol{x}_1(\delta)$ value corresponds to the seventh value in the ordered set of x_1-values. Suppose next that the cases that have their x_1-values less than $\boldsymbol{x}_1(\delta)$ are $\mathcal{I} = \{23, 9, 15, 33, 47, 3, 22\}$. Lastly, split set $\mathcal{I}$ into approximate halves as $\mathcal{I}_Y = \{23, 9, 15\}$ and $\mathcal{I}_{x_2} = \{33, 47, 3, 22\}$ such that Y_i is missing for $i \in \mathcal{I}_Y$ and x_{2j} is missing for $j \in \mathcal{I}_{x_2}$. This pattern of MAR is drastic since cases are lost (due to missingness) from the lower portion of the data when the data are ordered by $\boldsymbol{x}_1$.

For the data imputation, we used the `mice` package in `R`, where it is assumed that data are MAR and the imputation of any variable with missingness uses the predictive mean

matching (pmm) methods for imputation. The fully conditional specification for imputation of variables $\boldsymbol{x}_2$ and $\boldsymbol{y}$ is given by predictor matrices in Tables 1a,b for the `mice()` function for simulation settings where $k = 3$ and $k = 5$, respectively. These predictor matrices are indicator matrices (1=yes and 0=no) which inform the `mice()` function about which column variables will be used in the imputation model for each row variable.

Table 1: Predictor matrix for fully conditional specification (FCS) of the imputation model for the mice function for simulation configurations where (a) $k = 3$, and (b) $k = 5$.

(a)

	y	$x1$	$x2$	$x3$
$x2$	1	1	0	1
y	0	1	1	1
$x1$	0	0	0	0
$x3$	0	0	0	0

(b)

	y	$x1$	$x2$	$x3$	$x4$	$x5$
$x2$	1	1	0	1	1	1
y	0	1	1	1	1	1
$x1$	0	0	0	0	0	0
$x3$	0	0	0	0	0	0
$x4$	0	0	0	0	0	0
$x5$	0	0	0	0	0	0

This simulation study considered a total of 40 data configurations resulting from the combination of $2 \times 2 \times 2 \times 5$ factors. For each configuration, seven different methods were used to perform the likelihood ratio (LR) test for the hypothesis in equation (6.18). These methods included CCA with incomplete data and six imputation-based approaches: four from our proposed method utilizing transformation functions, as well as MR's method (referred to as MR92) and CM's modified-MR method with r_L^+ (referred to as CM.Pos). During each LR test, decisions to reject or not reject the null hypothesis were recorded for all seven methods. The entire process of data generation, inducing missingness, imputing data, and LR testing was repeated 1000 times.

Statistica Sinica: Newly accepted Paper

6.1 Simulation Study Results

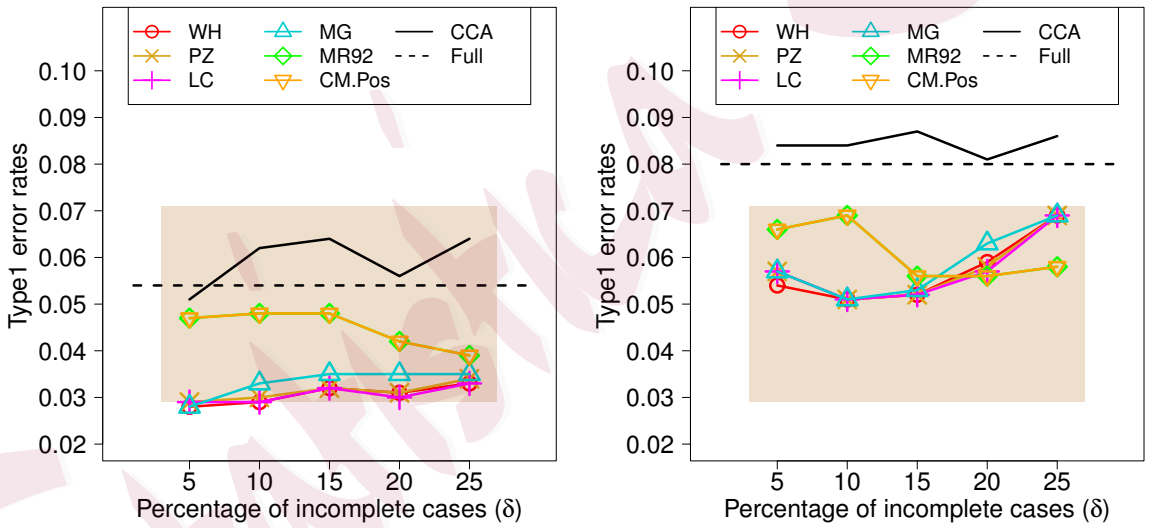
Since all tests were conducted at the pre-specified significance level of $\alpha = 0.05$, for simulation settings under the null hypothesis, Figures 3 and 4 include a shaded rectangle to represent the region of Type-I-error indifference. This was computed using confidence intervals for proportion α with the number of simulations $N = 1000$ to yield $\alpha \mp 3\sqrt{\frac{\alpha(1-\alpha)}{N}} \approx (0.029, 0.071)$. Additionally, subsequent figures present error and power rates for the original *full* data (with no missing values), labelled as "Full" in figure legends.

Type I Error Rates

In Figure 3a for $n = 50$ and $k = 3$, all methods had their observed Type-I error rates remain within the upper threshold of 7.1%, with results for the proposed method (using transformations WH, PZ, LC, and MG) being lower than the Type-I error rates of CCA, MR92, and CM.Pos. In comparison, in Figure 3b for $(n, k) = (50, 5)$, the proposed method had its Type-I error rates within the Type-I error indifference region, but its rates were higher than the Type-I error rates of MR92 and CM.Pos for $\delta > 15\%$. This setting of $(n, k) = (50, 5)$ is of particular interest since it represents a setting where the parameter dimension $k = 5$ is considered "comparable" to the sample size $n = 50$ in the sense of these values being closer when compared with the setting $(n, k) = (50, 3)$. For $(n, k) = (50, 5)$, the fact that the Type-I error rate for "Full" is greater than the upper threshold of the α-range again suggests that, in our simulated setting, the sample size $n = 50$ is not sufficiently larger than $k = 5$ for the asymptotic distribution of the Likelihood Ratio Test (LRT) statistic

Statistica Sinica: Newly accepted Paper

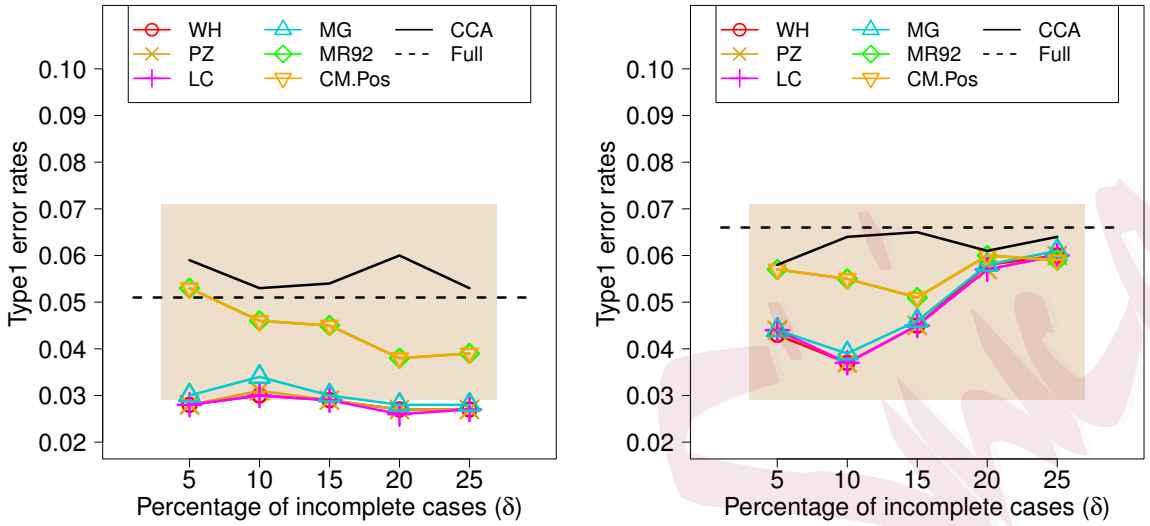
to be well-approximated by a chi-squared distribution. This also explains the larger than expected Type-I error values for CCA, where the sample size used in the analysis is smaller than $n = 50$ because this analysis is limited to complete cases only. For $n = 100$, Figures 4a-b show reduced variability compared to $n = 50$, which is expected. The conclusions from $n = 50$ were also observed in settings with $n = 100$ with the main difference being that for $(n, k) = (100, 5)$, CCA no longer has inflated Type-I error rates. Overall, conclusions from the Type-I error rates in Figures 3 and 4 suggest that CCA does not qualify as an α-level test, while other methods do qualify and are hence comparable in terms of their power.



(a) (b)

Figure 3: Comparing observed Type1 error rates of LRT using CCA, WH, PZ, MG, LC, MR and CM for $n = 50$ and (a) $k = 3$ and (b) $k = 5$.

Statistica Sinica: Newly accepted Paper



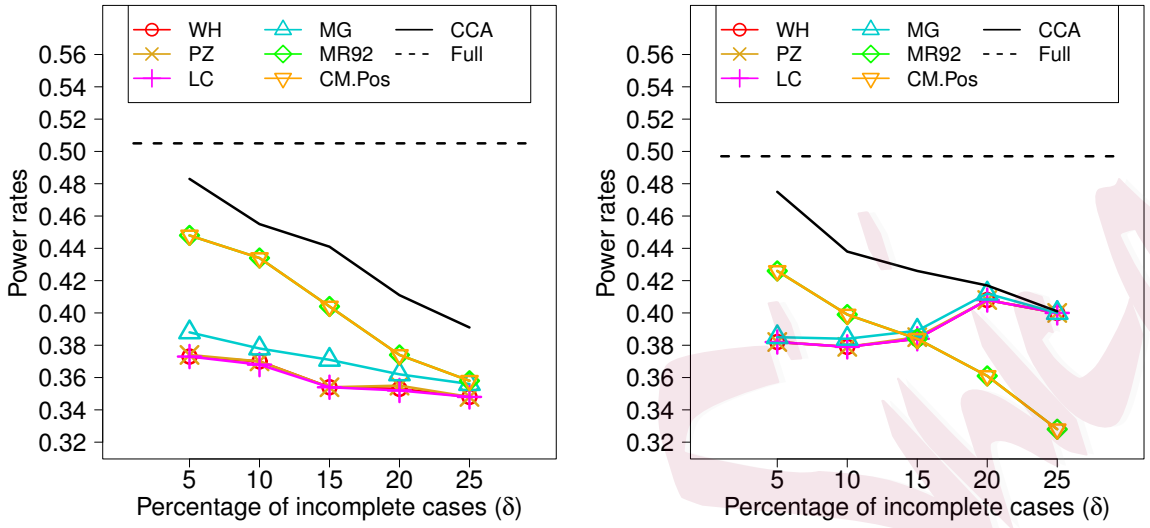
(a) (b)

Figure 4: Comparing observed Type1 error rates of LRT using CCA, WH, PZ, MG, LC, MR and CM for $n = 100$ and (a) $k = 3$ and (b) $k = 5$.

Power rates

Figures 5a-b and 6a-b compare power under the alternative hypothesis of non-zero association. For $(n, k) = (50, 3)$, in Figure 5a, CCA and the two competitor methods, MR92 and CM.Pos, have higher power at lower missingness levels. However, their performance exhibits steep declines as δ increases. On the other hand, the performance of our proposed method (using transformations WH, PZ, LC, and MG) remains stable for all rates of missingness, thereby exhibiting controlled Type-II error rates because of minimal variability across different missing percentages. Figure 5b shows a similar trend for $k = 5$, but power values are slightly lower overall due to the increased parameter dimension. An interesting observation, however, is that the power of proposed method is higher than that of competitors MR92 and

Statistica Sinica: Newly accepted Paper

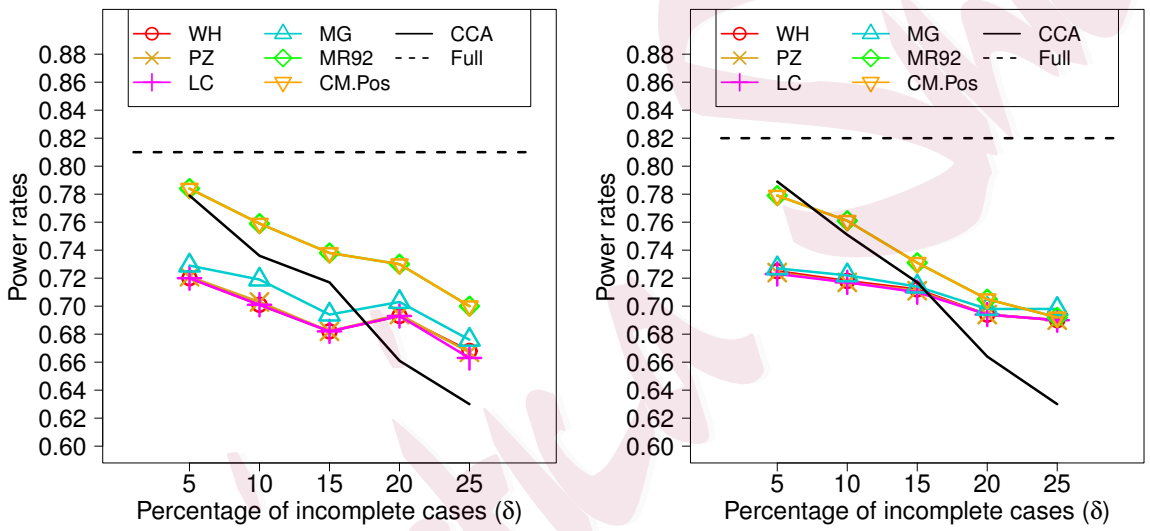


(a) (b)

Figure 5: Comparing observed Power rates of LRT using CCA, WH, PZ, MG, LC, MR and CM for $n = 50$ and (a) $k = 3$ and (b) $k = 5$.

CM.Pos for $\delta \geq 15\%$. For $n = 100$ in Figures 6a-b, both panels indicate improved power for all methods, with our proposed methods still showing a more stable pattern compared to other methods.

Statistica Sinica: Newly accepted Paper



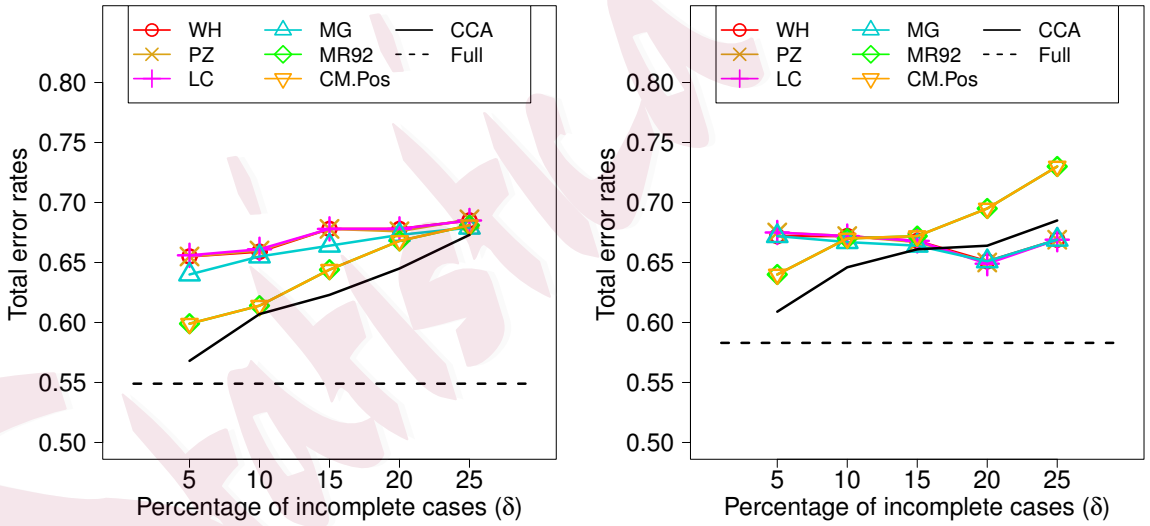
(a) (b)

Figure 6: Comparing observed Power rates of LRT using CCA, WH, PZ, MG, LC, MR and CM for $n = 100$ and (a) $k = 3$ and (b) $k = 5$.

Statistica Sinica: Newly accepted Paper

Total Error Rates

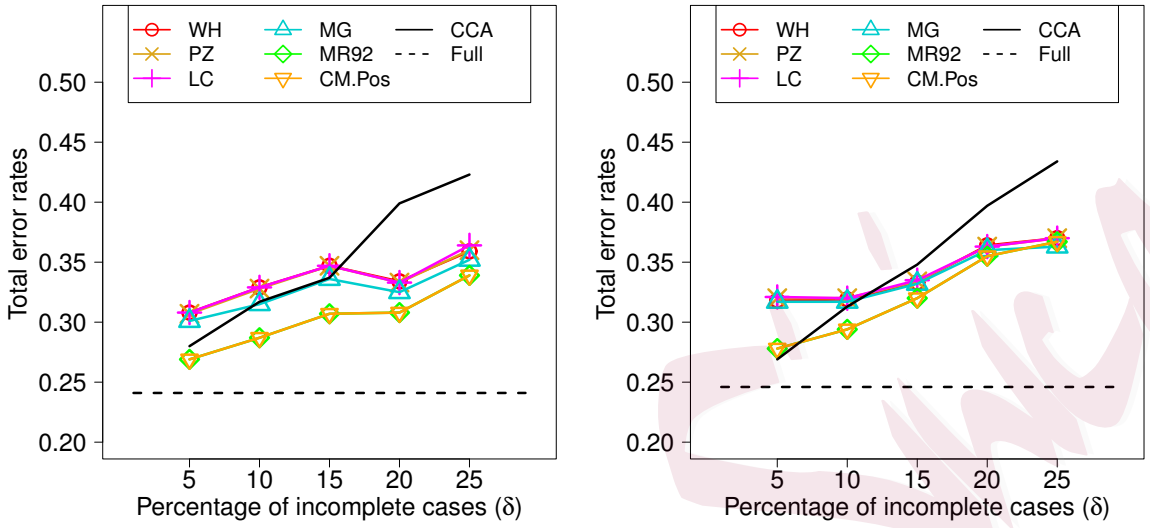
Figure 7a-b summarizes total error rates for $n = 50$. From Figure 7a for $(n, k) = (50, 3)$, it is observed that the total error increases with increasing δ for all methods, with the rate of increase being higher for CCA and the competing methods MR92 and CM.Pos compared with that of our proposed method. For $k = 5$ in Figure 7b, a similar pattern can be seen for CCA, but the total errors for our proposed method remain almost unchanged. Moreover, MR92 and CM.Pos consistently show higher total error for $\delta \geq 15\%$, reflecting their inefficiency as the percentage of missingness increases and the parameter dimension k approaches n.



(a) (b)

Figure 7: Comparing observed total error rates of LRT using CCA, WH, PZ, MG, LC, MR and CM for $n = 50$ and (a) $k = 3$ and (b) $k = 5$.

Statistica Sinica: Newly accepted Paper



(a) (b)

Figure 8: Comparing observed total error rates of LRT using CCA, WH, PZ, MG, LC, MR and CM for $n = 100$ and (a) $k = 3$ and (b) $k = 5$.

Key Observations

Let the term *small-to-moderate sample size* denote settings where $k < n$ but $k \not\ll n$. From the simulation study, we make the following overall observations. The proposed method effectively controls the Type I error rate and maintains stable power, resulting in stable total error rates across varying levels of missingness. However, in settings with small-to-moderate sample sizes and varying levels of missingness, CCA yields unreliable results, whereas our proposed method outperforms competing approaches — higher power and lower total error rates — when the missingness rate exceeds 15%.

The close alignment of the observed Type-I error rates and power values indicates that our proposed method utilizing transformations achieves inferential equivalence to the tra-

Statistica Sinica: Newly accepted Paper

ditional multivariate methods. Unlike the MR92 and CM.Pos procedures, our proposed method demonstrates this inferential reliability without the associated computational burden, estimation bias, model assumptions, or mathematical inconsistencies.

Recommendations and a Cautionary Note about CCA

In data applications with moderate to high rates of missingness and moderately large k, we recommend our proposed method using the WH, PZ, LC, or MG transformations, with preference given to MG for its simplicity and low total (Type-I + Type-II) error values. Additionally, the steep decline in the power values of MR92 and CM.Pos as missingness rates increase should give one pause about applying them. Lastly, our results demonstrate that one should be cautious about concluding that CCA is a reliable method based solely on its power function. The CCA method yields higher power values at the cost of inflated Type-I error values, and this is an artefact of wider confidence intervals caused by the reduced sample size from excluding incomplete cases. Purporting CCA as a reliable method on these grounds would imply that a trivial method with an infinite confidence interval $(-\infty, +\infty)$ is universally superior, which is a clear contradiction.

6.2 Comparison of Computation Times

Since our proposed method performs calculations with one-dimensional entities, certain efficiencies are expected in computational time when compared with the steps involved in the competing methods MR92 and CM.Pos. To demonstrate this computational efficiency, we recorded the computation times of MR92 and our proposed method using Goria's transfor-

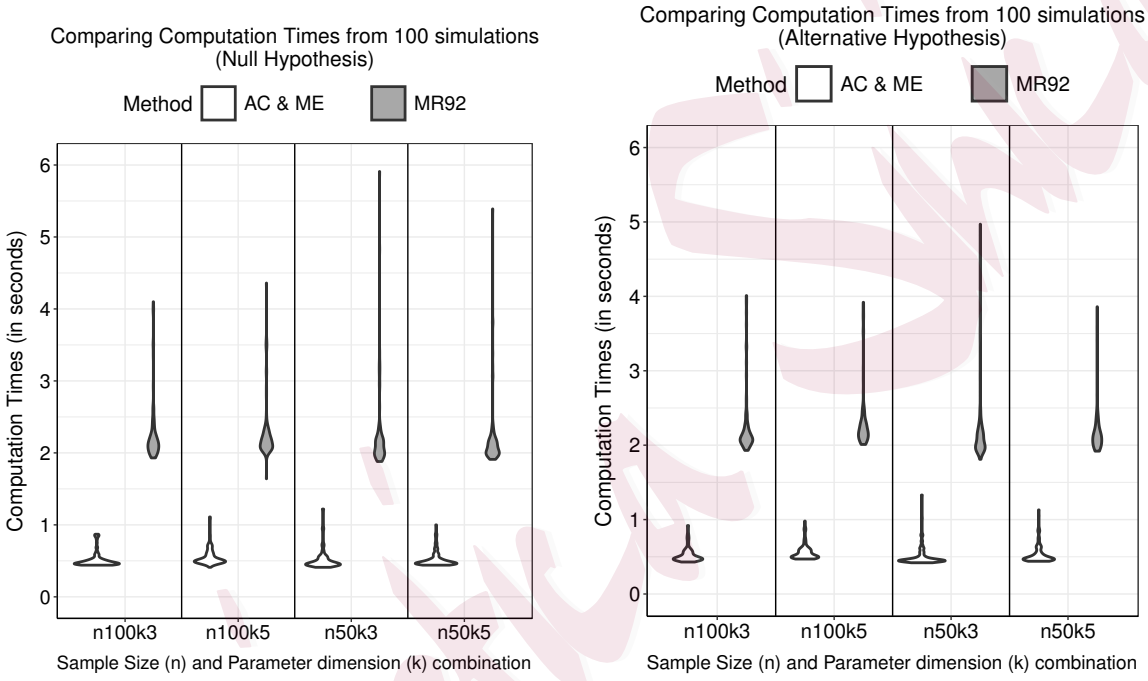
Statistica Sinica: Newly accepted Paper

mation for configurations of $(n, k, \delta = 25\%)$ under the null and alternative hypotheses, which are presented in Figure 9. These computation times (in seconds) are attributed purely to these individual methods after they are subjected to the same imputation object from using the `mice()` function on the incomplete data.

For our proposed method, we selected Goria's transformation since it showed superior results in Section 6.1, while MR92 was chosen since it is the competitive existing method that is computationally intensive. Since CM.Pos can be computed using elements from MR92, it is expected to have similar computation times and hence it was not considered for comparison in Figure 9.

For the null and alternative hypotheses, computation times for our proposed method across different combinations of (n, k) are consistently around 0.5 seconds (and mostly below 1 second) and approximately 20-25% lower than the computation time of MR92, which is around 2 seconds (and mostly below 6 seconds). In summary, these results suggest that computation times of our proposed method remains approximately constant as k increases from 3 to 5, consistent with $O(1)$ complexity.

Statistica Sinica: Newly accepted Paper



(a) (b)

Figure 9: Violin plots illustrating the distribution of computation times (obtained from 100 simulations) for the proposed AC & ME method (utilizing Goria's transformation) versus the MR92 method. Panels display results for combinations of $(n, k) \in \{(100, 3), (100, 5), (50, 3), (50, 5)\}$ with $\delta = 25\%$ under both the null and alternative hypotheses.

Statistica Sinica: Newly accepted Paper

7. Areas of application

Chi-square tests play a foundational role in various statistical models that support a broad range of applied disciplines. Our proposed method for combining chi-square statistics from multiply-imputed datasets addresses critical gaps in the existing literature by offering a computationally efficient and mathematically consistent approach that is superior in data with small-to-moderate sample sizes and substantial missingness. Next, we list a few examples from a broad spectrum of applications that the proposed method now makes available for analysis with multiply-imputed data:

Next, we list a few examples from a broad spectrum of applications that the proposed method now makes available for data analysis: **(i) Genetics:** the Chi-Square Goodness-of-Fit test can examine Mendelian inheritance patterns by comparing observed genetic crosses with expected ratios; **(ii) Marketing:** Pearson's chi-square test for independence can analyze contingency tables to assess consumer preferences; **(iii) Sociology:** McNemar's test can evaluate changes in public opinion over time using paired categorical data; and **(iv) Medical Research:** global and local tests within General Linear Models can assess the significance of multiple predictors in epidemiological studies.

Existing methods for combining chi-square statistics from multiply-imputed data are computationally demanding, mathematically incongruent, or reliant on restrictive assumptions, thereby limiting their utility for complex models or large-scale studies. Our proposed method addresses these challenges by offering a computationally efficient, reliable and eas-

ily implementable solution that applies to any setting where the test statistic is (approximately) chi-square distributed. By overcoming the theoretical issues and computational burdens associated with existing methods, our methodology advances the practical application of MI across various applied disciplines. A data application of our proposed method is demonstrated next with a case study, where the objective is to evaluate the conditional independence of infant survival and level of prenatal care, controlling for clinic via the use of Mantel-Haenszel test.

8. Application of the Cochran-Mantel-Haenszel Test to Infant Mortality Data

In this section, we analyze the incomplete infant mortality data from Little and Rubin (2002, Table 13.8, p. 284), where the clinic variable is missing for 26.3% (255 out of 970) of the cases. These data are presented in Table 2 as a $2 \times 2 \times 2$ contingency table for the i-th level of *prenatal care* (More or Less), j-th *survival* status (Died or Survived), and k-th *clinic* (A or B). The complete data in Table 2a is originally from Bishop (1969), while the data in Table 2b are the simulated cases with the artificially missing clinic by Little and Rubin (2002, Table 13.8, p. 284).

The dataset in Table 2 was previously utilized by Meng and Rubin (1992) and Chan (2022) to evaluate methods for combining LR test statistics from multiply-imputed data. In our analysis, we apply the Mantel-Haenszel test to assess the conditional independence between prenatal care and infant survival, stratified by clinic.

For the data presented in Table 2, let π_{ijk} denote the probability of classification for

Table 2: Partially classified 2^3 contingency table relating infant survival to the amount of prenatal care received across two clinics.

(a) Completely classified cases

		Survival, j		
Clinic, k	Prenatal care, i	Died	Survived	
A	Less	3	176	
	More	4	293	
B	Less	17	197	
	More	2	23	$N^{(a)} = 715$ cases

(b) [Artificial] Partially classified cases, clinic missing

		Survival, j		
Clinic, k	Prenatal care, i	Died	Survived	
?	Less	10	150	
	More	5	90	$N^{(b)} = 255$ cases

? denotes missing value.

cell (i, j, k). The observed cell counts, n_{ijk}, serve as the response variable. To test for conditional independence of prenatal care and survival given the clinic, we use the Cochran-Mantel-Haenszel (CMH) test statistic, defined as:

$$M^2 = \frac{\left(\sum_{k=1}^{K}(n_{11k} - \mu_{11k})\right)^2}{\sum_{k=1}^{K}\text{Var}(n_{11k})}$$

where $\mu_{11k} = \dfrac{n_{1+k} \cdot n_{+1k}}{n_{++k}}$ and $\text{Var}(n_{11k}) = \dfrac{n_{1+k} \cdot n_{2+k} \cdot n_{+1k} \cdot n_{+2k}}{n_{++k}^2(n_{++k} - 1)}$ with

n_{11k} as the observed count in the cell (1,1) of stratum k;

n_{1+k} and n_{2+k} as row totals for stratum k;

n_{+1k} and n_{+2k} as column totals for stratum k; and

n_{++k} as the total number of observations in stratum k.

Under the null hypothesis of conditional independence, M^2 is asymptotically chi-square distributed with one degree of freedom.

Assuming that missingness in the clinic variable are MAR, it was multiply-imputed using three different imputation algorithms. Specifically, we implemented imputation methods using (i) a fully Bayesian approach with a Beta prior following Meng and Rubin (1992), (ii) a logistic regression model, and (iii) the semi-parametric approach of predictive mean matching (pmm; Van Buuren, 2012; Van-Buuren and Oudshoorn, 1999). These imputation methods generate plausible values for missing values in the clinic variable, reflecting a range of modeling assumptions and imputation strategies. We next specify details for each imputation method.

For the fully Bayesian approach with a Beta prior, we followed the steps as outlined in Meng and Rubin (1992, Section 4.2, p.110). To impute the missing clinic for the l-th imputed dataset ($l = 1, 2, ..., m$), we assume a noninformative Beta$(1, 1)$ prior for the cell probabilities π_{ijk}. First, we draw the conditional cell probabilities, denoted as $\pi_{1|ij}$, from a Beta$(n^{(a)}_{ij1} + 1, n^{(a)}_{ij2} + 1)$ distribution, where $n^{(a)}_{ijk}$ denotes the observed cell count in cell (i, j, k) of Table 2(a). Using generated values for $\pi_{1|ij}$, we impute the counts $n^{(b)}_{ij1}$ for the marginal cells by drawing from a Binomial$(n^{(b)}_{ij+}, \pi_{1|ij})$ distribution, where $n^{(b)}_{ij+}$ is the observed marginal count for cell (i, j) of Table 2(b). The remaining count is set as $n^{(b)}_{ij2} = n^{(b)}_{ij+} - n^{(b)}_{ij1}$. Finally, the l-th imputed cell count for cell (i, j, k) is obtained by combining the observed and imputed values: $n^{(l)}_{ijk} = n^{(a)}_{ijk} + n^{(b)}_{ijk}$, for $(i, j, k = 1, 2)$.

For the other two imputation methods, we used the `mice()` function in `R` for ease of implementation, conditioning the imputation on the main effects of infant survival and prenatal care, first under a logistic regression model and then under the pmm framework. The pmm approach is particularly advantageous when parametric assumptions may not hold, or when it is important to preserve the empirical structure and bounds observed in the known data. For each imputation method, we used the following range for number of imputations: $m = \{5, 10, 20, 30, 50, 100, 500\}$. We chose this wide range because there is no clear consensus in the literature regarding the optimal number of imputations, and this remains an active area of research.

Table 3 presents the results of our proposed transformation method for the Cochran-Mantel-Haenszel chi-square test of association between variables Survival and Prenatal care controlling for Clinic. We calculated the CMH statistic for each imputed dataset, applied the transformation methods (from Section 3) to each statistic, and then combined the results using our proposed new combining rule. The PZ transformation is excluded from these results, as it requires $\nu > 1$, whereas in this example $\nu = 1$. In Table 3, for our proposed method (using transformations WH, MG, and LC), t_{stat}, v_{BR}, and p represents the calculated t-statistic, the degrees of freedom based on Barnard and Rubin (1999) method, and the corresponding p-value for different number of imputations denoted by m. Also included as a footnote in Table 3 are results from CCA of data in Table 2a and all observed data (in Table 2). Statistics for "Observed" are obtained from CMH test of the data where counts from Table 2b are distributed into counts of Table 2a using the EM algorithm under the

assumption that missingness is ignorable.

Table 3 shows that the proposed transformation-based methods led to same conclusions as the Observed and CCA methods, but tend to produce lower p-values. This is expected, as our simulation study demonstrated that the CCA method exhibits inflated type-I error rates under the null hypothesis, resulting in higher p-values. Based on the analysis of this data, we do not find evidence for association between infant survival and prenatal care after controlling for clinic.

Table 3: MI results for CMH test when using different imputation methods for infant mortality data.

Beta prior (Meng and Rubin, 1992)									
	WH			MG			LC		
m	t_{stat}	v_{BR}	p	t_{stat}	v_{BR}	p	t_{stat}	v_{BR}	p
5	-0.555	50.023	0.581	-0.597	35.423	0.554	-0.595	31.396	0.556
10	-0.785	194.103	0.433	-0.857	137.816	0.393	-0.858	118.997	0.393
20	-0.738	342.124	0.461	-0.805	260.113	0.421	-0.804	236.743	0.422
30	-0.625	303.211	0.532	-0.679	230.882	0.497	-0.687	204.095	0.493
50	-0.665	430.019	0.506	-0.724	339.177	0.469	-0.731	290.126	0.466
100	-0.679	537.352	0.498	-0.739	450.309	0.460	-0.748	403.761	0.455
500	-0.680	697.077	0.496	-0.741	631.819	0.459	-0.751	588.827	0.453
Logistic prior									
	WH			MG			LC		
m	t_{stat}	v_{BR}	p	t_{stat}	v_{BR}	p	t_{stat}	v_{BR}	p
5	-0.743	27.843	0.464	-0.802	18.356	0.433	-0.818	11.600	0.430
10	-0.898	144.427	0.371	-0.983	92.146	0.328	-0.993	53.900	0.325
20	-0.740	185.348	0.460	-0.806	131.169	0.421	-0.827	97.160	0.410
30	-0.732	268.126	0.465	-0.798	197.776	0.426	-0.816	155.323	0.416
50	-0.737	393.901	0.462	-0.804	301.738	0.422	-0.819	236.686	0.413
100	-0.679	518.008	0.497	-0.740	425.327	0.460	-0.753	360.452	0.452
500	-0.710	713.768	0.478	-0.774	649.395	0.439	-0.782	603.412	0.434
Predictive Mean Matching (pmm)									
	WH			MG			LC		
m	t_{stat}	v_{BR}	p	t_{stat}	v_{BR}	p	t_{stat}	v_{BR}	p
5	-0.125	20.741	0.902	-0.144	14.982	0.887	-0.193	11.743	0.850
10	-0.395	82.370	0.694	-0.427	59.692	0.671	-0.440	50.711	0.662
20	-0.384	160.102	0.702	-0.415	121.735	0.679	-0.429	113.867	0.668
30	-0.446	219.543	0.656	-0.483	169.347	0.630	-0.497	155.943	0.620
50	-0.493	300.224	0.623	-0.534	239.051	0.594	-0.550	223.798	0.583
100	-0.537	412.934	0.591	-0.583	340.190	0.560	-0.603	311.025	0.547
500	-0.406	586.609	0.685	-0.441	524.218	0.660	-0.461	511.305	0.645

Observed: $\chi^2 = 0.014$, $df = 1$, $p = 0.906$
CCA: $\chi^2 = 0.039$, $df = 1$, $p = 0.844$

Statistica Sinica: Newly accepted Paper

9. Discussion and Conclusion

This article introduces a novel, simplified method utilizing transformation functions to combine chi-square statistics from analysis of multiply-imputed data. Unlike existing methods by Chan and Meng (2022) and Meng and Rubin (1992), which suffer from mathematical inconsistencies (where the posited strictly positive statistic yields negative values), require unverifiable assumptions (such as *proportional and equal loss of information*), or are computationally burdensome (due to calculations involving matrices and access to imputed data for multiple passes of multivariate statistics), our proposed method circumvents these issues. Specifically, our approach involves transforming chi-square statistics to approximately normal, followed by pooling them using univariate combining rules. This approach eliminates the requirement in existing methods for access to the multiply-imputed data and multiple passes of statistics. Thus, our method allows researchers to implement standard univariate combining rules directly without any of the computational burden and limitations of existing methods.

Extensive simulation studies demonstrated that the proposed method delivers superior performance by effectively controlling Type I error rates and providing superior statistical power than competing methods in data settings with small-to-moderate sample size and missingness exceeding 15%. Furthermore, these statistical advantages are achieved alongside substantial computational efficiency such that our proposed method (on average) was four times faster with computational complexity of order $O(1)$, thereby presenting a significant

advantage over the higher computational costs of methods compared here.

While our method offers substantial practical and theoretical advantages for combining chi-square statistics from multiply-imputed data, future work includes modifying formulations and reference distributions of existing methods to improve their statistical power. Limitations in these components are likely the primary source of severe loss in statistical power in settings with small-to-moderate sample size and substantial missingness.

10. Acknowledgements

The authors thank the Co-Editor, the Associate Editor, and the two anonymous referees for seeing the value of our work in the initial submission and for their insightful comments that helped improve the final manuscript. This research was supported by Discovery Grant from the Natural Science and Engineering Research Council of Canada to Ashok Chaurasia [RGPIN-2018-05479]. The funding agency had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The opinions expressed in this manuscript are those of the authors and do not necessarily reflect the views of their respective institutions.

References

Barnard, J. and D. B. Rubin (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika 86*, 948–955.

Bishop, Y. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics*, 383–399.

Canal, L. (2005). A normal approximation for the chi-square distribution. *Computational statistics & data analysis 48*(4), 803–808.

Chan, K. W. (2022). General and feasible tests with multiply-imputed datasets. *The Annals of Statistics 50*(2), 930–948.

Statistica Sinica: Newly accepted Paper

Chan, K. W. and X.-L. Meng (2022). Multiple improvements of multiple imputation likelihood ratio tests. *Statistica Sinica 32*(3), 1489–1514.

Chaurasia, A. (2023). Combining rules for F- and Beta-statistics from multiply-imputed data. *Econometrics and Statistics 25*, 51–65.

Goria, M. N. (1992). On the fourth root transformation of chi-square. *Australian Journal of Statistics 34*(1), 55–64.

Grund, S., O. Lüdtke, and A. Robitzsch (2016). Pooling anova results from multiply imputed datasets. *Methodology*.

Li, K. H., T. E. Raghunathan, and D. B. Rubin (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. *Journal of the American Statistical Association 86*, 1065–1073.

Little, R. and D. Rubin (2002). *Statistical analysis with missing data* (2 ed.). New York: John Wiley & Sons.

Meng, X.-L. and D. B. Rubin (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika 79*(1), 103–111.

Peizer, D. B. and J. W. Pratt (1968). A normal approximation for binomial, F, Beta, and other common, related tail probabilities, I. *Journal of the American Statistical Association 63*(324), 1416–1456.

Raghunathan, T. and Q. Dong (2011). *Analysis of Variance from Multiply Imputed Data Sets*. Ann Arbor: University of Michigan.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika 94*(2), 502–508.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Van Buuren, S. (2012). *Flexible imputation of missing data*, Volume 10. CRC press Boca Raton, FL.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software 45*, 1–67.

Van-Buuren, S. and C. Oudshoorn (1999). *Flexible multivariate imputation by mice*. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

van Ginkel, J. R. (2019). Significance tests and estimates for R^2 for multiple regression in multiply imputed datasets: A cautionary note on earlier findings, and alternative solutions. *Multivariate Behavioral Research 54*(4), 514–529.

Wilson, E. B. and M. M. Hilferty (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences 17*(12), 684–688.

Ashok Chaurasia
School of Public Health Sciences,
University of Waterloo,
Waterloo, Ontario N2L 3G1, Canada.
E-mail: a4chaura[at]uwaterloo[dot]ca
ORCID: https://orcid.org/0000-0001-6745-0493

Mohsen Ebadi
School of Public Health Sciences,
University of Waterloo,
Waterloo, Ontario N2L 3G1, Canada.
Note: This author contributed to the work under the supervision of Ashok Chaurasia.
E-mail: m2ebadi[at]uwaterloo[dot]ca
ORCID: https://orcid.org/0000-0002-6793-9683

Statistica Sinica