The Annals of Statistics
2022, Vol. 50, No. 2, 930–948
https://doi.org/10.1214/21-AOS2132

GENERAL AND FEASIBLE TESTS WITH MULTIPLY-IMPUTED DATASETS

BY KIN WAI CHAN[a]

Department of Statistics, The Chinese University of Hong Kong, [a]*kinwaichan@cuhk.edu.hk*

Multiple imputation (MI) is a technique especially designed for handling missing data in public-use datasets. It allows analysts to perform incomplete-data inference straightforwardly by using several already imputed datasets released by the dataset owners. However, the existing MI tests require either a restrictive assumption on the missing-data mechanism, known as equal odds of missing information (EOMI), or an infinite number of imputations. Some of them also require analysts to have access to restrictive or nonstandard computer subroutines. Besides, the existing MI testing procedures cover only Wald's tests and likelihood ratio tests but not Rao's score tests, therefore, these MI testing procedures are not general enough. In addition, the MI Wald's tests and MI likelihood ratio tests are not procedurally identical, so analysts need to resort to distinct algorithms for implementation. In this paper, we propose a general MI procedure, called stacked multiple imputation (SMI), for performing Wald's tests, likelihood ratio tests and Rao's score tests by a unified algorithm. SMI requires neither EOMI nor an infinite number of imputations. It is particularly feasible for analysts as they just need to use a complete-data testing device for performing the corresponding incomplete-data test.

1. New thoughts on the old results.

1.1. *Introduction*. Missing data are usually encountered in real-data analysis, both in observational and experimental studies. Statistical inference of incomplete datasets is harder than that of complete datasets. Multiple imputation (MI), proposed by Rubin (1978), is one of the most popular ways of handling missing data. This method requires specifying an imputation model for filling in the missing data multiple times so that standard complete-data procedures can be straightforwardly applied to each of the imputed datasets; see Sections 1.2 and 1.3 for a review. Also see Rubin (1987), Carpenter and Kenward (2013), and Kim and Shao (2013) for an introduction. Although a fairly complete theory for performing MI tests is available, all existing results suffer from at least one of following three problems: (i) reliance on strong statistical assumptions, (ii) requirement of infeasible computer subroutines, and (iii) lack of unified combining rules for various types of tests. We will discuss these problems thoroughly in Section 1.4.

The major goal of this paper is to derive a handy MI test that resolves the aforementioned problems. This paper is structured as follows. In the remaining part of this section, the existing MI tests are reviewed. We also discuss their pros and cons. In Section 2, we motivate the proposal test statistics and present the plan to achieve our proposed test. In Section 3, our proposed methodology and principle are discussed. In Section 4, a novel theory for estimating the odds of missing information is presented. In Section 5, a new test based on multiply-imputed datasets is derived. In Section 6, applications and simulation experiments are illustrated. In Section 7, we conclude the paper and discuss possible future work. Proofs, auxiliary results, additional simulation results and an R-package `stackedMI` are included as Supplementary Material (Chan (2021)).

Received February 2021; revised August 2021.
MSC2020 subject classifications. Primary 62D05; secondary 62F03, 62E20.
Key words and phrases. Fraction of missing information, hypothesis testing, jackknife, missing data, stacking.

1.2. *Background and problem setup.* Let X be a dataset in the form of a $n \times p$ matrix consisting of n rows of independent units. Assume that X is generated from a probability density $f(X \mid \psi)$, where $\psi \in \Psi$ is the model parameter. The parameter of interest is a sub-vector of ψ denoted by $\theta \in \Theta \subseteq \mathbb{R}^k$. We are interested in testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ for some fixed $\theta_0 \in \Theta$. If the complete dataset $X_{\mathrm{com}} := X$ is available, we may perform the Wald's test, likelihood ratio (LR) test, and Rao's score (RS) test. The test statistics are

$$d_{\mathrm{W}}(X) := d_{\mathrm{W}}(\widehat{\theta}, \widehat{V}) := (\widehat{\theta} - \theta_0)^{\mathrm{T}} \widehat{V}^{-1} (\widehat{\theta} - \theta_0), \tag{1.1}$$

$$d_{\mathrm{L}}(X) := d_{\mathrm{L}}(\widehat{\psi}, \widehat{\psi}_0 \mid X) := 2 \log\{ f(X \mid \widehat{\psi}) / f(X \mid \widehat{\psi}_0) \}, \tag{1.2}$$

$$d_{\mathrm{R}}(X) := d_{\mathrm{R}}(\widehat{\psi}_0 \mid X) := u(\widehat{\psi}_0)^{\mathrm{T}} \{ I(\widehat{\psi}_0) \}^{-1} u(\widehat{\psi}_0), \tag{1.3}$$

respectively, where $\widehat{\theta} := \widehat{\theta}(X)$ is the maximum likelihood estimator (MLE) of θ; $\widehat{V} := \widehat{V}(X)$ is a variance estimator of $\widehat{\theta}$; $\widehat{\psi} := \widehat{\psi}(X)$ and $\widehat{\psi}_0 := \widehat{\psi}_0(X)$ are the unrestricted and H_0-restricted MLEs of ψ; $u(\psi) := u(\psi \mid X) := \partial \log f(X \mid \psi) / \partial \psi$ is the score function; and $I(\psi) := I(\psi \mid X) := -\partial^2 \log f(X \mid \psi) / \partial \psi \partial \psi^{\mathrm{T}}$ is the Fisher's information. See Lehmann and Romano (2005) for more details. Throughout the paper, we use $\aleph \in \{\mathrm{W}, \mathrm{L}, \mathrm{R}\}$ to abbreviate the name of the test in various subscripts. In each of (1.1)–(1.3), the mapping $X \mapsto d_{\aleph}(X)$ is a function of the dataset X only. We call such function $X \mapsto d_{\aleph}(X)$ a standard testing *device* (i.e., a testing *subroutine* or a testing *procedure*) in computer software. The device $d_{\aleph}(\cdot)$ is the only requirement for complete-data testing. Under standard regularity conditions (see, e.g., Section 4.4 of Serfling (2001)) and under H_0, we have, for any $\aleph \in \{\mathrm{W}, \mathrm{L}, \mathrm{R}\}$, that

$$d_{\aleph}(X)/k \Rightarrow \chi^2_k / k, \quad \text{as } n \to \infty, \tag{1.4}$$

where "$\Rightarrow$" denotes weak convergence; see Chapter 2 of van der Vaart (2000).

If a part of $X_{\mathrm{com}} = \{X_{\mathrm{obs}}, X_{\mathrm{mis}}\}$ is missing such that only X_{obs} is available, testing H_0 is more involved. One widely used method is *multiple imputation* (MI), which is a two-stage procedure.

- The first stage involves an *imputer* to handle the missing data. The imputer draws $X^1_{\mathrm{mis}}, \ldots, X^m_{\mathrm{mis}}$ from the conditional distribution $[X_{\mathrm{mis}} \mid X_{\mathrm{obs}}]$ so that the missing part X_{mis} can be filled in by $X^1_{\mathrm{mis}}, \ldots, X^m_{\mathrm{mis}}$ to form m completed datasets $X^\ell := \{X_{\mathrm{obs}}, X^\ell_{\mathrm{mis}}\}$ ($\ell = 1, \ldots, m$). Note that MI assumes the missing mechanism is ignorable (Rubin (1976)). See Remark 1.1 for more discussion of this stage.
- The second stage involves possibly many *analysts* to perform inference of their own interests. Each of them receives the same completed datasets $X^1, \ldots, X^m$ from the imputer. Then, s/he can repeatedly apply some standard *complete-data procedures* to $X^1, \ldots, X^m$, and obtain m preliminary results. The final result is obtained by appropriately combining them; see Section 1.3 for a review.

MI is an attractive method because it naturally divides imputation and analysis tasks into two separate stages so that analysts do not need to be trained for handling incomplete datasets. Indeed, it has been a very popular method in various fields; see, for example, Rubin (1987) Tu, Meng and Pagano (1993), Rubin (1996), Schafer (1999), King et al. (2001), Peugh and Enders (2004), Kenward and Carpenter (2007), Harel and Zhou (2007), Horton and Kleinman (2007), Rose and Fraser (2008), Holan et al. (2010), Kim and Yang (2017), and Yu et al. (2021).

REMARK 1.1. MI is originally designed for handling public-use datasets. Hence, the imputers in stage 1 are in general different from the analysts in stage 2; see, for example,

Parker and Schenker (2007). Consequently, analysts cannot produce arbitrarily many imputed datasets as they wish. For example, only multiply-imputed datasets are released in National Health Interview Survey (NHIS) conducted in the United States; see Schenker et al. (2006). In particular, only five imputed datasets are released in 2018 NHIS; see https://www.cdc.gov/nchs/nhis/nhis_2018_data_release.htm. Moreover, because the imputers usually belong to the organizations, for example, a census bureau of the government, who collect the data, they usually know better how the data are missing, and have auxiliary variables to impute the missing data. So, the ignorability could be a reasonable assumption. See Rubin (1996) for more comprehensive discussions of these matters.

1.3. *MI combining rules and reference null distribution.* Applying the functions $\widehat{\theta}(\cdot)$, $\widehat{V}(\cdot)$, $\widehat{\psi}(\cdot)$ and $\widehat{\psi}_0(\cdot)$ to each of the imputed datasets $X^1, \ldots, X^m$, an analyst obtains $\widehat{\theta}^\ell := \widehat{\theta}(X^\ell)$, $\widehat{V}^\ell := \widehat{V}(X^\ell)$, $\widehat{\psi}^\ell := \widehat{\psi}(X^\ell)$ and $\widehat{\psi}_0^\ell := \widehat{\psi}_0(X^\ell)$ for $\ell = 1, \ldots, m$. Define

$$\widetilde{d}'_{\mathrm{W}} := \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{W}}(\widehat{\theta}^\ell, \bar{V}), \qquad \widetilde{d}''_{\mathrm{W}} := d_{\mathrm{W}}(\bar{\theta}, \bar{V}), \tag{1.5}$$

$$\widetilde{d}'_{\mathrm{L}} := \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{L}}(\widehat{\psi}^\ell, \widehat{\psi}_0^\ell \mid X^\ell), \qquad \widetilde{d}''_{\mathrm{L}} := \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{L}}(\bar{\psi}, \bar{\psi}_0 \mid X^\ell), \tag{1.6}$$

where $\bar{\theta} := \sum_{\ell=1}^m \widehat{\theta}^\ell/m$, and $\bar{V}$, $\bar{\psi}$, $\bar{\psi}_0$ are similarly defined. The MI Wald's statistic (Li, Raghunathan and Rubin (1991)) and MI LR statistic (Meng and Rubin (1992)) are $\widetilde{D}_{\mathrm{W}}$ and $\widetilde{D}_{\mathrm{L}}$, respectively, where, for $\aleph \in \{\mathrm{W}, \mathrm{L}\}$,

$$\widetilde{D}_{\aleph} := \frac{\widetilde{d}''_{\aleph}}{k\{1 + (1 + \frac{1}{m})\widetilde{\mu}_{r,\aleph}\}} \quad \text{and} \quad \widetilde{\mu}_{r,\aleph} := \frac{\widetilde{d}'_{\aleph} - \widetilde{d}''_{\aleph}}{k(m-1)/m}. \tag{1.7}$$

The factor $\{1 + (1 + 1/m)\widetilde{\mu}_{r,\aleph}\}$ in $\widetilde{D}_{\aleph}$ is used to deflate $\widetilde{d}''_{\aleph}$ in order to adjust for the loss of information due to missingness. The LR test statistic $\widetilde{D}_{\mathrm{L}}$ may be negative. Chan and Meng (2021+) recently proposed a corrected version.

If m is fixed and $n \to \infty$, the limiting distribution of $\widetilde{D}_{\aleph}$ is notoriously complicated because of the dependence on the unknown matrix $F := I_{\mathrm{mis}} I_{\mathrm{com}}^{-1}$ in a tangled way, where

$$I_{\mathrm{com}} := \mathsf{E}\left\{-\frac{\partial^2 \log f(X \mid \psi)}{\partial\theta\,\partial\theta^{\mathrm{T}}}\right\} \quad \text{and} \quad I_{\mathrm{mis}} := \mathsf{E}\left\{-\frac{\partial^2 \log f(X \mid X_{\mathrm{obs}}, \psi)}{\partial\theta\,\partial\theta^{\mathrm{T}}}\right\}$$

are the complete-data and missing-data Fisher's information of θ, respectively. We also denote $I_{\mathrm{obs}} := I_{\mathrm{com}} - I_{\mathrm{mis}}$ as the observed-data Fisher's information of θ. Let the eigenvalues of F be $f_1 \geq \cdots \geq f_k$, and denote $r_j := f_j/(1 - f_j)$ for each j. The values f_j and r_j are known as the *fraction of missing information* (FMI) and the *odds of missing information* (OMI), respectively. There is no loss in information if $r_1 = \cdots = r_k = 0$, or, equivalently, $f_1 = \cdots = f_k = 0$. Under regularity conditions (RCs) and H_0,

$$\widetilde{D}_{\aleph} \Rightarrow \mathbb{D} := \frac{\frac{1}{k}\sum_{j=1}^k \{1 + (1 + \frac{1}{m})r_j\}G_j}{1 + \frac{1}{k}\sum_{j=1}^k (1 + \frac{1}{m})r_j H_j}, \tag{1.8}$$

as $n \to \infty$, where $G_1, \ldots, G_k \sim \chi_1^2$ and $H_1, \ldots, H_k \sim \chi_{m-1}^2/(m-1)$ are independent. The RCs and derivation of (1.8) are presented in Proposition 5.1. Since the distribution $\mathbb{D}$ depends on the nuisance parameters $r_1, \ldots, r_k$ in a complicated way, it is not immediately possible to use $\mathbb{D}$ as a reference null distribution. In order to mitigate this fundamental difficulty, it has been a common practice to assume some structure on $r_1, \ldots, r_k$ (see Condition 1 below), and/or resort to asymptotic ($m \to \infty$) approximation. We present these existing strategies one by one, and give a summary in Table 1.

TABLE 1

Summary of the asymptotic reference null distributions. The reference distribution refers to the common weak limit ($n \to \infty$ and $m > 1$) of the existing test statistic $\widetilde{D} \in \{\widetilde{D}_{\mathrm{W}}, \widetilde{D}_{\mathrm{L}}\}$ defined in (1.7) and the proposed test statistics $\widehat{D} \in \{\widehat{D}_{\mathrm{W}}, \widehat{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{R}}\}$ to be defined in (2.3)

Assumptions		Asymptotic null distribution of $\widetilde{D}$ or $\widehat{D}$		
Condition 1	$m \to \infty$	Exact	Approximated	Name
Required	Required	χ^2_k/k	χ^2_k/k	T1
Required	Not required	$\mathbb{D}_0$ in (1.9)	$\approx F(k, \widehat{q})$ (Li, Raghunathan and Rubin (1991))	T2
Not required	Required	$\mathbb{D}_\infty$ in (1.10)	$\approx \widehat{c}_0 + \widehat{c}_1 \chi^2_k/k$ (Meng (1990))	T3
Not required	Not required	$\mathbb{D}$ in (1.8)	$\approx \widehat{\mathbb{D}}$ (Algorithm 2.1)	T4 (proposal)

CONDITION 1 (Equal odds of missing information (EOMI)). There is μ_r such that $r_1 = \cdots = r_k = \mu_r$.

Condition 1 is equivalent to $f_1 = \cdots = f_k$, known as equal fraction of missing information (EFMI). Although it is a strong assumption and is almost always violated in real problems, it is widely used in the literature because of simplicity; see Rubin (1987). Under Condition 1, $\mathbb{D}$ can be represented as

$$\mathbb{D}_0 := \frac{\{1 + (1 + \frac{1}{m})\mu_r\}G}{1 + (1 + \frac{1}{m})\mu_r H}, \tag{1.9}$$

where $G \sim \chi^2_k/k$ and $H \sim \chi^2_{k(m-1)}/\{k(m-1)\}$ are independent. In this case, $\mathbb{D}_0$ depends only on one unknown μ_r, hence, approximating (1.9) is easier. The first approximation of $\mathbb{D}_0$ was provided by Rubin (1987). Then Li, Raghunathan and Rubin (1991) refined it to $F(k, \widehat{q})$, where $\widehat{q}$ is an estimate of

$$q := \begin{cases} 4 + (K_m - 4)\left[1 + (1 - 2/K_m)\left\{\left(1 + \dfrac{1}{m}\right)\mu_r\right\}^{-1}\right]^2, & \text{if } K_m > 4; \\ (m-1)\left[1 + \left\{\left(1 + \dfrac{1}{m}\right)\mu_r\right\}^{-1}\right]^2 (k+1)/2, & \text{otherwise,} \end{cases}$$

where $K_m := k(m-1)$. In practice, $\widehat{q}$ is constructed by plugging in an estimate of μ_r into q. Li, Raghunathan and Rubin (1991) and Meng and Rubin (1992) proposed to estimate μ_r by $\widetilde{\mu}_{r,\mathrm{W}}$ and $\widetilde{\mu}_{r,\mathrm{L}}$ if Wald's test and the LR test are used, respectively. Many software routines have implemented this approximation, for example, van Buuren and Groothuis-Oudshoorn (2011). Besides, some approximations of $\mathbb{D}_0$ designed for small n can be found in Rubin and Schenker (1986), Barnard and Rubin (1999) and Reiter (2007).

If Condition 1 does not hold but $m \to \infty$, then $\mathbb{D}$ is simplified to $\mathbb{D}_\infty$, which can be represented by

$$\mathbb{D}_\infty := \frac{1}{k(1+\mu_r)} \sum_{j=1}^{k} (1 + r_j) G_j, \quad \text{where } \mu_r := \frac{1}{k} \sum_{j=1}^{k} r_j. \tag{1.10}$$

Meng (1990) and Li, Raghunathan and Rubin (1991) proposed approximating $\mathbb{D}_\infty$ by $\widehat{c}_0 + \widehat{c}_1 \chi^2_k/k$ via matching their first two moments. The coefficients $\widehat{c}_0$ and $\widehat{c}_1$ are estimates of $c_1 := \{1 + \sqrt{\sigma_r^2/(1+\mu_r^2)}\}^{1/2}$ and $c_0 := 1 - c_1$, where $\sigma_r^2 := \sum_{j=1}^{k}(r_j - \mu_r)^2/k$. It is remarked that Meng (1990) is an improvement over Li, Raghunathan and Rubin (1991). Meng (1990) proposed estimating μ_r and σ_r^2 by $\widetilde{\widetilde{\mu}}_{r,\mathrm{W}} := \mathrm{tr}\{\widehat{B}\bar{V}^{-1}\}/k$ and $\widetilde{\widetilde{\sigma}}^2_{r,\mathrm{W}} :=$

$\mathrm{tr}\{(\widehat{B}\bar{V}^{-1})^2\}/k - \widetilde{\widetilde{\mu}}^2_{r,\mathrm{W}}\{1 + k/(m-1)\}$, respectively, where $\widehat{B} := \sum_{\ell=1}^m (\widehat{\theta}^\ell - \bar{\theta})(\widehat{\theta}^\ell - \bar{\theta})^{\mathrm{T}}/(m-1)$.

If both Condition 1 and $m \to \infty$ are satisfied, then $\mathbb{D} \sim \chi^2_k/k$, that is, the standard reference distribution in (1.4). This very rough approximation was discussed in Section 3.2 of Li, Raghunathan and Rubin (1991).

1.4. *Theoretical, implementational and universal concerns.* Although fairly complete theories for performing MI tests are available, they are not fully satisfactory because of the following three concerns.

The first concern is a theoretical consideration. As we discussed in Section 1.3, all existing MI tests (e.g., T1, T2 and T3 in Table 1) require Condition 1 and/or $m \to \infty$. In many real applications, they are very restrictive and can hardly be satisfied. Roughly speaking, Condition 1 means that all parameters in $\theta = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}$ are equally impacted by the missing data. It is violated in many applications, for example, linear regression (see Section 6.1), testing variance-covariance matrix (see Section B.3 of the Supplementary Material), etc. The assumption of $m \to \infty$ does not match the current practice as well. For public-use datasets, the dataset owners may refuse to release a large number of imputed datasets to the public due to storage problem, processing inconvenience, or privacy concern, therefore, $m \le 30$, or even $m \le 10$, is typically used; see Remark 1.1 for an example. Moreover, unlike typical simulation study, the analysts cannot arbitrarily generate a large number imputed datasets. Hence, the value of m is typically not very large.

The second concern is about implementation. MI tests are most useful if users only need to apply their intended complete-data testing device $X \mapsto d(X)$, that is, either (1.1), (1.2) or (1.3), repeatedly to the imputed datasets $X^1, \ldots, X^m$. Unfortunately, this ideal minimal requirement is not sufficient for most existing MI tests (e.g., T2, and T3 in Table 1). Instead of merely requiring the device $d(\cdot)$, they may need (i) the variance-covariance matrix estimator $\widehat{V}(X)$, or (ii) a nonstandard likelihood function. The device (i) is required for computing, for example, $\widetilde{\mu}_{r,\mathrm{W}}$ (Li, Raghunathan and Rubin (1991)), and $\widetilde{\widetilde{\mu}}_{r,\mathrm{W}}$ and $\widetilde{\widetilde{\sigma}}^2_{r,\mathrm{W}}$ (Meng (1990)). It is infeasible because, in some problems, $\widehat{V}(X)$ is typically unavailable in standard computer subroutines. For example, to perform a G-test, that is, the LR test for goodness of fit in contingency tables (see Section 6.4.3 of Shao (1999)), one may use the R function `GTest` in the package `DescTools`. But this function does not provide $\widehat{V}(X)$ in the output. In some problems, computing $\widehat{V}(X)$ is highly nontrivial, for example, testing variance-covariance matrices; see Section B.3 of the Supplementary Material. The device (ii) usually requires analysts' effort to build, so, it can be challenging or simply troublesome for them. For example, $(X, \psi_0, \psi_1) \mapsto d_{\mathrm{L}}(\psi_1, \psi_0 \mid X)$, instead of the standard $X \mapsto d_{\mathrm{L}}(X)$, is required in $\widetilde{\mu}_{r,\mathrm{L}}$ (Meng and Rubin (1992)). Arguably, most (if not all) LR test statistic subroutines are not built in this way. Although the recent work by Chan and Meng (2021+) only requires $X \mapsto d_{\mathrm{L}}(X)$ for performing MI tests, it assumes equal OMI and is restricted to LR tests. In order to handle unequal OMI, we need a substantially more sophisticated principle and technique, which are completely novel and have never been discussed in the literature.

The third concern is about universality. Computing $\widetilde{D}_{\mathrm{W}}$ and $\widetilde{D}_{\mathrm{L}}$ require different algorithms as the functional forms of $(\widetilde{d}'_{\mathrm{W}}, \widetilde{d}''_{\mathrm{W}})$ in (1.5) and $(\widetilde{d}'_{\mathrm{L}}, \widetilde{d}''_{\mathrm{L}})$ in (1.6) are different. It is inconvenient for users. In addition, the existing MI procedures only cover Wald's test and LR test but not RS test. Since many tests are RS tests in nature (see, e.g., Bera and Bilias (2001)), it reveals a gap between MI testing theory and practical usage. As discussed in Rao (2005), there are many reasons to use RS test. For example, RS test does not require fitting the full models, which are nonidentifiable or computationally intensive in some problems;

see Section B.6 of the Supplementary Material for an example. So, it is desirable to have a unified MI procedure for all tests.

This paper addresses these three problems. A general, unified and feasible MI test without requiring Condition 1 or $m \to \infty$ is proposed. It only requires the analysts to have a standard complete-data test device $X \mapsto d(X)$, where $d(\cdot)$ can be either $d_{\mathrm{W}}(\cdot)$, $d_{\mathrm{L}}(\cdot)$ or $d_{\mathrm{R}}(\cdot)$ defined in (1.1)–(1.3).

2. Motivation and plan of proposal.

2.1. *Motivation.* Complete-data Wald's and LR test statistics are asymptotically equivalent under H_0; see Section 4.4 of Serfling (2001). Meng and Rubin (1992) showed that this asymptotic equivalence continues to hold for the MI statistics defined in (1.5) and (1.6), that is, $\widetilde{d}'_{\mathrm{W}} \sim \widetilde{d}'_{\mathrm{L}}$ and $\widetilde{d}''_{\mathrm{W}} \sim \widetilde{d}''_{\mathrm{L}}$, where $A_n \sim B_n$ means that $A_n - B_n \xrightarrow{\mathrm{pr}} 0$, and "$\xrightarrow{\mathrm{pr}}$" denotes convergence in probability. So, we may asymptotically represent $\widetilde{d}'_{\mathrm{W}}$ and $\widetilde{d}'_{\mathrm{L}}$ as $\widetilde{d}'$, and represent $\widetilde{d}''_{\mathrm{W}}$ and $\widetilde{d}''_{\mathrm{L}}$ as $\widetilde{d}''$. Then $\widetilde{D}_{\mathrm{W}}$ and $\widetilde{D}_{\mathrm{L}}$ equal to $\widetilde{D}$ asymptotically, where

$$\widetilde{D} := \frac{\widetilde{d}''}{k\{1 + (1 + \frac{1}{m})\widetilde{\mu}_r\}} \quad \text{and} \quad \widetilde{\mu}_r := \frac{\widetilde{d}' - \widetilde{d}''}{k(m-1)/m}. \tag{2.1}$$

From (2.1), we know that the MI test statistic $\widetilde{D}$ depends on $X^1, \ldots, X^m$ only through $\widetilde{d}'$ and $\widetilde{d}''$. So, all information contained in $X^1, \ldots, X^m$ is summarized by $\widetilde{d}'$ and $\widetilde{d}''$. In general, the two-number summary $(\widetilde{d}', \widetilde{d}'')$ is not enough for estimating k unknown parameters $r_1, \ldots, r_k$. Hence, in order to estimate all individual $r_1, \ldots, r_k$, it is necessary to derive a more general class of MI statistics other than $\widetilde{d}'$ and $\widetilde{d}''$.

Besides, we would like to have a MI testing procedure that can be completed solely by the device $d(\cdot)$. In order to achieve this goal, we begin with representing the statistics $\widetilde{d}'$ and $\widetilde{d}''$ in terms of $d(\cdot)$. According to Xie and Meng (2017), we can asymptotically represent $\widetilde{d}'$ and $\widetilde{d}''$ as

$$\widetilde{d}' \sim \frac{1}{m}\sum_{\ell=1}^{m} d(X^{\ell}) \quad \text{and} \quad \widetilde{d}'' \sim \frac{1}{m} d\big(X^{\{1:m\}}\big), \tag{2.2}$$

where $X^{\{1:m\}} := [(X^1)^{\mathrm{T}}, \ldots, (X^m)^{\mathrm{T}}]^{\mathrm{T}}$ is a stacked dataset. Using (2.2), we may interpret $\widetilde{d}'$ and $\widetilde{d}''$ as summary statistics via stacking *one* and *all* imputed datasets, respectively. Stacking different numbers of imputed datasets produces distinct inferential tools. For example, $\widetilde{d}''$, the numerator of the MI test statistic $\widetilde{D}$ in (2.1), measures the amount of evidence against H_0; whereas $\widetilde{d}' - \widetilde{d}''$, which is proportional to the estimator $\widetilde{\mu}_r$ in (2.1), measures the amount of information loss due to missing data. Hence, it motivates us to derive new MI statistics by stacking imputed datasets in various ways.

2.2. *Overview and plan of proposal.* Following the motivations in Section 2.1, we propose a MI test statistic that admits the form

$$\widehat{D} := \frac{\widehat{d}^{\{1:m\}}}{k\{1 + (1 + \frac{1}{m})\widehat{\mu}_r\}}, \tag{2.3}$$

where $\widehat{d}^{\{1:m\}}$ and $\widehat{\mu}_r$ are some statistics to be derived so that (i) they can be computed solely by using the device $d(\cdot)$, and (ii) the asymptotic null distribution of $\widehat{D}$ is $\mathbb{D}$ (see (1.8)) without any ad-hoc approximation. The goals (i) and (ii) are completed in Sections 3 and 5, respectively. Since the distribution of $\mathbb{D}$ depends on $r_1, \ldots, r_k$, we propose to approximate them by their estimators $\widehat{r}_j$'s, which are derived in Section 4 by using various stacked statistics. Algorithm 2.1 computes our proposed test statistic $\widehat{D}$ and the corresponding p-value. We will explain the steps of Algorithm 2.1 in the subsequent sections.

Algorithm 2.1: Asymptotically correct MI test for H_0

Input:
(i) $X \mapsto d(X)$ – any complete-data testing device in (1.1)–(1.3);
(ii) $X^1, \ldots, X^m$ – m properly imputed datasets; and
(iii) k – the dimension of Θ.
begin
Stack $X^1, \ldots, X^m$ row-by-row to form $X^{\{1:m\}}$.
Compute $\widehat{d}^{\{1:m\}} \leftarrow d(X^{\{1:m\}})/m$.
for $\ell \in \{1, \ldots, m\}$ **do**
Stack $X^1, \ldots, X^{\ell-1}, X^{\ell+1}, \ldots, X^m$ row-by-row to form $X^{\{-\ell\}}$.
Compute $\widehat{d}^{\{-\ell\}} \leftarrow d(X^{\{-\ell\}})/(m-1)$.
Compute $\widehat{d}^{\{\ell\}} \leftarrow d(X^{\ell})$.
Compute $\widehat{T}_\ell \leftarrow (m-1)\widehat{d}^{\{-\ell\}} + \widehat{d}^{\{\ell\}} - m\widehat{d}^{\{1:m\}}$.
Compute $\widehat{t}_j \leftarrow \sum_{\ell=1}^m \widehat{T}_\ell^j/m$ for each $j = 1, \ldots, k$.
Compute $\widehat{r}_{1:k} \leftarrow M_1^{-1}(M_2^{-1}(\widehat{t}_{1:k}))$ according to Proposition 4.1.
Compute $\widehat{D}$ according to (2.3).
Draw $G_j^{(\iota)} \sim \chi_1^2$ and $H_j^{(\iota)} \sim \chi_{m-1}^2/(m-1)$ independently for $\iota = 1, \ldots, N$ and $j = 1, \ldots, k$.
Compute $\widehat{\mathbb{D}}^{(\iota)}$, $\iota = 1, \ldots, N$, according to (5.1). Set $N = 10^4$ by default.
Compute $\widehat{p} \leftarrow \sum_{\iota=1}^N \mathbb{1}(\widehat{\mathbb{D}}^{(\iota)} \geq \widehat{D})/N$.
return: $\widehat{p}$ – the p-value for testing H_0 against H_1.

3. Methodology and principle.

3.1. *Stacking principle.* In this section, we introduce a new class of MI statistics by stacking $X^1, \ldots, X^m$ in various ways. As we shall see in Theorem 3.1 below, stacking them differently extracts different information from $X^1, \ldots, X^m$. We refer this phenomenon to a *stacking principle*. Define the stacked dataset X^S by stacking $\{X^\ell : \ell \in S\}$ row-by-row for some nonempty $S \subseteq \{1, \ldots, m\}$. For example, $X^S = [(X^{\ell_1})^{\mathrm{T}}, \ldots, (X^{\ell_s})^{\mathrm{T}}]^{\mathrm{T}}$ is an $(ns) \times p$ matrix if $S = \{\ell_1, \ldots, \ell_s\}$. Define

$$\widehat{d}^S := \frac{1}{|S|} d(X^S), \tag{3.1}$$

where $|S|$ is the cardinality of S, and $d(\cdot)$ is any testing device in (1.1)–(1.3). In particular, we have $\widehat{d}^{\{\ell\}} = d(X^\ell)$, $\widehat{d}^{\{-\ell\}} = d(X^{\{-\ell\}})/(m-1)$, and $\widehat{d}^{\{1:m\}} = d(X^{\{1:m\}})/m$, where $\{1:m\} := \{1, \ldots, m\}$ and $\{-\ell\} := \{1, \ldots, m\} \setminus \{\ell\}$. We denote $\widehat{d}^S$ by $\widehat{d}_{\mathrm{W}}^S$, $\widehat{d}_{\mathrm{L}}^S$ and $\widehat{d}_{\mathrm{R}}^S$ to emphasize that d_{W}, d_{L} and d_{R} are used, respectively.

It is worth mentioning that "stacking" is a *universal* operation. Since it is problem-independent, the analysts can apply this operation to all kinds of testing problems universally. This nature is similar to some well-known procedures, for example, bootstrapping (Efron (1979)), jackknife resampling (Quenouille (1956)), subsampling (Politis, Romano and Wolf (1999)), etc. All these procedures are model-free and fully nonparametric.

The usefulness of $\widehat{d}^S$ can be seen from its asymptotic distribution. To derive it, we need some RCs.

CONDITION 2. The observed-data MLE $\widehat{\theta}_{\mathrm{obs}}$ of θ satisfies $T^{-1/2}(\widehat{\theta}_{\mathrm{obs}} - \theta^\star) \Rightarrow \mathrm{N}_k(0_k, I_k)$, where $T := I_{\mathrm{obs}}^{-1}$ is well-defined, $\theta^\star$ is the true value of θ; 0_k is a k-vector of zeros; and I_k is a $k \times k$ identity matrix.

CONDITION 3. The imputed statistics $(\widehat{\theta}^1, \widehat{V}^1), \ldots, (\widehat{\theta}^m, \widehat{V}^m)$ are conditionally independent given X_{obs}. Moreover, for each $\ell = 1, \ldots, m$, they satisfy $\{B^{-1/2}(\widehat{\theta}^\ell - \widehat{\theta}_{\mathrm{obs}}) \mid$

$X_{\mathrm{obs}}\} \Rightarrow \mathrm{N}_k(0_k, I_k)$ and $\{T^{-1}(\widehat{V}^\ell - V) \mid X_{\mathrm{obs}}\} \xrightarrow{\mathrm{pr}} O_k$, where $B := I_{\mathrm{obs}}^{-1} - I_{\mathrm{com}}^{-1}$ and $V := I_{\mathrm{com}}^{-1}$ are well-defined, and O_k is a $k \times k$ matrix of zeros.

Condition 2 is satisfied under the usual RCs that guarantee asymptotic normality of MLEs; see, for example, Wang and Robins (1998) and Kim and Shao (2013). Condition 3 is satisfied if a proper imputation model (Rubin (1987)) is used. The posterior predictive distribution $f(X_{\mathrm{mis}} \mid X_{\mathrm{obs}})$ is an example of proper imputation models, which have been widely used and adopted in MI; see Sections 2.4–2.7 of Rubin (1996) for a comprehensive discussion of this assumption. We emphasize that the analysts do not need to compute or know the imputed statistics $(\widehat{\theta}^\ell, \widehat{V}^\ell)$ by themselves. Condition 3 guarantees that the imputer does his/her imputation job correctly. It is remarked that if the analyst's and imputer's models are uncongenial (Meng (1994)), then the standard Rubin's MI procedure may not be valid. Generalizing MI procedures to the uncongenial case is not completely solved yet. Interested readers are referred to a recent discussion article (Xie and Meng (2017)) for a simple remedy when $m \to \infty$. Extending our proposed method to the uncongenial case is left for further study.

DEFINITION 1. Let $\theta^\star$ be the true value of θ, and θ_0 be the null value of θ specified in H_0. The difference $\theta^\star - \theta_0$ satisfies that $\sqrt{n}A(\theta^\star - \theta_0) \to \delta \equiv (\delta_1, \ldots, \delta_k)^{\mathrm{T}}$ for some δ and invertible matrix A.

Definition 1 defines a sequence of local alternative hypotheses; see, for example, van der Vaart (2000). Note that it is required for proving asymptotic equivalence of the test statistics $d_{\mathrm{W}}(X)$, $d_{\mathrm{L}}(X)$ and $d_{\mathrm{R}}(X)$; see, for example, Serfling (2001) and Lehmann and Romano (2005). This general setting allows us to prove the validity of MI estimators, even when H_0 is not true. Moreover, practical testing problems are more challenging under a local alternative hypothesis than under an obviously wrong fixed alternative hypothesis because, under a fixed alternative hypothesis, all three test statistics obviously diverge to infinity, and have power one asymptotically. Hence, our setting is sufficient for most practical applications. The following theorem states the asymptotic distribution of $\widehat{d}^S$.

THEOREM 3.1. *Assume Conditions 2–3. Let $W_j, Z_{j1}, \ldots, Z_{jm}$, $j = 1, \ldots, k$, be independent* $\mathrm{N}(0, 1)$ *random variables, and $\bar{Z}_{j(S)} = \sum_{\ell \in S} Z_{j\ell}/|S|$ be the average of $\{Z_{j\ell} : \ell \in S\}$. Then, for any nonempty multiset S from $\{1, \ldots, m\}$,*

$$\widehat{d}^S \Rightarrow \mathbb{d}^S := \sum_{j=1}^{k} \{\delta_j + (1 + r_j)^{1/2} W_j + r_j^{1/2} \bar{Z}_{j(S)}\}^2, \tag{3.2}$$

where the convergence is true jointly for all S.

Theorem 3.1 is true for any multiset S, for example, $S = \{1, 1, 2, 3\}$. In this case, $\bar{Z}_{j(S)} = (Z_{j1} + Z_{j1} + Z_{j2} + Z_{j3})/4$. We emphasize that Theorem 3.1 requires neither Condition 1 nor $m \to \infty$. So, it is in line with the practical situation. The convergence (3.2) is with respect to the regime $n \to \infty$, hence, it is simply a usual large-sample asymptotic result. More importantly, Theorem 3.1 sheds light on performing hypothesis tests because of two reasons. First, the limiting distribution $\mathbb{d}^S$ depends on δ. When H_0 is true, that is, $\delta_1 = \cdots = \delta_k = 0$, the statistic $\widehat{d}^S$ converges weakly to a nondegenerated distribution. When $\delta_j \to \infty$ for some j, $\widehat{d}^S$ diverges to infinity. So, the statistic $\widehat{d}^S$ can be used to test H_0 for every nonempty S. Second, $\mathbb{d}^S$ depends on $r_1, \ldots, r_k$, but the dependence on $r_1, \ldots, r_k$ varies among S. Consequently, pooling information from different $\widehat{d}^S$ may help to estimate $r_1, \ldots, r_k$.

However, $\widehat{d}^S$ is not immediately useful because of two reasons. First, $\mathbb{d}^S$ depends on $r_1, \ldots, r_k$ in a complicated way. In particular, the jth summand on the right-hand side of

(3.2) depends on r_j *nonlinearly*. Thus, it is not clear how to use $\widehat{d}^S$ to estimate $r_1, \ldots, r_k$. Second, for any nonempty S_1 and S_2, the statistics $\widehat{d}^{S_1}$ and $\widehat{d}^{S_2}$ are asymptotically dependent through *both* $\{W_j\}$ and $\{Z_{j\ell}\}$. The random variables $W_1, \ldots, W_k$ always appear in the limiting distributions of $\widehat{d}^{S_1}$ and $\widehat{d}^{S_2}$, no matter how S_1 and S_2 are chosen. So, $\widehat{d}^{S_1}$ and $\widehat{d}^{S_2}$ are too correlated to be useful if S_1 and S_2 are blindly selected.

3.2. *Stacked multiple imputation—a new class of MI procedures.* In this section, we propose a novel methodology for properly using $\widehat{d}^S$. According to the discussion in Section 3.1, we know that $\widehat{d}^S$ is too complex to be useful because of two sources of dependence: (i) the nonlinear dependence of the limiting distribution of $\widehat{d}^S$ on $r_1, \ldots, r_k$, and (ii) the strong probabilistic dependence among different $\widehat{d}^S$ through the random variables $W_1, \ldots, W_k$. In this section, we propose a method to get rid of all these two unwanted sources of dependence.

For any nonempty sets $S_1, S_2 \subseteq \{1, \ldots, m\}$ such that $S_1 \neq S_2$, define

$$\widehat{T}_{S_1,S_2} = \frac{|S_1| + |S_2|}{|S_1| + |S_2| - 2|S_1 \cap S_2|}\{|S_1|\widehat{d}^{S_1} + |S_2|\widehat{d}^{S_2} - (|S_1| + |S_2|)\widehat{d}^{S_1 \oplus S_2}\},$$

where $S_1 \oplus S_2$ is the multiset addition, for example, $\{1, 3\} \oplus \{1, 2\} = \{1, 1, 2, 3\}$. We call $\widehat{T}_{S_1,S_2}$ a *stacked multiple imputation* (SMI) statistic. Note that the SMI statistic $\widehat{T}_{S_1,S_2}$ can be computed solely by stacking the imputed datasets and applying the complete-data testing device $d(\cdot)$. Besides, as we shall see in Proposition 3.2 below, $\widehat{T}_{S_1,S_2}$ is free of the two aforementioned sources of unwanted dependence. Hence, the SMI statistic $\widehat{T}_{S_1,S_2}$ has nice computational and theoretical properties. Consequently, it is qualified to be a building block for all MI procedures proposed in this paper. Let

$$R_\tau := \sum_{j=1}^{k} r_j^\tau, \quad \tau = 1, \ldots, k. \tag{3.3}$$

The asymptotic distribution of $\widehat{T}_{S_1,S_2}$ and its properties are shown below.

PROPOSITION 3.2. *Assume Conditions 2–3. Let $S_1, S_2 \subseteq \{1, \ldots, m\}$ be any nonempty and nonidentical sets. Define $W_j, Z_{j1}, \ldots, Z_{jm}$, $j = 1, \ldots, k$ as in Theorem 3.1.*

1. *$\widehat{T}_{S_1,S_2} \Rightarrow \mathbb{T}_{S_1,S_2}$, where $\mathbb{T}_{S_1,S_2}$ is represented as*

$$\mathbb{T}_{S_1,S_2} := \frac{|S_1| \times |S_2|}{|S_1| + |S_2| - 2|S_1 \cap S_2|} \sum_{j=1}^{k} r_j\{\bar{Z}_{j(S_1)} - \bar{Z}_{j(S_2)}\}^2. \tag{3.4}$$

2. *$\mathbb{T}_{S_1,S_2}$ has the same marginal distribution as $\mathbb{T} := \sum_{i=1}^{k} r_i U_i$, where $U_1, \ldots, U_k \sim \chi_1^2$ independently.*

3. *Let $t_\tau := \mathsf{E}(\mathbb{T}^\tau)$ for $\tau \in \{1, \ldots, k\}$, and $t_0 := 1$. Then $t_1, \ldots, t_k$ can be found iteratively as follows*:

$$t_1 = R_1 \quad \text{and} \quad t_\tau = \sum_{j=1}^{\tau} \frac{(\tau - 1)!}{(\tau - j)!} 2^{j-1} R_j t_{\tau - j} \tag{3.5}$$

for $\tau = 2, \ldots, k$. In particular, $\mathsf{E}(\mathbb{T}) = R_1$ and $\mathsf{Var}(\mathbb{T}) = 2R_2$.

According to Proposition 3.2(1), the limiting distribution $\mathbb{T}_{S_1,S_2}$ depends on $r_1, \ldots, r_k$ linearly, and is independent on the random variables $W_1, \ldots, W_k$ that appear in (3.2). Hence, the SMI statistic $\widehat{T}_{S_1,S_2}$ "filters" out the two unwanted dependence structures. Consequently,

$\widehat{T}_{S_1,S_2}$ is easier to work with. Proposition 3.2(2) states that $\widehat{T}_{S_1,S_2}$'s are asymptotically identically distributed over different S_1, S_2. Proposition 3.2(3) states that the τth moments of $\mathbb{T}$ depends on $r_1, \ldots r_k$ only through $R_1, \ldots, R_\tau$ for each $\tau = 1, \ldots, k$.

Proposition 3.2 implies that the sample τth moments of $\widehat{T}_{S_1,S_2}$ over a set of pairs of (S_1, S_2) is a natural estimator of t_τ. Let $\Lambda \subseteq \mathcal{L} := \{(S_1, S_2) \subseteq \{1, \ldots, m\}^2 : S_1, S_2 \neq \varnothing; S_1 \neq S_2\}$ so that Λ is a set of appropriately chosen pairs of (S_1, S_2). For $\tau \in \{1, \ldots, k\}$, define an estimator of t_τ by

$$\widehat{t_\tau} := \widehat{t_\tau}(\Lambda) := \frac{1}{|\Lambda|} \sum_{(S_1,S_2)\in\Lambda} \widehat{T}^\tau_{S_1,S_2}, \tag{3.6}$$

where the short notation $\widehat{t_\tau}$ is used when Λ is clear in the context. Some examples of Λ are given below.

EXAMPLE 3.1. The following selection rules of Λ are suggested. Let

$$\begin{aligned} \Lambda_{\text{Jack}} &:= \big\{(\{\ell\}, \{-\ell\})\big\}_{1\le\ell\le m}, \qquad \Lambda_{\text{Full}} := \big\{(\{\ell\}, \{1:m\})\big\}_{1\le\ell\le m}, \\ \Lambda_{\text{Pair}} &:= \big\{(\{\ell\}, \{\ell'\})\big\}_{1\le\ell<\ell'\le m} \end{aligned} \tag{3.7}$$

be the *Jackknife*, *full* and *pair* selection rules for Λ. Note that $|\Lambda_{\text{Jack}}| = |\Lambda_{\text{Full}}| = m$, whereas $|\Lambda_{\text{Pair}}| = m(m-1)/2$. Putting $\Lambda = \Lambda_{\text{Jack}}, \Lambda_{\text{Full}}, \Lambda_{\text{Pair}}$ into (3.6), we obtain the following three estimators of t_τ:

$$\widehat{t_\tau}(\Lambda_{\text{Jack}}) = \frac{1}{m}\sum_{\ell=1}^{m} \widehat{T}^\tau_{\{\ell\},\{-\ell\}}, \qquad \widehat{t_\tau}(\Lambda_{\text{Full}}) = \frac{1}{m}\sum_{\ell=1}^{m} \widehat{T}^\tau_{\{\ell\},\{1:m\}},$$

$$\widehat{t_\tau}(\Lambda_{\text{Pair}}) = \frac{2}{m(m-1)} \sum_{\ell'=2}^{m}\sum_{\ell=1}^{\ell'-1} \widehat{T}^\tau_{\{\ell\},\{\ell'\}},$$

respectively. Since $\widehat{t_\tau}(\Lambda_{\text{Pair}})$ requires stacking at most two datasets, it should be used when computing $\mathcal{d}(X)$ is difficult for a large dataset X. Although the device $\mathcal{d}(\cdot)$ has to be implemented $3m(m-1)/2 = O(m^2)$ times, the computations can be parallelized easily. On the other hand, computing $\widehat{t_\tau}(\Lambda_{\text{Jack}})$ and $\widehat{t_\tau}(\Lambda_{\text{Full}})$ requires implementing the device $\mathcal{d}(\cdot)$ $3m = O(m)$ times only. It is preferable when m is large.

The message behind Example 3.1 is that stacked MI is flexible enough to allow users to choose the most computationally viable statistics according to their problems. Although testing on stacked datasets is more computationally intensive and requires more computing memory, these computational requirements are usually affordable by standard laptop computers nowadays. Indeed, the increase in *computing cost* is used to exchange for a decrease in *human time cost* of deriving or searching nonstandard computing devices required in the exiting tests T2 and T3 stated in Table 1.

3.3. *Asymptotic properties*. In this section, we study the asymptotic properties of $\widehat{t_\tau}(\Lambda)$ in the $\mathcal{L}^2$ sense. From now on, we denote the indicator function by $\mathbb{1}(\cdot)$. The following condition is required for developing $\mathcal{L}^2$ convergence results.

CONDITION 4. For any nonempty and nonidentical $S_1, S_2 \subseteq \{1, \ldots, m\}$, denote $\widehat{T}_{S_1,S_2}(n) = \widehat{T}_{S_1,S_2}$ as a sequence indexed by the sample size n. The sequence $\{\widehat{T}^{2\tau}_{S_1,S_2}(n) : n \in \mathbb{N}\}$ is assumed uniformly integrable, that is,

$$\lim_{C\to\infty} \sup_{n\in\mathbb{N}} \mathsf{E}\big[\widehat{T}^{2\tau}_{S_1,S_2}(n)\mathbb{1}\big\{\big|\widehat{T}_{S_1,S_2}(n)\big| > C\big\}\big] = 0.$$

The following theorem derives the limits of $\mathsf{E}\{\widehat{t}_\tau(\Lambda)\}$ and $\mathsf{Var}\{\widehat{t}_\tau(\Lambda)\}$ as $n \to \infty$.

THEOREM 3.3. *Assume Conditions 2–4. Let $m > 1$, $\Lambda \subseteq \mathcal{L}$ and $\tau \in \mathbb{N}$. Then the following hold.* (1) $\mathsf{E}\{\widehat{t}_\tau(\Lambda)\} \to t_\tau$ *as* $n \to \infty$. (2) *For* $(S_1, S_2), (S_3, S_4) \in \Lambda$, *define*

$$\rho(S_1, S_2, S_3, S_4) := (s_1 + s_2 - 2s_{12})^{-1/2}(s_3 + s_4 - 2s_{34})^{-1/2} \times \left(\frac{s_{13}}{s_1 s_3} - \frac{s_{14}}{s_1 s_4} - \frac{s_{23}}{s_2 s_3} + \frac{s_{24}}{s_2 s_4}\right),$$

where $s_a = |S_a|$ *and* $s_{ab} = |S_a \cap S_b|$ *for* $a, b \in \{1, 2, 3, 4\}$. *Then*

$$\mathsf{Var}\{\widehat{t}_\tau(\Lambda)\} \to \frac{1}{|\Lambda|^2} \sum_{(S_1,S_2)\in\Lambda} \sum_{(S_3,S_4)\in\Lambda} \rho^2(S_1, S_2, S_3, S_4) C_\tau(S_1, S_2, S_3, S_4), \tag{3.8}$$

where $C_\tau(S_1, S_2, S_3, S_4)$ *satisfies* $\sup_{(S_1,S_2),(S_3,S_4)\in\Lambda} |C_\tau(S_1, S_2, S_3, S_4)| \le C_\tau$ *for some finite* C_τ *which depends on* $\tau, r_1, \ldots, r_k$. *In particular,* $C_1(S_1, S_2, S_3, S_4) = 2R_2$.

From Theorem 3.3, the choice of Λ only affects the limit of $\mathsf{Var}\{\widehat{t}_\tau(\Lambda)\}$ but not the limit of $\mathsf{E}\{\widehat{t}_\tau(\Lambda)\}$. The asymptotic variance of $\widehat{t}_\tau(\Lambda)$ is affected by two factors: (i) $1/|\Lambda|^2$ and (ii) the double summation in (3.8). If $|\Lambda|$ is too small, the first factor $1/|\Lambda|$ is large. On the other hand, if $|\Lambda|$ is too large, then $\{\widehat{T}_{S_1,S_2} : (S_1, S_2) \in \Lambda\}$ may be highly correlated in the sense that most of the $\rho(S_1, S_2, S_3, S_4)$'s in (3.8) are large. Hence, Λ needs to be chosen carefully so that the estimator $\widehat{t}_\tau(\Lambda)$ has a small asymptotic variance. Theorem 3.3 is easy to use. Once the set Λ is fixed, one can easily compute $\rho(S_1, S_2, S_3, S_4)$. By simple counting, the order of magnitude of the asymptotic variance of $\widehat{t}_\tau(\Lambda)$ can also be found. In particular, we show that $\widehat{t}_\tau(\Lambda_{\text{Jack}})$, $\widehat{t}_\tau(\Lambda_{\text{Full}})$ and $\widehat{t}_\tau(\Lambda_{\text{Pair}})$ in Example 3.1 are all good estimators of t_τ.

COROLLARY 3.4. *Define* Λ_{Jack}, Λ_{Full} *and* Λ_{Pair} *according to* (3.7). *Assume Conditions 2–4. For any* $\mathcal{S} \in \{\text{Jack}, \text{Full}, \text{Pair}\}$ *and any* $\tau \in \{1, \ldots, k\}$, *we have* $\mathsf{Var}\{\widehat{t}_\tau(\Lambda_{\mathcal{S}})\} \to V_{\mathcal{S},\tau}(m)$ *as* $n \to \infty$, *where* $V_{\mathcal{S},\tau}(m) = O(1/m)$ *as* $m \to \infty$. *In particular,* $V_{\text{Jack},1}(m) = V_{\text{Full},1}(m) = V_{\text{Pair},1}(m) = 2R_2/(m-1)$.

Corollary 3.4 shows that, asymptotically, the precision of $\widehat{t}_\tau(\Lambda_{\mathcal{S}})$ increases with m for any selection rule $\mathcal{S} \in \{\text{Jack}, \text{Full}, \text{Pair}\}$. Together with Theorem 3.3(1), the mean squared error (MSE) of $\widehat{t}_\tau(\Lambda_{\mathcal{S}})$, that is, $\mathsf{MSE}\{\widehat{t}_\tau(\Lambda_{\mathcal{S}})\} := \mathsf{E}\{\widehat{t}_\tau(\Lambda_{\mathcal{S}}) - t_\tau\}^2$ decreases in the order of $O(1/m)$ as m increases.

Unless otherwise stated, we use $\Lambda = \Lambda_{\text{Jack}}$ by default. Then the estimator $\widehat{t}_\tau(\Lambda_{\text{Jack}})$ is as simple as

$$\widehat{t}_\tau := \frac{1}{m}\sum_{\ell=1}^{m} \widehat{T}_\ell^\tau, \quad \text{where } \widehat{T}_\ell := \widehat{d}^{\{\ell\}} + (m-1)\widehat{d}^{\{-\ell\}} - m\widehat{d}^{\{1:m\}}.$$

4. Estimation of OMI. This section proposes estimators for all individual $r_1, \ldots, r_k$. For notational simplicity, we denote $r_{1:k} = (r_1, \ldots, r_k)^{\mathrm{T}}$. We also abbreviate other variables similarly, for example, $t_{1:k} = (t_1, \ldots, t_k)^{\mathrm{T}}$.

From Proposition 3.2, $t_{1:k}$ are defined via the following two-step mapping:

$$r_{1:k} \overset{M_1}{\mapsto} R_{1:k} \overset{M_2}{\mapsto} t_{1:k}, \tag{4.1}$$

where the maps M_1 and M_2 are defined according to (3.3) and (3.5), respectively. From Theorem 3.3, $\widehat{t}_{1:k}$ are good estimators of $t_{1:k}$. Our goal is to estimate $r_{1:k}$, through "reverse engineering" the two-step transformation (4.1). However, it is impossible unless $r_{1:k}$ is uniquely determined by $t_{1:k}$. The following proposition shows that (4.1) is a one-to-one function, that is, the function inverse of $M_2(M_1(\cdot))$ always exists.

PROPOSITION 4.1. *Define the functions $M_1(\cdot)$ and $M_2(\cdot)$ according to (3.3) and (3.5), respectively.*

1. *The inverse function $r_{1:k} = M_1^{-1}(R_{1:k})$ exists, that is, $r_{1:k}$ are uniquely determined by $R_{1:k}$. Let*

$$A := \left[\begin{array}{c|cccc} 0_{k-1} & I_{k-1} \\ \hline a_1 & a_2 & \cdots & a_k \end{array}\right], \tag{4.2}$$

where $a_k := R_1$ and $a_{k-j+1} := (R_j - \sum_{i=1}^{j-1} R_i a_{k-j+i+1})/j$, for $j = 2, \ldots, k$. Then r_j is the jth largest modulus of the eigenvalue of A for $j = 1, \ldots, k$.

2. *The inverse function $R_{1:k} = M_2^{-1}(t_{1:k})$ exists, that is, $R_{1:k}$ are uniquely determined by $t_{1:k}$ as follows: $R_1 = t_1$ and*

$$R_\tau = \frac{t_\tau}{(\tau-1)!2^{\tau-1}} - \sum_{j=1}^{\tau-1} \frac{t_{\tau-j} R_j}{(\tau-j)!2^{\tau-j}}, \quad \tau = 2, \ldots, k. \tag{4.3}$$

Proposition 4.1 implies that the parameters $r_{1:k}$ are identifiable via the moment conditions $\mathsf{E}(\mathbb{T}^\tau) = t_\tau$ $(\tau = 1, \ldots, k)$ in such a way that $r_{1:k} = M_1^{-1}(M_2^{-1}(t_{1:k}))$. Since the function $M_1^{-1}(M_2^{-1}(\cdot))$ is continuous and does not depend on any unknown, we can estimate $R_{1:k}$ and $r_{1:k}$ by

$$\widehat{R}_{1:k} := M_2^{-1}(\widehat{t}_{1:k}) \quad \text{and} \quad \widehat{r}_{1:k} := M_1^{-1}(\widehat{R}_{1:k}), \tag{4.4}$$

respectively. The corollary below states the large-n asymptotic MSEs of $\widehat{r}_j$.

COROLLARY 4.2. *Let Λ be either Λ_{Jack}, Λ_{Full} or Λ_{Pair}. Under Conditions 2–4, as $n \to \infty$, $\mathsf{MSE}(\widehat{r}_j) := \mathsf{E}(\widehat{r}_j - r_j)^2 \to V(m)$, where $V(m)$ is a function of m such that $V(m) \to 0$ as $m \to \infty$.*

Corollary 4.2 guarantees that the estimators $\widehat{r}_{1:k}$ have small MSEs when m, n are sufficiently large. The step-by-step procedure for computing $\widehat{r}_{1:k}$ with $\Lambda = \Lambda_{\text{Jack}}$ is shown in Algorithm 2.1. This algorithm is user-friendly for analysts as only the complete-data testing device $X \mapsto d(X)$ is required. It is, indeed, the minimal requirement even for compete-data testing.

Besides, it is sometimes informative to summarize $r_1, \ldots, r_k$ through their mean and variance, that is,

$$\mu_r := \frac{1}{k}\sum_{j=1}^{k} r_j \quad \text{and} \quad \sigma_r^2 := \frac{1}{k}\sum_{j=1}^{k}(r_j - \mu_r)^2.$$

These two values are required for approximating the limiting null distribution $\mathbb{D}$ in (1.8) for testing H_0, for example, T2 and T3 in Table 1; see Meng (1990), Li, Raghunathan and Rubin (1991), and Meng and Rubin (1992) for details. In Section A.1 of the Supplementary Material, we show that

$$\widehat{\mu}_r := \frac{\widehat{t}_1}{k} \quad \text{and} \quad \widehat{\sigma}_r^2 := \frac{\{k(m-1)+2\}\widehat{t}_2 - (m-1)(k+2)\widehat{t}_1^2}{2k^2(m-2)} \tag{4.5}$$

are asymptotically unbiased estimators of μ_r and σ_r^2, respectively. The precise statement and their properties are deferred to the Supplementary Material due to space constraint. There are two immediate applications of (4.5).

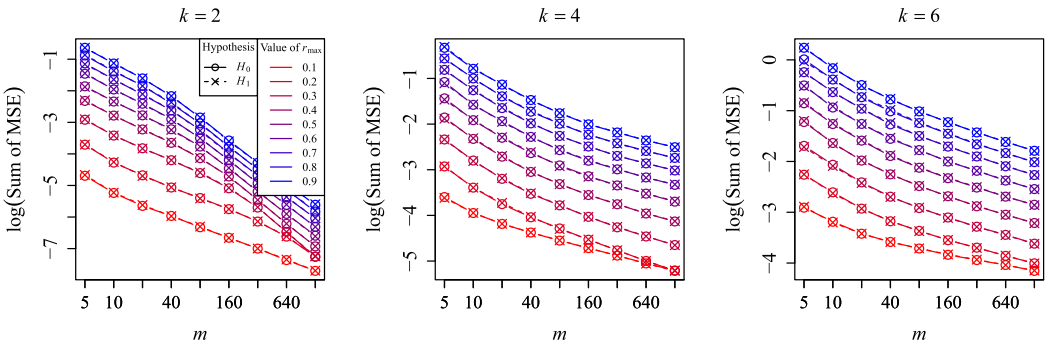

FIG. 1. *The log of sum of MSEs of* $\widehat{r}_{1:k}$, *that is*, $\log \sum_{j=1}^{k} \mathsf{E}(\widehat{r}_j - r_j)^2$; *see Example* 4.1.

1. The estimators $\widehat{\mu}_r$ and $\widehat{\sigma}_r^2$ can be used to compute the approximated null distributions of T2 and T3 in Table 1 as the distributions depend only on μ_r and σ_r^2. We emphasize that the original approximations T2 (Li, Raghunathan and Rubin (1991)) and T3 (Meng (1990)) work for Wald's test only. Although Meng and Rubin (1992) extended T2 to LR test, T2 and T3 were still inapplicable to RS test. With the proposed estimators in (4.5), the generalized T2 and T3 support Wald's, LR and RS tests.

2. The statistic $\widehat{\sigma}_r^2/\widehat{\mu}_r^2$ estimates the squared coefficient of variation of $\{r_j\}$. It can be used to construct a formal test for the validity of EOMI; see Section A.2 of the Supplementary Material for details.

We conclude this subsection with an example.

EXAMPLE 4.1. Let $r_1, \ldots, r_k$ be evenly spread in $[0.1, r_{\max}]$, that is,

$$r_j = 0.1 + (r_{\max} - 0.1)\frac{j-1}{k-1}, \quad j = 1, \ldots, k, \tag{4.6}$$

where $r_{\max} \in \{0.1, 0.2, \ldots, 0.9\}$ and $k \in \{2, 4, 6\}$. The coefficient of variation (CV) of $r_{1:k}$ increases with the value of $r_{\max}$. We evaluate $\widehat{r}_{1:k}$ under H_0 and H_1. Under H_0, $\delta_1 = \cdots = \delta_k = 0$ in (3.2). Under H_1, we set $\delta_1 = \cdots = \delta_k = 1$ in particular. In the simulation experiments, we assume that the sample size $n \to \infty$ but the number of imputation m is fixed. So, the experiments assess solely the performance of the MI procedure instead of the performance of the large-n χ^2-approximation (1.4). Applications and simulation experiments with finite n are studied in Section 6.

The sum of the MSEs of $\widehat{r}_{1:k}$, that is, $E := \sum_{j=1}^{k} \mathsf{E}(\widehat{r}_j - r_j)^2$ is shown in Figure 1. The values of E under H_0 and H_1 are nearly identical, implying that it is safe to use $\widehat{r}_{1:k}$, no matter H_0 is true or not. Second, the value of E decreases when m increases. It verifies Corollary 4.2. However, the performance of $\widehat{r}_{1:k}$ declines when $r_{\max}$ or k increases. It is reasonable as the CV of $r_{1:k}$ increases with $r_{\max}$, and the number of estimands ($r_{1:k}$) increases with k. In either case, the estimation problem is harder by nature.

5. General multiple imputation procedures.

5.1. *Hypothesis testing of model parameters.* We denote $\widehat{D}$ by $\widehat{D}_{\mathrm{W}}$, $\widehat{D}_{\mathrm{L}}$, $\widehat{D}_{\mathrm{R}}$ to emphasize that $d = d_{\mathrm{W}}, d_{\mathrm{L}}, d_{\mathrm{R}}$ is used, respectively. The limiting distribution of $\widehat{D}$ is stated below.

PROPOSITION 5.1. *Assume Conditions* 2–3. *Let* $\widehat{D} \in \{\widehat{D}_{\mathrm{W}}, \widehat{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{R}}\}$ *and* $m > 1$. *Under* H_0, *we have, as* $n \to \infty$, *that* (1) $\widehat{D} - \widetilde{D} \xrightarrow{\mathrm{pr}} 0$, *where* $\widetilde{D} \in \{\widetilde{D}_{\mathrm{W}}, \widetilde{D}_{\mathrm{L}}\}$; *and* (2) $\widehat{D} \Rightarrow \mathbb{D}$, *where* $\mathbb{D}$ *is defined in* (1.8).

Proposition 5.1 states that $\widehat{D}$ and $\widetilde{D}$ are asymptotically ($n \to \infty$) equivalent for any m, and have the same limiting null distribution $\mathbb{D}$. We emphasize again that computing $\widetilde{D}$ is not feasible as it requires problem-specific devices other than $d(\cdot)$. However, computing our proposed $\widehat{D}$ requires only $d(\cdot)$.

We propose approximating the limiting null distribution $\mathbb{D}$ by substituting $r_j = \widehat{r}_j$ into (1.8). Although it is not a named distribution, we can easily compute its quantile via Monte Carlo methods. Precisely, we first generate $G_j^{(\iota)} \sim \chi_1^2$ and $H_j^{(\iota)} \sim \chi_{m-1}^2/(m-1)$ independently for $j = 1, \ldots, k$ and $\iota = 1, \ldots, N$. Upon conditioning on $\widehat{r}_1, \ldots, \widehat{r}_k$, we can generate N random replicates of $\widehat{\mathbb{D}}$ as follows:

$$\widehat{\mathbb{D}}^{(\iota)} := \frac{\frac{1}{k}\sum_{j=1}^{k}\{1 + (1+\frac{1}{m})\widehat{r}_j\}G_j^{(\iota)}}{1 + \frac{1}{k}\sum_{j=1}^{k}(1+\frac{1}{m})\widehat{r}_j H_j^{(\iota)}}, \quad \iota = 1, \ldots, N. \tag{5.1}$$

The $100(1-\alpha_0)\%$ quantile of $\mathbb{D}$ can then be estimated by the $100(1-\alpha_0)\%$ sample quantile of $\{\widehat{\mathbb{D}}^{(1)}, \ldots, \widehat{\mathbb{D}}^{(N)}\}$, where $\alpha_0 \in (0,1)$. The sample quantile can be served as a critical value for testing $H_0 : \theta = \theta_0$. Similarly, the p-value can be found by $\widehat{p} = \sum_{\iota=1}^{N} \mathbb{1}\{\widehat{\mathbb{D}}^{(\iota)} \geq \widehat{D}\}/N$. The null hypothesis H_0 is rejected at size α_0 if $\widehat{p} < \alpha_0$. We emphasize that the proposed MI test is asymptotically correct with or without Condition 1. Step-by-step procedure for computing $\widehat{p}$ is presented in Algorithm 2.1.

In Section A.3 of the Supplementary Material, we present several alternative approximation schemes by projecting $\mathbb{D}$ to some distributions that depend only on μ_r and σ_r^2 instead of $r_{1:k}$. This idea is similar to Meng (1990). Such approximations can be used if one only wants to estimate μ_r and σ_r^2. Although these approximations are algorithmically simpler, the resulting MI tests control type-I error rates substantially worse than our proposal (5.1). A quick simulation example is presented in Section B.1 of the Supplementary Material for illustration.

5.2. *Discussion on applications.* Our proposed method uses p-value as a one-number summary for assessing variability of estimators in the presence of missing data. It can be applied not only to hypothesis testing but also other statistical procedures that require variability assessment or use p-value as a part of the automatic procedures.

Let $\widehat{p}(\theta_0)$ be the p-value returned by Algorithm 2.1 for testing $H_0 : \theta = \theta_0$. The function $\widehat{p}(\cdot) : \Theta \to [0,1]$ is called a p-value function (Fraser (2019)). It measures the degree of falsity of $H_0 : \theta = \theta_0$. Similar concepts include confidence curves (Birnbaum (1961)), confidence distributions (Xie and Singh (2013), Xie, Singh and Strawderman (2011)), significance functions (Fraser (1991)), plausibility functions (Martin (2015)), etc. There are many automatic procedures that are built on the p-value function; see, for example, Martin (2017). We are not able to exhaust all applications. Only four examples are presented here.

First, one obvious application is confidence regions (CR) construction. By the duality of hypothesis testing and CR (Section 5.4 of Lehmann and Romano (2005)), a $100(1-\alpha_0)\%$ CR for θ is $\mathcal{C} = \{\theta_0 \in \Theta : \widehat{p}(\theta_0) \geq \alpha_0\}$, which can be obtained by repeatedly using Algorithm 2.1. Second, if researchers prefer not to fix the confidence level in advance, it is possible to report the p-value function as an estimator of θ; see Infanger and Schmidt-Trucksäss (2019) for some examples in medical studies. Third, p-value has been a commonly-used tool for combining evidence in meta-analysis (Heard and Rubin-Delanchy (2018)). For example, if the p-value for testing $H_0 : \theta = \theta_0$ by the gth dataset is $\widehat{p}_g(\theta_0)$ for $g = 1, \ldots G$, a possible combined p-value is $\sum_{g=1}^{G} \log \widehat{p}_g(\theta_0)$ (Fisher (1934)). Fourth, p-values can be used for stepwise variable selection in generalized linear model (Section 4.6.1 of Agresti (2015)).

In a nutshell, classical Rubin's rule use variance as a medium for assessing uncertainty, whereas our proposed method use p-value. Our proposal is useful not only for the standard null hypothesis testing but also for other statistical procedures that require variability assessment.

6. Monte Carlo experiments and applications. Incomplete-data testing of linear regression coefficient and region estimation of a probability vector are presented in Sections 6.1 and 6.2, respectively. Due to space constraints, some additional simulation experiments and real-data examples are deferred to the Supplementary Material. They include (i) inference of variance-covariance matrix in Section B.3, (ii) variable selection in generalized linear model in Section B.4, (iii) contingency tables in Sections B.5, and (iv) logistic regression in Section B.6.

6.1. *Linear regression.* Let y_i and $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ be a univariate random response and p deterministic covariates of the ith unit, respectively, where $i = 1, \ldots, n$. Consider the linear regression model: $y_i = \beta_0 + x_i^{\mathrm{T}}\beta + \epsilon_i$ for each i, where $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$. The full set of model parameters is $\psi = (\beta_0, \beta^{\mathrm{T}}, \sigma)^{\mathrm{T}}$; and the parameter of interest is $\theta = \beta$. We want to test $H_0 : \theta = 0_p$ against $H_1 : \theta \neq 0_p$. Clearly, in this case, $k = p$.

For each i, suppose y_i and x_{i1} are always observed, while $x_{i2}, \ldots, x_{ip}$ may be missing. Let $I_{ij} = 0$ if x_{ij} is missing, otherwise $I_{ij} = 1$. Suppose further that I_{ij} follows a logistic regression model:

$$\mathsf{P}(I_{ij} = 1 \mid x_{i,j-1}, I_{i,j-1} = a) = \operatorname{expit}(\gamma_0 + \gamma_1 x_{i,j-1})\mathbb{1}(a = 1) \tag{6.1}$$

for $i = 1, \ldots, n$ and $j = 2, \ldots, p$, where $\operatorname{expit}(t) := 1/(1 + e^{-t})$. The missing data $X_{\text{mis}} = \{x_{ij} : I_{ij} = 0\}$ are then imputed $m \in \{10, 30\}$ times by a Bayesian model; see Section B.2 of the Supplementary Material for details.

In the simulation experiment, all three devices d_{W}, d_{L} and d_{R} are studied; see Section B.2 for their formulas. We compute the MI test statistics $\widehat{D}_{\mathrm{W}}$, $\widehat{D}_{\mathrm{L}}$ and $\widehat{D}_{\mathrm{R}}$, and refer them to four different approximated null distributions in Table 1, that is, T1 (χ^2_k/k), T2 (Li, Raghunathan and Rubin (1991)), T3 (Meng (1990)), and T4 (the proposal in Algorithm 2.1). Note that the unknowns μ_r and σ_r^2 in T2 and T3 are estimated by (4.5) with the devices d_{W}, d_{L} and d_{R} for $\widehat{D}_{\mathrm{W}}$, $\widehat{D}_{\mathrm{L}}$ and $\widehat{D}_{\mathrm{R}}$, respectively.

We consider $n = 1000$, $p = 5$, $x_i \overset{\text{iid}}{\sim} \mathrm{N}_p(1_p, \Sigma_x)$ with $(\Sigma_x)_{ab} = 2^{-|a-b|}$ for each $1 \leq a, b \leq p$, $\sigma^2 = 1$ and $\beta_0 = 1$, where 1_p is a p-vector of ones. Two sets of $(\gamma_0, \gamma_1) \in \{(1, 0), (0, 1)\}$ are considered. Note that the data are missing completely at random (MCAR) when $\gamma_1 = 0$, whereas the data are missing at random (MAR) when $\gamma_1 = 1$. Note that the fractions of missing covariates are about (0, 16%, 29%, 41%, 50%) and (0, 24%, 40%, 51%, 61%) in the MCAR and MAR cases, respectively.

We record the sizes of the tests (denoted by α) at various nominal size $\alpha_0 \in [1\%, 5\%]$. The results for $m = 30$ are shown in Figure 2. The proposed test T4 controls the size substantially more accurately than all other competitors in all cases. Although the widely used T2 performs well when Condition 1 is true (see Li, Raghunathan and Rubin (1991)), it only has a marginal improvement over T1 when Condition 1 does not hold. Besides, our proposed test is a unified method for Wald's test, LR test, and RS test. So, users only need to change the testing device (i.e., d_{W}, d_{L}, d_{R}) in Algorithm 1. It is worth mentioning again that our proposal is the only MI procedure that can handle RS tests.

The results for $m = 10$ are deferred to Figure 5 of the Supplementary Material. The patterns are similar to Figure 2 except for T3. Note that T3 is not trustworthy unless m is large. In this example, its performance converges to an increasingly bad state as m increases. The powers of the MI tests are not directly comparable as their sizes are not equally accurate. Their size-adjusted powers are identical because they are based on the same test statistic $\widehat{D}$ (or its asymptotic equivalent). For reference, the power curves are shown in Section B.2 of the Supplementary Material. Additional discussions about the effects of m are deferred to Section B.2.

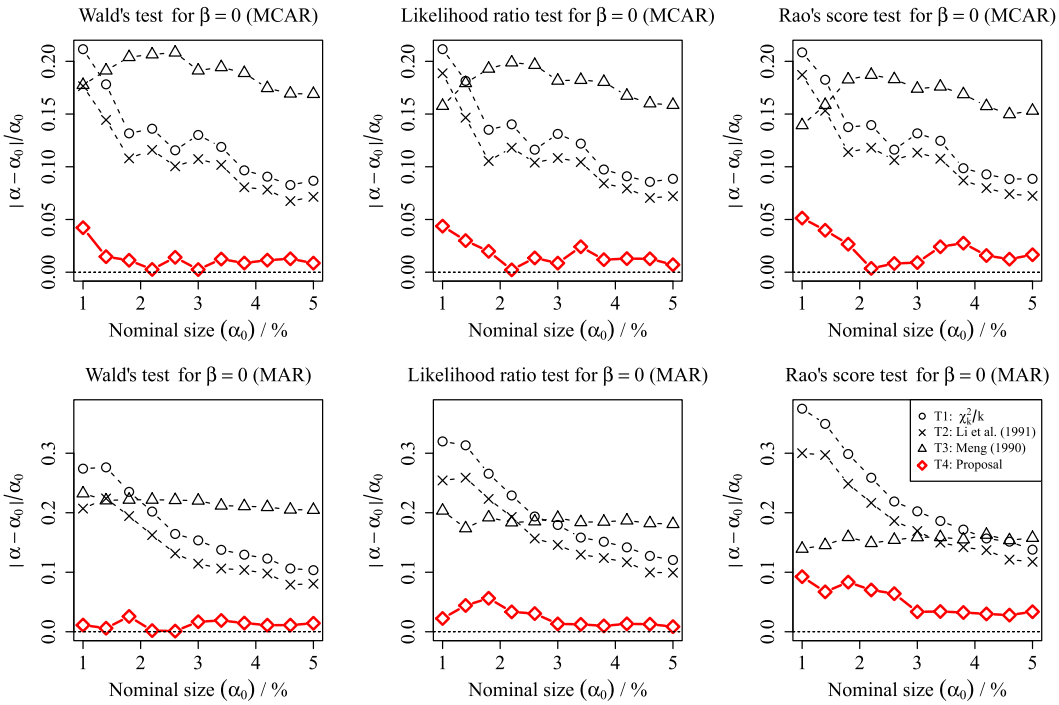

FIG. 2. *Size accuracy of the MI tests regarding regression coefficients in Section 6.1.*

6.2. *Confidence region and p-value function of a probability vector.* Let $[y_i \mid x_i] \sim \text{Bern}(\theta_1^{x_i}\theta_0^{1-x_i})$ independently for $i = 1, \ldots, n$, where $x_1, \ldots, x_n \in \{0, 1\}$ are fixed binary covariates, and $\theta_0, \theta_1 \in (0, 1)$ are unknown parameters. The covariates $x_1, \ldots, x_n$ are always observed, but the responses $y_1, \ldots, y_n$ may be missing. If y_i is observed, we denote $I_i = 1$, otherwise $I_i = 0$. Suppose that the missing mechanism is $[I_i \mid x_i] \sim \text{Bern}(\pi_1^{x_i}\pi_0^{1-x_i})$, where $\pi_0, \pi_1 \in (0, 1)$ are unknown. The goal is to estimate $\theta = (\theta_0, \theta_1)^{\mathrm{T}}$ with the incomplete dataset $\{(x_i, y_i I_i)\}_{i=1}^n$. In the simulation study, $n = 100$ and around 40% of the x_i's are 1. The unknown true values are $\theta_0 = 0.15$, $\theta_1 = 0.75$, $\pi_0 = 0.9$ and $\pi_1 = 0.1$. Note that y_i is very likely to be missing when $x_i = 1$. As discussed in Section 5.2, we could construct a CR or p-value function as an estimator of θ. We try both in this example.

Traditionally, one may construct the Wald's CR by Rubin's rule:

$$\mathcal{C}_{\mathrm{W}} = \left\{\theta \in \mathbb{R}^2 : (\bar{\theta} - \theta)^{\mathrm{T}}\left(\widehat{B} + \frac{m+1}{m}\bar{V}\right)^{-1}(\bar{\theta} - \theta) \le c\right\},$$

where the critical value c can be found according to Li, Raghunathan and Rubin (1991). It is well-known that the Wald's CR $\mathcal{C}_{\mathrm{W}}$ must be an ellipse, which is restrictive in some problems. Moreover, $\mathcal{C}_{\mathrm{W}}$ may not be a subset of the support $\Theta = (0, 1)^2$. Figure 3(a) visualizes these two problems. Alternatively, one may invert the LR test to construct a CR $\mathcal{C}_{\mathrm{L}}$ (say). It does not have the two aforementioned structural problems; see Figure 3(a) again.

We assess the performance of the likelihood-based $100(1 - \alpha_0)\%$ CRs for θ by using T1–T4. We compute the actual noncoverage rates $\widehat{\alpha}$ for different methods, and compare them with the nominal value α_0. The relative error $|\widehat{\alpha} - \alpha_0|/\alpha_0$ is reported for each method in Figure 3(b) under various $\alpha_0 \in [0.01, 0.05]$. The relative error of the Wald's CR is also computed for reference. Our proposed method has the lowest error uniformly. Figure 3(c) shows the p-value function $\widehat{p}(\theta)$ produced by our proposed T4 and Algorithm 2.1. It offers an alternative way for estimating θ with variability assessment but without specifying the confidence level

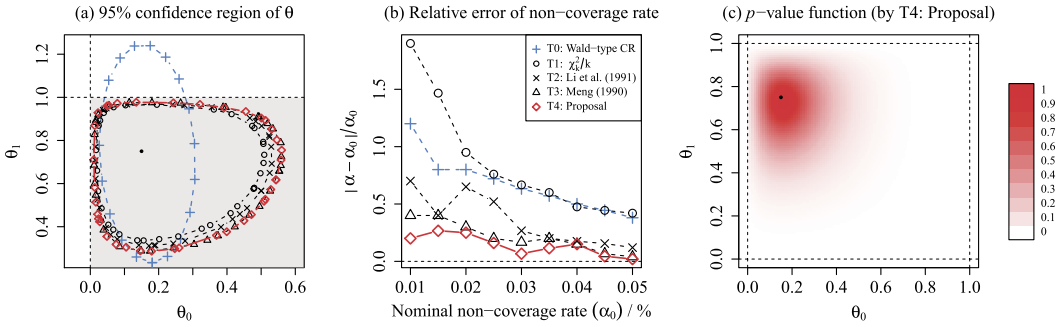

FIG. 3. (*a*) *One typical realization of* 95% *CRs of* θ *using different methods, where the grey region is the support* $(0, 1)^2$ *of* θ. (*b*) *Coverage accuracy of different CRs*. (*c*) *Heat-map of the* p*-value function by using the proposed T*4. *The solid black dots in plots* (*a*) *and* (*c*) *denote the true value of* $\theta = (0.15, 0.75)^{\mathrm{T}}$. *See Section* 6.2 *for detailed descriptions*.

in advance. From the above example, our proposed method is particularly useful for handling nonnormal data and parameters with bounded supports.

7. Conclusion, discussion and future work. The proposed test for handling multiply-imputed datasets is general as it does not require $m \to \infty$ or equal odds of missing information. So, it is particularly suitable for handling public-use datasets that have nontrivial missingness structures; see Remark 1.1 and Section 6 for some examples. The test is feasible in the sense that only a standard complete-data testing device is needed for performing incomplete-data Wald's, likelihood ratio, and Rao's score tests; see Algorithm 2.1. Besides, the proposed method is also useful for general statistical procedures that use p-value as an assessment tool; see Section 5.2. So, it has a wide range of applications. Although the proposed test improves the existing counterparts, further studies are needed in the following directions.

First, this paper assumes the sample size n is large enough so that the standard large-sample χ^2 approximation kicks in. Small-sample approximations (e.g., approaches similar to Barnard and Rubin (1999) and Reiter (2007)) are likely to further improve the proposed test. Second, we need to perform more computationally expensive tests on several stacked datasets. We increase the computing cost in order to minimize the human time cost needed to build nonstandard computing functions required in tests T2 and T3; see Table 1. Given the computing power of the current computers, we believe that computing time cost is a lesser constraint than human time cost. However, it is still desirable to further reduce the computing cost.

Acknowledgments. A part of the theoretical results in this article are partially developed from the author's Ph.D. thesis under the supervision of Xiao-Li Meng, who provided many insightful ideas that greatly contribute to this paper. The author would also like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the scope and presentation of the paper.

Funding. The author acknowledges the financial support from the Early Career Scheme (24306919) provided by the University Grant Committee of Hong Kong.

SUPPLEMENTARY MATERIAL

Supplement to "General and feasible tests with multiply-imputed datasets" (DOI: 10.1214/21-AOS2132SUPP; .pdf). The supplementary note includes proofs, supplementary results and additional examples. An R-package `stackedMI` is also provided.

REFERENCES

AGRESTI, A. (2015). *Foundations of Linear and Generalized Linear Models*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3308143

BARNARD, J. and RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86** 948–955. MR1741991 https://doi.org/10.1093/biomet/86.4.948

BERA, A. K. and BILIAS, Y. (2001). Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: An essay on historical developments and some new results. *J. Statist. Plann. Inference* **97** 9–44.

BIRNBAUM, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *J. Amer. Statist. Assoc.* **56** 246–249. MR0121904

CARPENTER, J. R. and KENWARD, M. G. (2013). *Multiple Imputation and Its Application*. Wiley, New York.

CHAN, K. W. (2022). Supplement to "General and feasible tests with multiply-imputed datasets." https://doi.org/10.1214/21-AOS2132SUPP

CHAN, K. W. and MENG, X.-L. (2021+). Multiple improvements of multiple imputation likelihood ratio tests. *Statist. Sinica* To appear.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

FISHER, R. A. (1934). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh. 5 edn.

FRASER, D. A. S. (1991). Statistical inference: Likelihood to significance. *J. Amer. Statist. Assoc.* **86** 258–265. MR1137116

FRASER, D. A. S. (2019). The p-value function and statistical inference. *Amer. Statist.* **73** 135–147. MR3925719 https://doi.org/10.1080/00031305.2018.1556735

HAREL, O. and ZHOU, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Stat. Med.* **26** 3057–3077. MR2380504 https://doi.org/10.1002/sim.2787

HEARD, N. A. and RUBIN-DELANCHY, P. (2018). Choosing between methods of combining p-values. *Biometrika* **105** 239–246. MR3768879 https://doi.org/10.1093/biomet/asx076

HOLAN, S. H., TOTH, D., FERREIRA, M. A. R. and KARR, A. F. (2010). Bayesian multiscale multiple imputation with implications for data confidentiality. *J. Amer. Statist. Assoc.* **105** 564–577. MR2759932 https://doi.org/10.1198/jasa.2009.ap08629

HORTON, N. J. and KLEINMAN, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Amer. Statist.* **61** 79–90. MR2339152 https://doi.org/10.1198/000313007X172556

INFANGER, D. and SCHMIDT-TRUCKSÄSS, A. (2019). P value functions: An underused method to present research results and to promote quantitative reasoning. *Stat. Med.* **38** 4189–4197. MR3999273 https://doi.org/10.1002/sim.8293

KENWARD, M. G. and CARPENTER, J. (2007). Multiple imputation: Current perspectives. *Stat. Methods Med. Res.* **16** 199–218. MR2371006 https://doi.org/10.1177/0962280206075304

KIM, J. K. and SHAO, J. (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press/CRC, Boca Raton, FL.

KIM, J. K. and YANG, S. (2017). A note on multiple imputation under complex sampling. *Biometrika* **104** 221–228. MR3626470 https://doi.org/10.1093/biomet/asw058

KING, G., HONAKER, J., JOSEPH, A. and SCHEVE, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am. Polit. Sci. Rev.* **95** 49–69.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

LI, K. H., RAGHUNATHAN, T. E. and RUBIN, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *J. Amer. Statist. Assoc.* **86** 1065–1073. MR1146352

MARTIN, R. (2015). Plausibility functions and exact frequentist inference. *J. Amer. Statist. Assoc.* **110** 1552–1561. MR3449054 https://doi.org/10.1080/01621459.2014.983232

MARTIN, R. (2017). A statistical inference course based on p-values. *Amer. Statist.* **71** 128–136. MR3668699 https://doi.org/10.1080/00031305.2016.1208629

MENG, X.-L. (1990). *Towards Complete Results for Some Incomplete-Data Problems*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Harvard University. MR2685717

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9** 538–573.

MENG, X.-L. and RUBIN, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79** 103–111. MR1158520 https://doi.org/10.1093/biomet/79.1.103

PARKER, J. D. and SCHENKER, N. (2007). Multiple imputation for national public-use datasets and its possible application for gestational age in United States Natality files. *Paediatr. Perinat. Epidemiol.* **21** 97–105. https://doi.org/10.1111/j.1365-3016.2007.00866.x

PEUGH, J. L. and ENDERS, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Rev. Educ. Res.* **74** 525–556.

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. *Springer Series in Statistics*. Springer, New York. MR1707286 https://doi.org/10.1007/978-1-4612-1554-7

QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360. MR0081040 https://doi.org/10.1093/biomet/43.3-4.353

RAO, C. R. (2005). Score test: Historical review and recent developments. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*. *Stat. Ind. Technol.* 3–20. Birkhäuser, Boston, MA. MR2111500 https://doi.org/10.1007/0-8176-4422-9_1

REITER, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests for multiple imputation for missing data. *Biometrika* **94** 502–508. MR2380575 https://doi.org/10.1093/biomet/asm028

ROSE, R. A. and FRASER, M. W. (2008). A simplified framework for using multiple imputation in social work research. *Soc. Work Res.* **32** 171–178.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

RUBIN, D. B. (1978). Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc.* 20–34.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

RUBIN, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.

RUBIN, D. B. and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Amer. Statist. Assoc.* **81** 366–374. MR0845877

SCHAFER, J. L. (1999). Multiple imputation: A primer. *Stat. Methods Med. Res.* **8** 3–15. https://doi.org/10.1177/096228029900800102

SCHENKER, N., RAGHUNATHAN, T. E., CHIU, P.-L., MAKUC, D. M., ZHANG, G. and COHEN, A. J. (2006). Multiple imputation of missing income data in the National Health interview survey. *J. Amer. Statist. Assoc.* **101** 924–933. MR2324093 https://doi.org/10.1198/016214505000001375

SERFLING, R. J. (2001). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SHAO, J. (1999). *Mathematical Statistics*. *Springer Texts in Statistics*. Springer, New York. MR1670883

TU, X. M., MENG, X.-L. and PAGANO, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *J. Amer. Statist. Assoc.* **88** 26–36.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge.

WANG, N. and ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948. MR1666715 https://doi.org/10.1093/biomet/85.4.935

XIE, X. and MENG, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? (with discussion). *Statist. Sinica* **27** 1485–1594.

XIE, M. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Stat. Rev.* **81** 3–39. MR3047496 https://doi.org/10.1111/insr.12000

XIE, M., SINGH, K. and STRAWDERMAN, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *J. Amer. Statist. Assoc.* **106** 320–333. MR2816724 https://doi.org/10.1198/jasa.2011.tm09803

YU, P. S. Y., CHAN, K. W., LAU, R. W. H., WAN, I. Y. P., CHEN, G. G. and NG, C. S. H. (2021). Uniportal video-assisted thoracic surgery for major lung resection is associated with less immunochemokine disturbances than multiportal approach. *Sci. Rep.* **11** 10369.