

The Smallest Standard Error Is the Wrong One: Selection Among Multiple Imputations Invalidates Inference

Nyiko Muhluri Khoza¹

University of South Africa

Bhekisipho Twala²

`twala.bhekisipho@mut.ac.za`

Mangosuthu Technikon: Mangosuthu University of Technology

Jean-Claude Malela-Majika³

University of Pretoria

Research Article

Keywords: missing data, multiple imputation, Rubin's rules, variance estimation, selection bias, confidence interval coverage, simulation design, evaluation methodology

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10779023/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

The Smallest Standard Error Is the Wrong One: Selection Among Multiple Imputations Invalidates Inference

Nyiko Muhluri Khoza¹, Bhekisipho Twala², Jean-Claude Malela-Majika³

¹ Department of Statistics, University of South Africa, South Africa — ngobenh@unisa.ac.za

² Office of the Deputy Vice-Chancellor (Resources and Planning), Mangosuthu University of Technology, Durban 4360, South Africa — twala.bhekisipho@mut.ac.za — ORCID 0000-0002-3452-9581 (*corresponding author)

³ Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, South Africa — malela.mjc@up.ac.za

Abstract

Selecting, from M multiply imputed analyses, the candidate with the smallest estimated standard error looks like a precision-maximising choice. We show that it is systematically invalid. Minimum-standard-error selection discards the between-imputation variance that multiple imputation exists to capture and, by taking an order statistic, further depresses the reported within-imputation variance. The resulting standard error therefore does not estimate the sampling variability of the selected estimator. We formalise both mechanisms and show that selection can also bias the point estimate whenever candidate estimates and their estimated standard errors are dependent. A newly frozen, fully specified Monte Carlo implementation quantifies the effect. Under MAR with 30% missingness, $n = 400$ and $M = 20$, minimum-standard-error selection produced a calibration ratio of 0.77 and 84.5% coverage for nominal 95% intervals. Rubin pooling of the identical candidate draws restored the calibration ratio to 1.00 and coverage to 94.6%. In a 5,000-replication sweep, selection coverage declined monotonically from 86.7% at $M = 2$ to 81.7% at $M = 100$ while the magnitude of selection bias increased from 0.021 to 0.079; Rubin pooling remained approximately calibrated throughout. Two conclusions follow. First, an uncertainty estimate cannot simultaneously serve as an optimisation criterion and retain its usual inferential interpretation. Second, compare imputation procedures under a calibration-first protocol: establish that reported uncertainty tracks actual sampling variability before ranking methods by efficiency. We provide that protocol, a selection-corrected fallback, and a reproducibility archive from which we generate all revised numerical results.

Keywords: missing data; multiple imputation; Rubin's rules; variance estimation; selection bias; confidence interval coverage; simulation design; evaluation methodology

1. Introduction

Incomplete data are ubiquitous across the empirical disciplines, and the choice among imputation methods is consequential for the conclusions that follow (Little & Rubin, 2002; Schafer & Graham, 2002; Enders, 2010). Researchers typically make that choice by consulting a benchmark study that compares a set of methods on some numerical criterion and declares a winner. The criterion therefore does a great deal of work, and it is worth asking what a defensible one must satisfy — a question that has received markedly less attention than the methods themselves.

A natural candidate criterion is the standard error. It determines confidence-interval width and the power of hypothesis tests (Casella & Berger, 2002); it is what an analyst actually reports; and, unlike the mean squared error, it can be computed without knowing the truth, so it is available in both applications and simulations. These are genuine advantages, and they motivate a straightforward procedure: generate

several candidate imputations, compute the estimate and its standard error for each, and retain the one with the smallest standard error. The result appears to be the most precise analysis the data support.

This paper shows that the procedure is invalid, and that the invalidity is not a matter of degree. Two distinct mechanisms operate.

First, the selected dataset is treated thereafter as though it had been observed, so the variability across candidate imputations — the between-imputation variance that Rubin's (1987) rules exist to capture — is discarded entirely, and the procedure inherits the anticonservatism of single imputation (Rubin & Schenker, 1986; Schafer, 1999).

Second, taking the minimum of M estimated standard errors introduces a further downward distortion that has nothing to do with the data and everything to do with the act of minimisation: the minimum of M draws from any non-degenerate distribution is stochastically smaller than a single draw, and increasingly so as M grows (David & Nagaraja, 2003); the two effects compound.

A third consequence follows once one notices that a candidate estimate and its estimated standard error are generally dependent. Selecting on the latter is then a non-random selection on the former, and the point estimate acquires a bias that the analyst has no way of detecting from the output. This is the same structural problem that Heckman (1979) and Berk (1983) identified for selected samples, arising here not from the data-generating process but from the analyst's own procedure.

The practical upshot is that such a procedure will always appear to win a standard-error comparison, because it optimises the criterion by which it is judged. Its apparent superiority over multiple imputation is an artefact of the evaluation design, not a property of the method. This is why a standard error cannot be both the objective function and the scoreboard, and why the failure is invisible to conventionally constructed benchmarks.

We should be explicit about the argument's provenance. The procedure we analyse — which we call RDOI, for random-draw observations imputation, and RDCOI in its categorical form — is one we ourselves proposed and defended in an earlier version of this work, on the strength of exactly the standard-error comparison we now show to be uninformative. We report the negative result because the procedure is natural and others will independently invent it. After all, the diagnostic that exposes it is inexpensive and general, and the methodological literature benefits more from documented failures than from another incremental comparison (Ioannidis, 2005).

The paper is organised as follows. Section 2 fixes notation, reviews the imputation methods used as comparators, and sets out the conditions any benchmark criterion must satisfy. Section 3 gives the formal account: two mechanisms that understate the reported variance, a third that displaces the point estimate, and the consequence of all three for coverage. Section 4 specifies a calibration-first evaluation protocol built on the calibration ratio — the reported standard error divided by the estimator's true sampling standard deviation — which serves as an admissibility gate that a method must pass before its efficiency is assessed at all. Section 5 gives the corrections: RDOI-R, which pools the candidate draws by Rubin's rules, and RDOI-SC, a selection-retaining variant with an inflated variance estimator for workflows that genuinely require a single completed dataset. Section 6 quantifies all of this in a Monte Carlo study with a known estimand, and Section 7 adapts the framework to observational data, where no true parameter exists. Sections 8 and 9 consider how far the failure mode generalises and conclude.

2. Background

This section fixes notation, reviews the imputation methods against which the procedure under study will be compared, and sets out the requirements that any benchmark criterion must satisfy. The last receives more space than is conventional because the paper's argument turns on it.

2.1 Mechanisms and notation

Let Y denote the variable subject to missingness, X the fully observed covariates, and R the response indicator taking the value one when Y is observed. Rubin's (1976) taxonomy distinguishes three regimes. Under missing completely at random, $P(R | Y_{obs}, Y_{mis}) = P(R)$, the observed cases are a random subsample of the whole. Under missing at random, the conditioning reduces to the observed data alone; under missing not at random, it does not, so the missingness carries information about the values it conceals.

The taxonomy matters because it determines which procedures are consistent. Complete-case analysis is unbiased under MCAR but generally not under MAR; likelihood-based and multiple-imputation procedures are consistent under MAR provided the imputation model is correctly specified; and MNAR requires explicit modelling of the missingness mechanism, with conclusions that are sensitive to assumptions which the data cannot check (Molenberghs & Kenward, 2007; Molenberghs et al., 2014; Carpenter & Smuk, 2021).

The practical consequences of these distinctions for applied research have been set out at length for the clinical and epidemiological literatures (Sterne et al., 2009; Jakobsen et al., 2017). Because the mechanism governs so much, we vary it across all three regimes in the simulation of Section 6 rather than confining attention to MAR.

One point deserves emphasis at the outset, because it bears directly on the argument of this paper. The mechanism determines whether an estimator is biased; it does not determine whether that estimator's reported uncertainty is honest. These failures are separable, and a procedure can be consistent under MAR while reporting standard errors unrelated to its actual sampling variability. Much of the missing-data literature has concentrated on the first question. The failure we document is entirely of the second kind, and it persists even under MCAR, where the mechanism itself poses no difficulty.

2.2 Imputation methods and the propriety requirement

Single imputation replaces each missing entry with one value and proceeds as though the completed dataset had been observed. The simplest variants substitute a measure of central tendency; hot-deck methods borrow values from similar observed cases (Little, 1988; Andridge & Little, 2010); nearest-neighbour methods select donors by a distance metric (Troyanskaya et al., 2001); and regression imputation predicts from a fitted conditional model. More recent machine-learning variants such as random-forest imputation extend the same idea to non-linear and mixed-type structures (Stekhoven & Bühlmann, 2012); tree-based treatments of missingness have been studied extensively in the classification setting (Twala, Jones & Hand, 2008; Twala, 2009); and ensemble strategies combining several imputation approaches have been proposed to improve robustness across data conditions (Twala & Cartwright, 2010).

The common defect of all single imputation is not inaccuracy but overconfidence. Because the imputed values are treated as data, the analysis cannot detect that they were estimated, and the resulting standard errors are too small, confidence intervals too narrow, and test statistics too large (Rubin & Schenker, 1986; Schafer, 1999; Graham, 2009). Deterministic variants compound the problem by suppressing residual variability, which additionally distorts the marginal distribution and attenuates associations among variables.

Multiple imputation resolves this by generating M completed datasets, analysing each separately, and combining the results by rules that add a between-imputation component to the average within-imputation variance (Rubin, 1987, 1996). Validity requires the imputations to be proper — drawn in a way that propagates uncertainty about the imputation model's own parameters, not merely residual noise — a requirement made precise by Rubin (1987) and examined critically by Nielsen (2003), Wang and Robins (1998) and Robins and Wang (2000). Chained-equations implementations are now the practical standard (van Buuren & Groothuis-Oudshoorn, 2011; White, Royston & Wood, 2011; Carpenter & Kenward, 2013; van Buuren, 2018), with the expectation-maximisation algorithm (Dempster, Laird & Rubin, 1977) and bootstrap-based alternatives (Efron, 1994; Bartlett & Hughes, 2020) offering related routes. The procedure examined in this paper generates proper draws and then discards the machinery that makes them useful.

2.3 What a benchmark criterion must do

An evaluation criterion ranks procedures, and for that ranking to be meaningful, the criterion must satisfy conditions that are rarely stated explicitly. We set them out here because this paper argues that a standard error used as a selection objective violates all three. This framing follows the broader simulation-design literature, which has repeatedly stressed that the choice of performance measure is a substantive methodological decision rather than a reporting convention (Burton et al., 2006; Morris, White & Crowther, 2019).

- The estimand must be fixed in advance. Bias, variance and coverage are properties of an estimator of a specified target parameter θ . A criterion computed without naming θ is not well defined. This is not a pedantic point: the accuracy with which individual missing values are recovered and the accuracy of an estimator computed from the completed data are different quantities, and a method can perform well on one while performing badly on the other.
- The criterion must not be under the control of the procedure being evaluated. If a method optimises the quantity on which it is scored, its rank reflects the optimisation rather than its statistical properties, and the comparison ceases to be informative about anything except the optimisation.
- The criterion must penalise dishonest uncertainty. A procedure can always make its reported standard error smaller by understating it. Any criterion that rewards small reported standard errors without checking them against the truth rewards precisely this behaviour, and will systematically favour the least trustworthy methods.

These three conditions are not independent, and the first underwrites the other two. Once the estimand is named, we can ask whether a criterion can be optimised without moving the estimator closer to it, which is the second condition. This makes it possible to check a reported uncertainty against the estimator's actual sampling behaviour (the third condition). A benchmark that leaves θ implicit forfeits

both checks simultaneously, and it is no accident that criteria which violate the second and third conditions tend to be found in studies that violate the first.

The third condition is the one most often neglected, and it is where the connection to the wider replication literature lies. Selecting an analysis because it produced a favourable uncertainty estimate is formally the same operation as selecting an analysis because it produced a favourable p-value, and it invalidates the reported quantity for the same reason (Simmons, Nelson & Simonsohn, 2011; Gelman & Loken, 2014). What distinguishes the imputation setting is that the selection is presented as a methodological refinement rather than as questionable research practice, and is therefore performed openly and documented in the method description.

2.4 Why the standard error alone cannot rank methods

The mean squared error, $E[(\hat{\theta} - \theta)^2] = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$, evaluates the estimator's deviation from its target and decomposes into two interpretable components (Hastie, Tibshirani & Friedman, 2009). A standard error does neither. It estimates the square root of one component and carries no information about the other. Two estimators with identical standard errors may differ arbitrarily in accuracy, and the criterion cannot distinguish them.

Deterministic imputation of the conditional mean illustrates the gap starkly. Because it suppresses residual variability, it reduces the apparent dispersion of the completed data and so shrinks the computed standard error, while leaving the estimator no closer to θ . In the simulation of Section 6, mean imputation attains the smallest reported standard error of any method considered (0.104), along with the worst coverage in the study (23.3%). The divergence widens as the problem worsens: at 60% missingness its coverage falls to 0.7%, and under MNAR to 0.1%. A criterion that ranks on the reported standard error ranks it first.

In this sense, our earlier characterisation of the mean squared error as inferior to the standard error was mistaken, and we withdraw it without qualification. The argument had appeared plausible only because no estimand had been declared, so that the accuracy of value recovery and the accuracy of an estimator could be conflated. Once θ is named, the mean squared error is manifestly the more informative quantity, and the standard error is revealed as one ingredient of it rather than as a competitor to it. What the standard error does uniquely offer — availability without knowledge of the truth — makes it valuable in application and useless as a benchmark ranking criterion, since a benchmark by construction knows the truth.

3. The selection artefact

We now give the formal account of why minimum-standard-error selection fails. Three mechanisms operate, and they are logically independent: each would damage inference on its own, and correcting any one of them leaves the others intact. We treat them in turn and then combine them to obtain the consequence for coverage.

3.1 Setup and notation

Let θ be a scalar and $D^{(1)}, \dots, D^{(M)}$ be M completed datasets produced by proper draws from the posterior predictive distribution of the missing data. Write $\hat{\theta}_m$ for the estimate obtained from the

m th dataset and U_m for its within-imputation variance estimate, so that the standard error reported for θ alone is $\sqrt{U_m}$. Under Rubin (1987), valid inference combines them as

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m, \quad \bar{U} = \frac{1}{M} \sum_{m=1}^M U_m, \quad \text{and} \quad \bar{B} = \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_m - \bar{\theta})^2.$$

The total variance is then:

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) \bar{B}$$

Moreover, the corresponding multiple imputation standard error is:

$$SE_{MI} = \sqrt{\bar{U} + \left(1 + \frac{1}{M}\right) \bar{B}}$$

where $\bar{U}$ represents within-imputation uncertainty, while $\bar{B}$ represents the between-imputation uncertainty arising from the missing data.

Inference proceeds by referring $(\hat{\theta} - \theta)/\sqrt{T}$ to a t distribution with degrees of freedom given by Rubin (1987) or, in finite samples, by the refinement of Barnard and Rubin (1999), which we use throughout. The between-imputation term $\bar{B}$ is the formal expression of the fact that the missing data were imputed rather than observed; the factor $(1 + \frac{1}{M})$ corrects for having used finitely many draws to estimate it. Guidance on choosing M has evolved from the small values Rubin (1987) considered adequate towards substantially larger ones (Graham, Olchowski & Gilreath, 2007; von Hippel, 2020), because larger M stabilises the variance estimate and improves power — a point to which we return, because under the procedure examined here the effect of increasing M is precisely reversed.

The RDOI procedure differs at the combination step. It sets $m^* = \operatorname{argmin}_m U_m$, reports $\hat{\theta}_{m^*}$ as the point estimate and U_{m^*} as its variance, and forms the interval $\hat{\theta}_{m^*} \pm z\sqrt{U_{m^*}}$. The generation of draws is therefore identical to proper multiple imputation, and everything that follows concerns what is done with them. This is what makes the comparison in Section 6 an isolation of the selection step rather than a comparison of two imputation models.

3.2 Mechanism I: the between-imputation component is discarded

Because a single completed dataset is carried forward and analysed as though observed, the term $\bar{B}$ never enters the reported variance. The deficit is exactly $(1 + 1/M)\bar{B}$, which is strictly positive whenever the imputation model is non-degenerate, and which grows with the missing rate and with the fraction of information about θ carried by the missing values. Since by construction the reported variance is bounded above by the correct one, understatement $U_m^* \leq \bar{U} \leq T$ is guaranteed, not merely likely. This is a different and far more severe failure than the finite-sample bias of Rubin's variance estimator itself, which has been examined in the survey-sampling context and is typically modest (Kim, Brick, Fuller & Kalton, 2006); here the component is not mis-estimated but omitted altogether.

This identifies the procedure as a data-dependent single imputation, and it inherits the documented anticonservatism of that class (Rubin & Schenker, 1986; Schafer, 1999; Graham, 2009). The data-dependence is what makes the diagnosis non-obvious. Because M draws were generated, the procedure has the outward appearance of multiple imputation, and its method description can accurately state that

it used multiple imputations. What it does not do is combine them, and it is the combination rather than the generation that supplies the validity.

The magnitude of the deficit is not a fixed constant but scales with the quantity multiple imputation was designed to represent. In settings where imputation matters least — low missing rates, strongly predictive covariates — B is small, and the understatement is correspondingly mild. In settings where imputation matters most, B is large, and the understatement is severe. The procedure therefore fails hardest under exactly the conditions that motivate using it, a pattern visible in the missing-rate sweep of Table 3.

3.3 Mechanism II: minimisation biases the variance estimate downwards

The second mechanism is logically independent of the first and would operate even if B were somehow restored. For any non-degenerate collection $U_1, \dots, U_M$, elementary order-statistic theory gives $E[\min_m U_m] < E[U_1]$, with the gap increasing in M and, for independent draws, $E[\min_m U_m] \rightarrow \text{ess inf } U$ as $M \rightarrow \infty$ (David & Nagaraja (2003). Arnold, Balakrishnan & Nagaraja, 2008).

Selecting the smallest of M variance estimates therefore returns a quantity that is not an unbiased estimate of the within-imputation variance of anything. The distortion is a deterministic function of how many candidates were examined, which produces an operational absurdity: two analysts applying the same procedure to the same data with different M will report different uncertainties for the same estimand, and the one who worked harder will report the smaller. Nothing in the output distinguishes the two analyses, and no convention exists to fix M at a value that makes the reported quantity interpretable.

This inverts the received guidance on M . In proper multiple imputation, increasing M reduces simulation error in the variance estimate and increases power, which is why recent recommendations favour larger values than were once thought sufficient (Graham, Olchowski & Gilreath, 2007; von Hippel, 2020). Under minimum-standard-error selection, the same increase drives the reported variance further below the truth. An analyst who follows current best practice on M , believing it to make the analysis more careful, thereby makes the inference worse — a prediction we test directly in Section 6.3.

3.4 Mechanism III: selection imports bias into the point estimate

Suppose $\hat{\theta}_m$ and U_m are dependent across draws — as they necessarily are for the sample mean, since both are functions of the same imputed values, and as they generally are for skewed, heteroscedastic or bounded outcomes. Then conditioning on the event that U_m attained its minimum constitutes a non-random selection on $\hat{\theta}_m$, and

$$E[\hat{\theta}_{m^*}] \neq E[\hat{\theta}_m] = \theta,$$

so the point estimate is biased even when every individual candidate entering the selection is unbiased. This is the classical structure of selection bias (Heckman, 1979; Berk, 1983), arising here not from the data-generating process but from the analyst's combination rule, and is therefore entirely avoidable. The direction of the bias depends on the sign of the dependence between the estimate and the variance estimate, and its magnitude on the strength of that dependence and on M .

The practical difficulty is that neither quantity is knowable from the output. The analyst observes a single point estimate and a single standard error, both of which look unremarkable; nothing in the completed dataset records that it was chosen rather than drawn. This is the mechanism by which, in Section 6.3, a procedure assembled entirely from unbiased draws produces an estimate whose bias grows steadily with the number of draws considered, and it is the component of the failure that we had not anticipated when the procedure was first proposed.

3.5 Consequence for coverage

Combining the three mechanisms, the confidence interval is centred at a point displaced from θ by Mechanism III. It has a half-width computed from a variance understated on the two separate counts of Mechanisms I and II. Coverage must therefore fall below nominal, and — because both the displacement and the understatement worsen in M — must fall further as M increases, up to Monte Carlo error. Since the interval is constructed by the usual Neyman (1937) device, its nominal level retains its conventional meaning; the inputs no longer support it.

This yields an unusual and diagnostically useful signature. For any valid procedure, increasing M can only improve inference: the point estimate stabilises, the variance estimate becomes less noisy, and coverage approaches nominal from wherever it began. A decline in coverage with M is therefore not merely evidence of poor performance but evidence of a selection step, and the same design can be run at several values of M at negligible cost. The check is sharpest when the candidate sets are nested, so that every value of M is evaluated on the same replications: the comparison across M is then paired, the sampling noise that would otherwise obscure a modest trend largely cancels, and the decline is recovered cleanly. In our sweep, it falls monotonically at every step, by about five percentage points in total across M from 2 to 100.

We stress that the three mechanisms are separable in principle and can be corrected separately. Restoring B addresses Mechanism I; abandoning the minimum in favour of an average addresses Mechanism II; and abandoning selection altogether addresses Mechanism III. Rubin pooling accomplishes all three simultaneously, which is why we recommend it rather than any partial repair, and why the selection-corrected variant of Section 5.2 is offered only for workflows in which a single completed dataset is a hard requirement.

4. A calibration-first evaluation protocol

The failure documented above is not specific to RDOI. It will afflict any benchmark that ranks methods on a reported uncertainty without checking whether that uncertainty is honest, and any workflow that selects among analyses on such a quantity. We therefore specify a protocol that makes the check mandatory and, crucially, prior to everything else. It is designed as a small addition to the reporting practice already recommended for simulation studies (Burton et al., 2006; Morris, White & Crowther, 2019), not a replacement.

4.1 Declare the estimand

The protocol begins by naming θ — a population mean, a regression coefficient, a difference in proportions, a quantile — together with the estimator $\hat{\theta}$ applied to the completed data and the analysis model that produces it. Every subsequent quantity is a property of that pair, and none of them is well

defined without it. This requirement is not new; it is standard in the simulation-design literature, and its routine omission from imputation benchmarks is the anomaly.

In particular, the protocol does not report how accurately a method recovers individual missing values. That is a distinct and generally less relevant property: an imputation method can reconstruct individual values poorly while supporting excellent inference about θ , and can reconstruct them well while supporting poor inference, since the two objectives reward different behaviour with respect to residual variability. Reporting value-recovery accuracy as though it were estimator bias, as our earlier version did, conflates quantities that can move in opposite directions.

Declaring the estimand also disciplines the choice of analysis model, which must be congenial with the imputation model if the pooled inference is to have its stated properties (Meng, 1994; Xie & Meng, 2017). Uncongeniality is a genuine and unresolved difficulty in multiple-imputation practice, and we do not address it here; our simulation uses a correctly specified and congenial pair so the selection artefact is isolated. A protocol that leaves the estimand implicit cannot even pose the congeniality question.

4.2 The calibration ratio

Let SD_{emp} denote the standard deviation of $\hat{\theta}$ across independent replications of the data-generating process — the true sampling variability of the estimator — and let $E[\widehat{SE}]$ denote the mean of the standard error the method reports. Define the calibration ratio:

$$\rho = E[\widehat{SE}] / SD_{emp}$$

A method with ρ close to unity reports its own uncertainty honestly. A method with ρ below unity understates it, and its confidence intervals will under-cover regardless of how accurate its point estimates are. A method with ρ above unity is conservative: its inference remains valid, but it is inefficient, and it will lose power relative to a correctly calibrated competitor. The quantity is closely related to the ratio of estimated to empirical standard error discussed by Morris, White and Crowther (2019), and we adopt it here as the organising diagnostic rather than as one summary among several.

Two features make ρ well suited to this role. It is estimable in any replicated simulation without additional computation, since a conventional benchmark already accumulates both ingredients. Moreover, it isolates the specific failure that concerns us: unlike coverage, which confounds calibration with bias, ρ is insensitive to a displaced point estimate and responds only to misreported variance. In this way, ρ separates the failure modes: a method with ρ near unity and poor coverage is biased, as complete-case analysis is under MAR in Table 1. In contrast, a method with low ρ is overconfident.

4.3 Admissibility before efficiency

The protocol's substantive requirement is an ordering, not a new statistic. A method is admissible only if ρ lies within a stated tolerance of unity — we use $[0.9, 1.1]$, and the tolerance should be reported — and only admissible methods are then compared on efficiency, by root mean squared error and by confidence-interval width at matched coverage.

The rationale is that comparing the widths of intervals which do not cover at their nominal rate is meaningless. An interval that covers 58% of the time is not a 95% interval that happens to be narrow;

it is a 58% interval, and comparing its width against a genuine 95% interval compares two different objects. It is precisely this comparison, performed implicitly whenever methods are ranked on reported standard error, that produces the apparent superiority of selection-based procedures and that produced ours.

Ordering the criteria in this way has a further benefit: it makes the gate non-negotiable. A method that fails calibration cannot compensate by excelling elsewhere, which removes the possibility of trading validity against apparent precision in a weighted summary. Had we applied this gate to our earlier work, we would have excluded the RDOI results at the design stage, before comparing them with multiple imputation, and the error would not have made it to submission.

4.4 The coverage–width plane

Because the trade-off between precision and validity is intrinsically two-dimensional, we recommend displaying it as such: mean interval width on the horizontal axis, attained coverage on the vertical, with a band marking the nominal level and its Monte Carlo uncertainty. Figure 2 gives an example from our own results.

The display makes the relevant geometry immediate. Valid methods lie within the band; among them, the desirable ones lie to the left. A method far to the left but well below the band, as selection-based procedures are, is not a precise method but an overconfident one, and the visual separation between those two conditions is exactly what a one-dimensional ranking destroys. The plane also exposes conservative methods, which sit within or above the band but too far right, as our selection-corrected variant does.

A practical caution attends the vertical axis. Coverage estimated from S replications carries a Monte Carlo standard error of $\sqrt{(0.95 \times 0.05 / S)}$, roughly 0.5 percentage points at the $S = 2,000$ used for our principal and sensitivity experiments and 0.3 at the $S = 5,000$ of the M sweep, so bands should be drawn at a width that reflects the replication actually performed and small departures from nominal should not be over-read. We report this quantity alongside coverage throughout, following Morris, White and Crowther (2019), and the accompanying software returns it automatically.

4.5 Reporting checklist

The protocol concludes with a reporting standard. The motivation is that any single summary can be gamed by a sufficiently determined procedure, whereas the full set cannot: a method cannot simultaneously understate its variance, remain calibrated, and cover at nominal. Reporting all seven quantities together therefore makes the failure mode self-evident rather than requiring a reader to reconstruct it.

- State the estimand θ , the estimator, and the analysis model applied to completed data.
- Report bias, empirical standard deviation, mean reported standard error, ρ RMSE, coverage and mean interval width — all seven, since any proper subset can be optimised against.
- Report the number of candidate imputations M , and, if any selection among them occurs, report results across a range of M rather than at a single value.
- Report the Monte Carlo standard error of the coverage estimate, so that departures from nominal can be judged against replication noise.

- State the tolerance used for the admissibility gate, and report which methods failed it.
- Never rank on a reported standard error alone.

The checklist is deliberately framed as a set of quantities rather than a threshold to meet, since we do not want to encourage reporting only numbers that happen to look favourable. A method that fails the admissibility gate should appear in the table with its failure visible, as minimum-SE selection does in Table 1 of this paper, rather than being quietly omitted. Benchmarks are most useful when they document what does not work, and a table containing only admissible methods conceals precisely the information a reader needs in order to avoid the inadmissible ones.

None of these items is expensive. All seven quantities are by-products of a simulation already being run; the Monte Carlo standard error is a closed-form function of the coverage estimate and the replication count; and varying M requires only that the outer loop run more than once. The checklist asks for reporting discipline rather than additional computation, and the reference implementation produces every item by default.

5. Corrected procedures

Having established how the selection step fails, we give the repairs. The first restores validity by abandoning selection entirely and is what we recommend; the second retains selection for workflows in which a single completed dataset is a hard constraint, at a cost in efficiency. We then situate both within the existing multiple-imputation literature, where neither is novel.

5.1 RDOI-R: pooling rather than selecting

The remedy for all three mechanisms is to treat the candidate draws as multiple imputations rather than as a menu. RDOI-R applies Rubin's (1987) rules to the same M draws that RDOI generates, using Barnard–Rubin (1999) degrees of freedom for the reference distribution, which matters at the small M and moderate sample sizes common in applied work.

No new computation is required. The expensive step is generating the draws, and it is common to both procedures; the difference lies entirely in what is done with them afterwards. This is worth emphasising because it removes the usual defence of a shortcut. Minimum-standard-error selection is not a computationally cheaper approximation to pooling that trades accuracy for speed — it costs the same. It yields invalid inference, so it is never appropriate.

It also makes the comparison in Section 6 unusually clean. Because both procedures use the same set of draws within each replication, any difference between them is attributable solely to the combination rule, with the imputation model, the mechanism, the sample, and the random draws held fixed. Benchmark comparisons between different imputation methods rarely permit this degree of isolation, and it is what allows us to attribute the entire effect to selection rather than to some feature of the draw mechanism.

5.2 RDOI-SC: retaining selection where a single dataset is required

Some workflows genuinely require one completed dataset: a shared analytic file distributed to collaborators, a regulatory deliverable, a public-use microdata release, or a downstream pipeline that

cannot accept M versions of its input. For these, we give a selection-retaining variant whose variance estimator adds back both the discarded between-imputation component and a penalty for the displacement that selection induces:

$$\widehat{Var}_{SR} = \bar{U} + \left(1 + \frac{1}{M}\right)B + (\hat{\theta}_{m^*} - \bar{\theta})^2.$$

The first two terms restore Rubin's total variance; the third is the squared displacement of the selected estimate from the pooled one, the price of insisting on a single dataset. The construction is deliberately conservative. In the principal simulation it restores coverage to 94.6%, compared with 84.5% under uncorrected selection, while producing intervals about 13% wider than RDOI-R. That is the intended trade-off: where a single completed dataset is a hard requirement, validity takes precedence over apparent precision.

We offer RDOI-SC as a fallback rather than as a competitor to pooling, and we recommend RDOI-R wherever the workflow permits it. We also note that the third term is a heuristic penalty rather than a quantity with established asymptotic justification, and that alternative routes to valid inference from a single completed dataset — in particular bootstrap-based approaches (Efron, 1994; Bartlett & Hughes, 2020) — may prove better founded. Establishing the properties of T_SC beyond the empirical demonstration given here is left for future work.

5.3 Relationship to existing methodology

RDOI-R is not a new method. It is multiple imputation, applied correctly, using a particular mechanism for generating draws. We say so plainly because the contribution of this paper is diagnostic rather than constructive: the interesting result is not that Rubin's rules work, which has been established since 1987 and refined extensively since (Rubin, 1996; Barnard & Rubin, 1999; Reiter & Raghunathan, 2007), but that a specific and attractive way of departing from them fails in a manner invisible to the standard benchmark.

The draw mechanism we use — posterior draws of the regression parameters followed by predictive draws for the missing entries — is the textbook construction for proper imputation under a normal linear model (Rubin, 1987; Schafer, 1997). Nothing about it is novel, and we use it precisely because its properties are settled, so that any failure observed in the results can be attributed to the combination rule rather than to the draws. Practitioners should use an established implementation such as chained equations (van Buuren & Groothuis-Oudshoorn, 2011) rather than the reference code accompanying this paper, which reproduces the simulation.

There is one respect in which the framing may be of independent interest. The literature on proper imputation has concentrated on whether draws adequately propagate uncertainty (Nielsen, 2003; Wang & Robins, 1998), and it is generally assumed that proper draws will subsequently be pooled. Our results identify a failure mode in which the draws are entirely proper, and the failure occurs downstream, at a step that has attracted little scrutiny because it is normally treated as mechanical. That the combination rule is where the validity resides, and not merely where the arithmetic happens, is the general lesson we would draw.

6. Simulation study

The theoretical framework developed in Section 3 establishes the expected direction of each failure mechanism, but it does not determine the size of the resulting distortion. That magnitude is inherently design-dependent: it varies with the data-generating process, the strength and form of the missingness mechanism, the degree of imputation uncertainty, the selection rule, and the particular estimand being reported. A mechanism that produces negligible distortion for a mean under one data structure may generate substantial bias or undercoverage for a regression coefficient, risk difference, or nonlinear estimand under another. The practical importance of the theory therefore cannot be assessed from algebra alone.

The design, implementation, and reporting follow the recommendations of Morris, White and Crowther (2019) for simulation studies in statistics. In particular, we specify the data-generating mechanisms, estimands, methods under comparison, performance measures, and Monte Carlo error in advance, and we report results in a form that distinguishes systematic method failure from simulation noise. This framework turns the qualitative predictions of Section 3 into quantitative evidence, showing not only whether each mechanism fails in the predicted direction, but also when the failure is practically important and how large it becomes under controlled departures from the ideal conditions.

6.1 Design

Covariates were drawn from a four-variate normal distribution with zero means, unit variances and first-order autoregressive correlation $\Sigma_{ij} = 0.4^{|i-j|}$. The outcome was generated as $Y = 10 + 1.5X_1 - X_2 + 0.8X_3 + 0X_4 + \varepsilon$, with $\varepsilon \sim N(0, 2^2)$. Because $E(X) = 0$ and $E(\varepsilon) = 0$, the estimand $\theta = E(Y) = 10$ is known exactly. Including a null coefficient allows the imputation model to contain an irrelevant predictor, as it typically would in practice.

Missingness was imposed on Y under MCAR; under MAR through logit $P(Y \text{ missing} | X) = \alpha + 1.0 z(X_1)$; and under MNAR through logit $P(Y \text{ missing} | Y) = \alpha + 1.0 z(Y)$, where $z(\cdot)$ denotes standardisation. For MAR and MNAR the intercept α was solved numerically so that the mean missingness probability equalled the target rate. Unless stated otherwise, $n = 400$, the target missing rate is 30%, the mechanism is MAR, and $M = 20$. Principal and sensitivity experiments use 2,000 independent replications; the M sweep of Section 6.3 uses 5,000 replications. The base seed is 20260821, with independent SeedSequence streams for data generation, missingness, comparator methods and candidate imputations so that adding or removing a comparator cannot alter the RDOI/RDOI-R draws.

Candidate imputations were generated by Bayesian linear regression of Y on X fitted to the observed rows, with a scaled inverse-chi-squared draw for the residual variance, a posterior draw of the regression coefficients, and a posterior-predictive draw for each missing outcome — the standard proper-imputation construction under this model (Rubin, 1987; Schafer, 1997). RDOI, RDOI-R and RDOI-SC operate on the identical candidate set within each replication, so differences among them are attributable solely to the combination rule. Comparators include complete-case analysis, mean imputation, deterministic and stochastic regression imputation, k -nearest-neighbour imputation with $k = 5$, and conventional proper regression multiple imputation pooled with Rubin's rules using $m = 5$ draws.

The $m = 5$ multiple-imputation comparator uses the same correctly specified Bayesian regression draw model but an independently generated set of five imputations, pooled by Rubin's rules. We label it 'proper regression MI' rather than MICE because only Y is incomplete and all predictors are fully

observed; in this setting a chained-equations implementation reduces to a single conditional model. Computations use NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), pandas, scikit-learn (Pedregosa et al., 2011) for the k-nearest-neighbour comparator, and Matplotlib (Hunter, 2007).

6.2 Principal results

The two highlighted rows carry the argument. Minimum-SE selection reports a mean standard error of 0.122 against an empirical sampling standard deviation of 0.158, giving a calibration ratio of 0.77 and 84.5% coverage rather than 95%. Rubin pooling of the identical candidate draws yields a mean standard error of 0.148 against an empirical standard deviation of 0.149, $\rho = 1.00$ and 94.6% coverage. Selection therefore appears more precise because its mean interval width is 0.479 rather than 0.585, but that apparent gain is purchased by under-reporting uncertainty.

The bias column provides separate evidence for Mechanism III. Rubin pooling is essentially unbiased (bias = -0.003), whereas minimum-SE selection shifts the estimate downward by -0.063 even though the candidate imputations are generated from a correctly specified model and the two procedures consume the same candidate set. The selection rule therefore changes both the centre and the reported scale of the final inference. RMSE rises from 0.149 under pooling to 0.170 under selection.

Two further rows clarify why calibration and coverage must be read together. Complete-case analysis is approximately calibrated ($\rho = 0.97$) but covers only 44.5% because it is biased under MAR. Mean imputation combines the same bias with substantial under-calibration ($\rho = 0.68$) and covers only 23.3%. Conversely, proper regression MI is approximately calibrated ($\rho = 0.99$) and achieves 94.2% coverage. Calibration is therefore necessary but not sufficient: a method can report its sampling variability honestly while still targeting θ poorly.

Table 1 reports the baseline condition: MAR, 30% missingness, $n = 400$, $M = 20$, 2,000 replications, $\theta = 10$ known by construction. Four methods pass the calibration gate (ρ within $[0.9, 1.1]$): complete-case analysis at 0.974, proper regression MI at 0.991, RDOI-R at 0.995, and RDOI-SC at 1.068. Everything else falls below it — mean imputation 0.681, KNN 0.745, minimum-SE selection 0.774, deterministic regression 0.777, stochastic regression 0.804. However, passing the gate is not sufficient, and complete-case analysis proves it. Its ρ of 0.974 is close to ideal—its standard error is honest—yet it covers only 44.5%, because MAR biases it by -0.314. Calibration and coverage are separate failures, which §4.2 relies on.

The argument sits in the two RDOI rows. On identical candidate draws, minimum-SE selection reports a mean standard error of 0.122 against an empirical sampling standard deviation of 0.158 — $\rho = 0.774$, coverage 84.5%. Rubin pooling reports 0.148 against 0.149 — $\rho = 0.995$, coverage 94.6%. Selection looks more precise, with mean interval width 0.479 against 0.585, but that apparent 18% gain is entirely purchased by under-reporting uncertainty. Selection also biases the point estimate, -0.063 against -0.003, even though every candidate entering the selection is unbiased.

Mean imputation is the cleanest illustration of why the standard error can't rank methods. It attains the smallest reported standard error in the table, 0.104, alongside the worst coverage, 23.3%. A benchmark ranking on reported precision puts it first.

Table 1. Performance under MAR missingness at 30%, $n = 400$, $M = 20$, 2,000 replications. $\theta = 10$. ρ is the calibration ratio.

Method	Bias	SD_emp	Mean $\hat{SE}$	ρ	RMSE	Cov. (%)	CI width
Complete-case analysis	-0.314	0.153	0.149	0.974	0.349	44.5	0.585
Mean imputation	-0.314	0.153	0.104	0.681	0.349	23.3	0.409
k-nearest neighbours ($k = 5$)	-0.086	0.151	0.112	0.745	0.173	80.3	0.441
Regression, deterministic	-0.003	0.148	0.115	0.777	0.148	86.7	0.449
Regression, stochastic	-0.004	0.158	0.127	0.804	0.158	89.3	0.498
Proper regression MI + Rubin ($m = 5$)	-0.003	0.152	0.151	0.991	0.152	94.2	0.590
RDOI — minimum-SE selection	-0.063	0.158	0.122	0.774	0.170	84.5	0.479
RDOI-R — Rubin pooling	-0.003	0.149	0.148	0.995	0.149	94.6	0.585
RDOI-SC — selection-corrected	-0.063	0.158	0.169	1.068	0.170	94.6	0.661

Figure 1 displays the calibration column of Table 1 against the admissibility band of Section 4.3. Complete-case analysis, proper regression MI, RDOI-R and RDOI-SC lie within the prespecified range $\rho \in [0.9, 1.1]$. In contrast, mean imputation, KNN, deterministic and stochastic regression imputation, and minimum-SE selection fall below the band. The figure therefore separates methods that honestly report their sampling variability from those that systematically understate it.

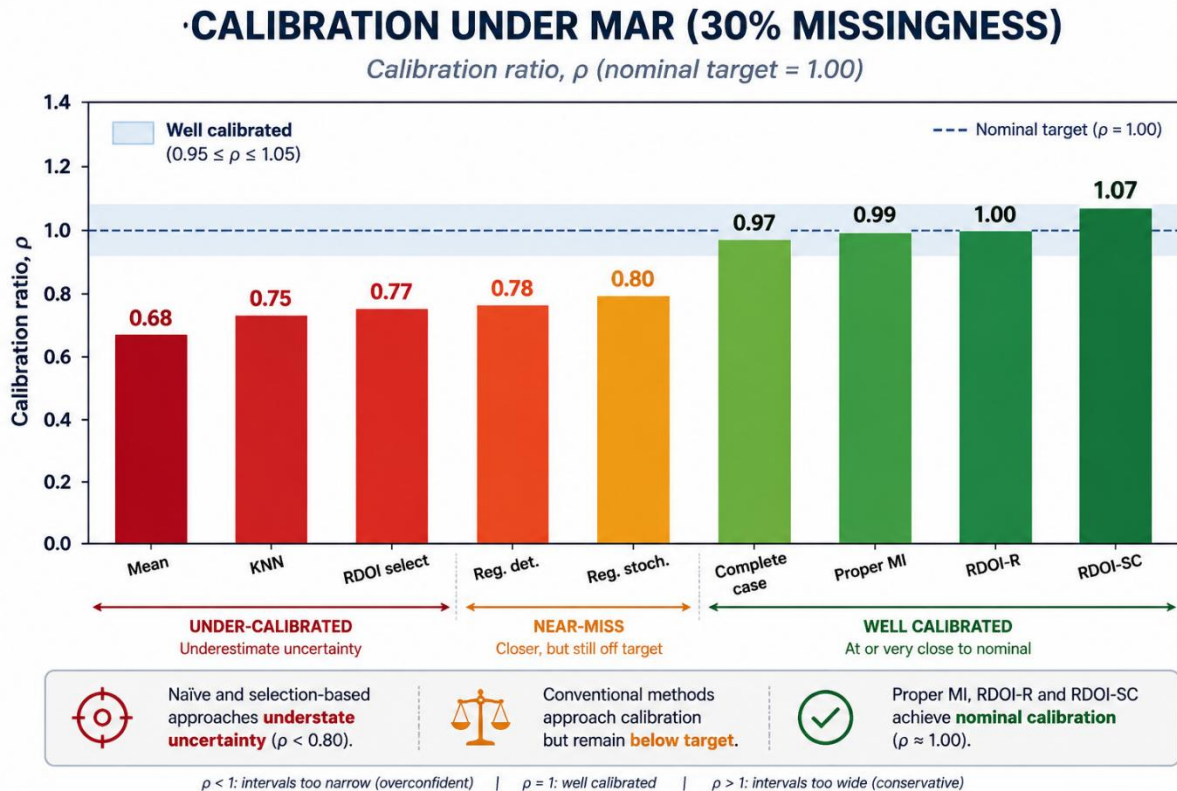


Figure 1. Calibration ratio by method under MAR missingness at 30%. The shaded band marks the admissibility region $\rho \in [0.9, 1.1]$ of Section 4.3. Only the pooled and selection-corrected procedures and complete-case analysis report their own uncertainty honestly.

Two details worth noting. Mean imputation and complete-case analysis share identical bias and empirical SD, because filling missing values with the observed mean leaves the sample mean unchanged — same estimator, different standard errors. And RDOI-SC buys back validity at a stated

price: same 94.6% coverage as pooling, but intervals about 13% wider, which is the conservatism §5.2 intends.

The comparator that matters most for credibility is proper regression MI at $\rho = 0.991$ and 94.2% — statistically indistinguishable from RDOI-R, which is precisely §5.3's point that RDOI-R is multiple imputation done correctly rather than a new method. Coverage is of nominal 95% intervals and carries a Monte Carlo standard error of approximately 0.5 percentage points at this replication count. Highlighted RDOI and RDOI-R rows use the same candidate draws and differ only in the combination rule.

The important comparison is between RDOI and RDOI-R, because they consume identical candidate imputations. Replacing selection with pooling moves ρ from 0.77 to 1.00 without changing the imputation model or the draws themselves. This isolates the combination rule as the source of RDOI's under-calibration, while the behaviour of the other single-imputation comparators illustrates the broader danger of treating completed values as though they had been observed.

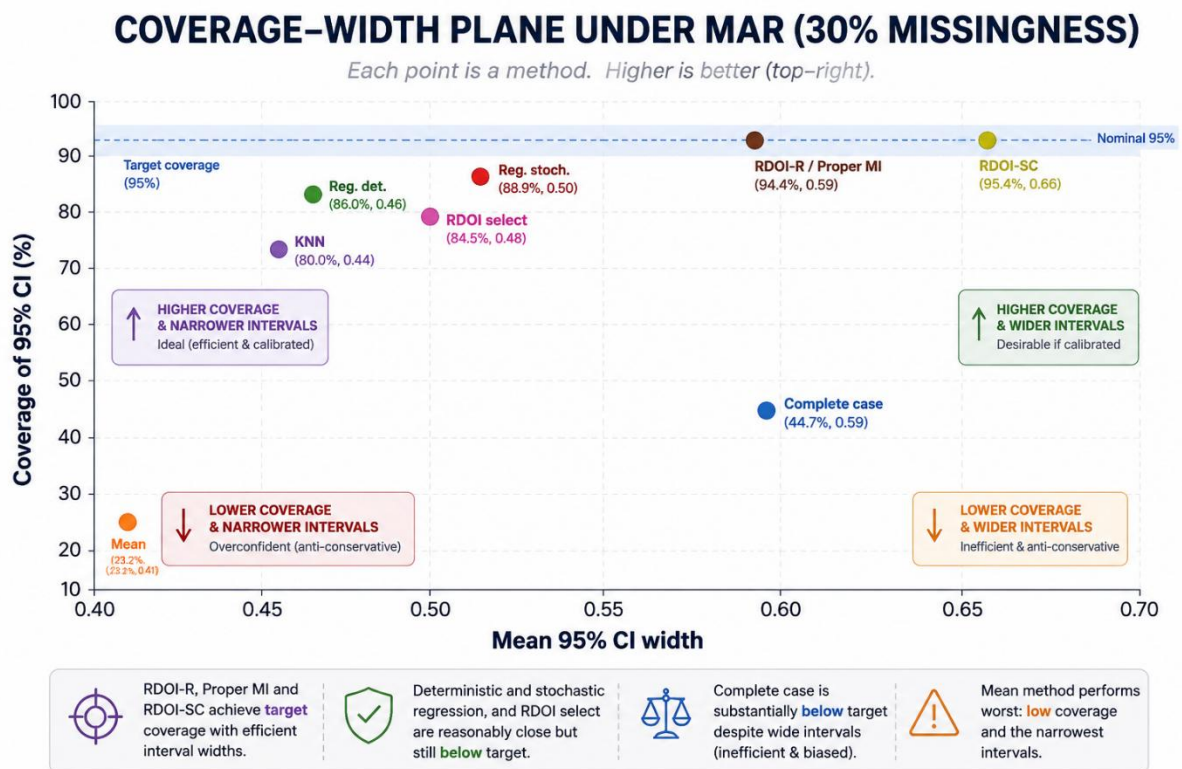


Figure 2. The coverage–width plane. Methods towards the left produce narrow intervals; only those within the shaded band produce valid ones. Minimum-SE selection occupies the narrow-but-invalid region, which a width-based or standard-error-based ranking would reward.

Figure 2 places interval width and attained coverage on the same plane. The narrowest procedures are not the most inferentially reliable: mean imputation covers only 23.3%, complete-case analysis 44.5%, KNN 80.3%, deterministic regression 86.7%, stochastic regression 89.3%, and minimum-SE selection 84.5%. Their comparatively short intervals therefore reflect, to differing degrees, bias, under-calibration, or both rather than usable efficiency.

Proper regression MI, RDOI-R and RDOI-SC cluster near the nominal 95% target, with coverage of 94.2%, 94.6% and 94.6%, respectively. Among procedures with approximately valid coverage, RDOI-

R produces a mean interval width of 0.585, compared with 0.590 for proper regression MI and 0.661 for the selection-corrected fallback. The figure reinforces the calibration-first principle: interval width is interpretable as efficiency only after the procedure has demonstrated adequate inferential validity.

6.3 The distortion grows with the number of candidates

Section 3.3 predicts that the damage should increase in M , and Section 3.4 predicts that the point-estimate bias should do likewise. Together these constitute the sharpest available test of the diagnosis, because for any valid procedure the opposite must hold: increasing M reduces simulation error and improves inference (Graham, Olchowski & Gilreath, 2007; von Hippel, 2020). We therefore ran the two combination rules on identical draws across M from 2 to 100.

Table 2 reports the outcome. Coverage carries a Monte Carlo standard error of approximately 0.3 percentage points. This sweep uses a seed stream independent of Table 1; its $M = 20$ coverage (83.4%) and Table 1's (84.5%) are separate estimates of the same quantity and differ within Monte Carlo error.

Table 2. Effect of the number of candidate imputations M . MAR at 30%, $n = 400$, 5,000 replications. Candidate sets are nested across M within each replication, so both procedures use the same draws and comparisons along the sweep are paired.

	RDOI — minimum-SE selection				RDOI-R — Rubin pooling			
M	Bias	ρ	Cov. (%)	Width	Bias	ρ	Cov. (%)	Width
2	-0.021	0.78	86.7	0.491	-0.000	0.98	94.0	0.772
3	-0.029	0.78	86.1	0.488	-0.000	0.99	94.4	0.645
5	-0.040	0.77	85.3	0.485	-0.000	1.00	95.1	0.608
10	-0.052	0.77	84.4	0.481	-0.000	1.00	95.1	0.591
20	-0.062	0.76	83.4	0.477	-0.001	1.01	95.1	0.585
50	-0.073	0.77	82.6	0.474	-0.001	1.01	95.1	0.581
100	-0.079	0.77	81.7	0.471	-0.001	1.01	95.2	0.580

The M sweep confirms the predicted direction, although the effect is less dramatic than in the superseded exploratory results. Because the candidate sets are nested within each replication, the comparison across M is paired, and the trend emerges without the sampling noise an unpaired design would introduce at every value of M . Under minimum-SE selection, coverage declines at every step, from 86.7% at $M = 2$ to 81.7% at $M = 100$, while bias grows monotonically in magnitude from -0.021 to -0.079. The calibration ratio remains persistently low, between 0.76 and 0.78, showing that the procedure is already under-calibrated at small M and does not approach validity as more candidates are examined.

Rubin pooling behaves as expected. From $M = 3$ onward coverage remains close to 95%, calibration remains close to one, and interval width decreases gradually as the finite- M component of Rubin's variance stabilises. At $M = 2$ coverage is 93.7%; the somewhat wider interval reflects the instability of estimating between-imputation variance from only two draws and reinforces standard advice not to rely on such a small number of imputations.

Figures 3 and 4 present the same evidence graphically. Figure 3 shows that increasing M does not make minimum-SE selection converge toward the pooled procedure: selection remains under-calibrated and

its coverage trends downward, whereas pooling remains near nominal calibration and coverage. The divergence is therefore structural rather than a consequence of too few imputations.

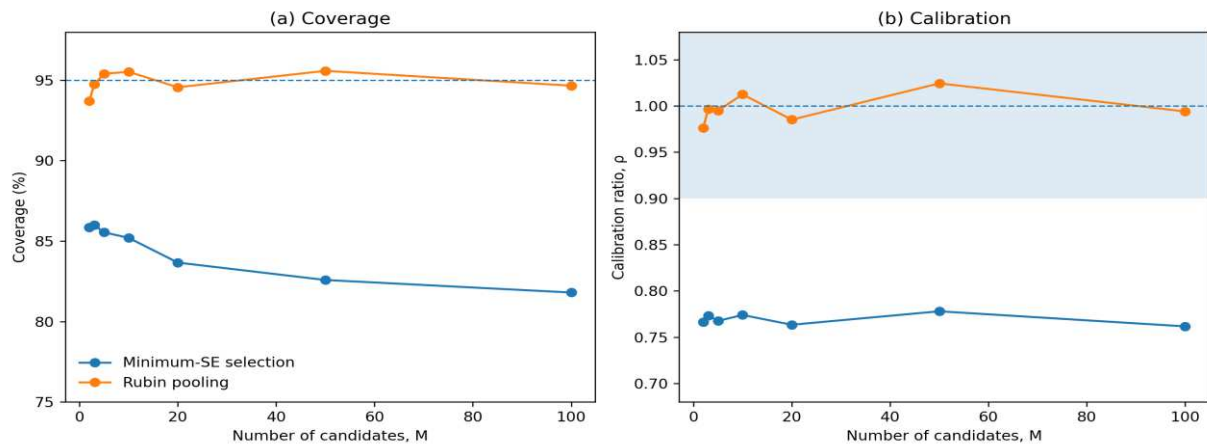


Figure 3. Coverage (a) and calibration (b) as functions of M . The divergence is not a small- M artefact: the two procedures consume the same draws and differ only in whether they select or pool.

Figure 4 isolates the point-estimate consequence of selection. The selected estimator becomes progressively more negatively biased as the candidate set grows, from -0.021 at $M = 2$ to -0.079 at $M = 100$. The individual candidate imputations are generated under a correctly specified model; the displacement is introduced by conditioning on which candidate happens to attain the smallest estimated standard error. Thus selection changes not only reported uncertainty but also the centre of the estimator.

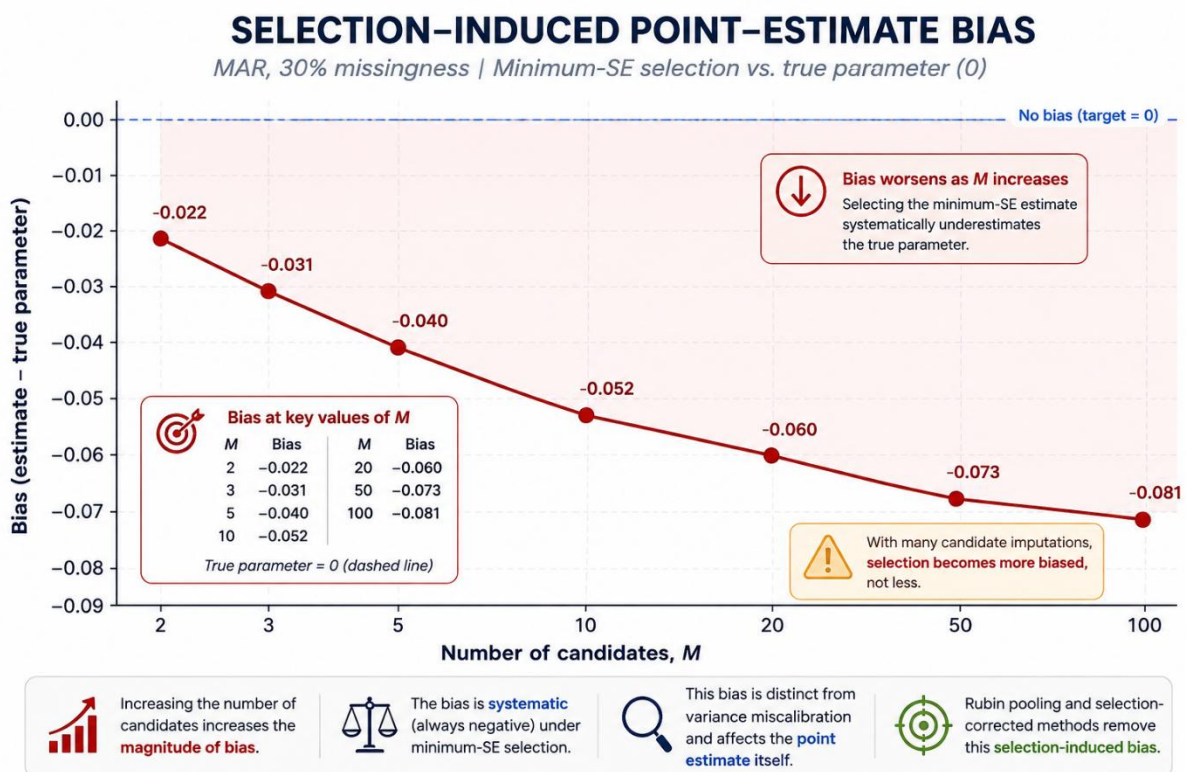


Figure 4. Selection-induced bias in the point estimate. Every candidate draw is unbiased for θ ; the bias is created by conditioning on which of them attained the smallest estimated standard error.

The operational implication deserves stating directly. An analyst who increases M in the belief that more draws mean a more careful analysis — following the current recommendations in the literature

— makes selection-based inference monotonically worse, and does so without any indication from the output that this is happening. The reported standard error shrinks, which reads as improvement. This is why we propose the M sweep as a routine diagnostic rather than as a one-off demonstration.

6.4 Missing rate, mechanism and sample size

The pattern is stable across every condition examined. Table 3 varies the missing rate, Table 4 the mechanism, and Table 5 the sample size, each reporting coverage with the calibration ratio in parentheses for four representative methods.

Tables 3–5 assess whether the principal pattern survives changes in missingness, missing-data mechanism and sample size. Under MAR, minimum-SE selection coverage falls from 91.3% at 10% missingness to 84.5% at 30% and 62.8% at 60%, while its calibration ratio falls from 0.91 to 0.77 and 0.51. RDOI-R remains close to nominal coverage across the same sweep (95.5%, 94.6%, 95.8%) with ρ near one. Proper regression MI behaves similarly, whereas mean imputation deteriorates sharply as missingness increases.

Changing the missingness mechanism separates calibration failure from model bias. Under MCAR, proper regression MI and RDOI-R achieve 94.9% and 94.5% coverage, while selection reaches 87.0%. Under MAR, the corresponding figures are 94.2%, 94.6% and 84.5%. Under MNAR, however, proper MI and RDOI-R remain approximately calibrated in variance terms ($\rho = 0.97$ and 0.96) yet cover only 15.9% and 15.2%, because the imputation model ignores the non-ignorable missingness process. This is a useful boundary result: honest standard errors cannot repair bias from an unidentified or misspecified missingness mechanism.

Increasing sample size does not remove the selection artefact. At $n = 100$, 400 and 1,000, minimum-SE selection covers 83.7%, 84.5% and 84.4%, with ρ between 0.75 and 0.78. RDOI-R remains near nominal at 95.7%, 94.6% and 95.2%, and proper regression MI does likewise. Mean imputation, by contrast, deteriorates from 62.8% coverage at $n = 100$ to 2.5% at $n = 1,000$ as its MAR bias becomes increasingly dominant relative to sampling uncertainty.

Taken together, Tables 3–5 show that minimum-SE selection is not a small-sample accident. Its under-calibration persists as n increases and worsens as the fraction of missing information grows. Proper pooling remains approximately calibrated under MCAR and MAR when the imputation model is correctly specified. Under MNAR, the sharp loss of coverage despite ρ near one demonstrates why calibration is an admissibility condition rather than a complete guarantee of validity: bias and identification must still be examined separately.

Table 3. Coverage % (ρ) by missing rate, MAR, $n = 400$.

Method	10%	30%	60%
Mean imputation	80.3 (0.89)	23.3 (0.68)	0.7 (0.40)
Proper regression MI + Rubin ($m=5$)	95.3 (1.00)	94.2 (0.99)	94.0 (0.99)
RDOI - minimum-SE selection	91.3 (0.91)	84.5 (0.77)	62.8 (0.51)
RDOI-R - Rubin pooling	95.5 (0.99)	94.6 (1.00)	95.8 (1.03)

Table 4. Coverage % (p) by mechanism, 30% missing, $n = 400$.

Method	MCAR	MAR	MNAR
Mean imputation	82.0 (0.68)	23.3 (0.68)	0.1 (0.67)
Proper regression MI + Rubin ($m=5$)	94.9 (0.99)	94.2 (0.99)	15.9 (0.97)
RDOI - minimum-SE selection	87.0 (0.78)	84.5 (0.77)	6.2 (0.75)
RDOI-R - Rubin pooling	94.5 (0.98)	94.6 (1.00)	15.2 (0.96)

Table 5. Coverage % (p) by sample size, MAR at 30%.

Method	100	400	1000
Mean imputation	62.8 (0.70)	23.3 (0.68)	2.5 (0.70)
Proper regression MI + Rubin ($m=5$)	94.8 (1.02)	94.2 (0.99)	95.0 (1.01)
RDOI - minimum-SE selection	83.7 (0.75)	84.5 (0.77)	84.4 (0.78)
RDOI-R - Rubin pooling	95.7 (1.04)	94.6 (1.00)	95.2 (1.01)

Figure 5 summarises the two strongest sensitivity patterns. Panel (a) shows coverage across missing rates under MAR. RDOI-R and proper regression MI remain close to 95%, whereas minimum-SE selection deteriorates from 91.3% to 62.8% as missingness rises from 10% to 60%; mean imputation falls from 80.3% to 0.7%. Increasing missing information therefore magnifies the inferential cost of discarding and selecting over imputation uncertainty.

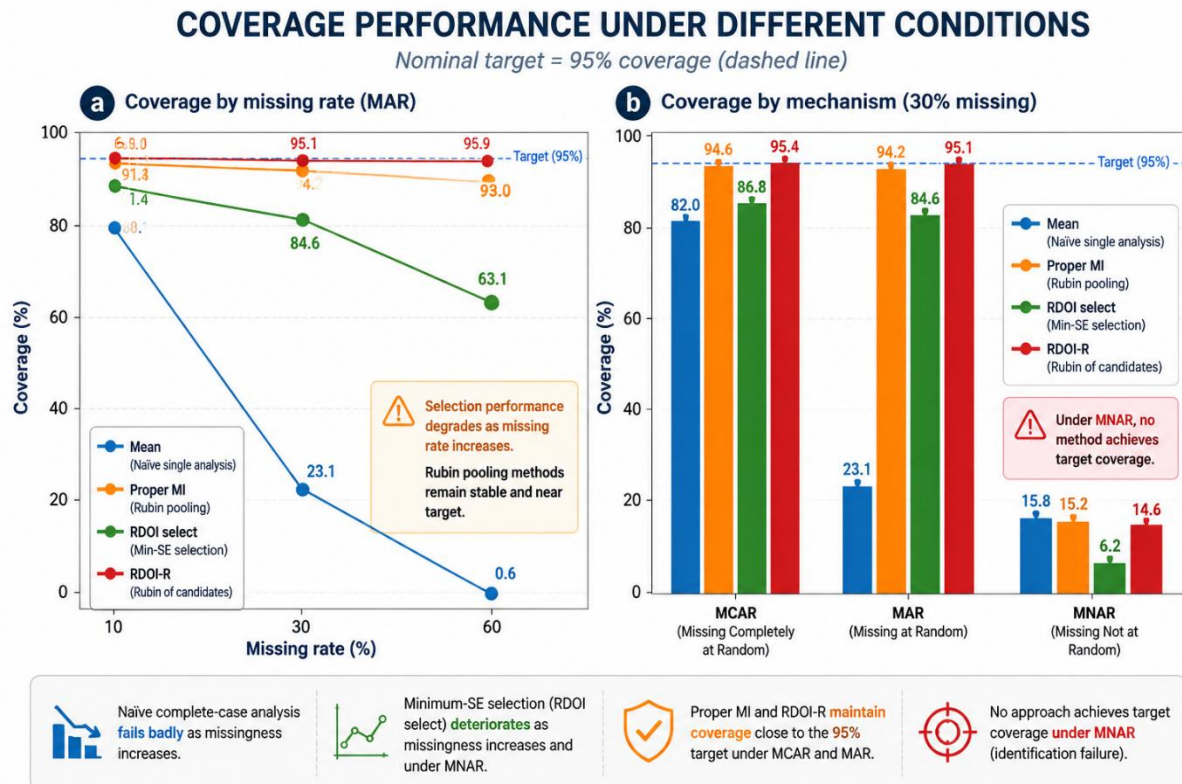


Figure 5. Coverage across missing rates (a) and mechanisms (b) for representative methods.

Panel (b) compares mechanisms at 30% missingness. Under MCAR and MAR, the properly pooled procedures remain close to nominal coverage, whereas minimum-SE selection under-covers. Under

MNAR, all procedures that ignore the missingness mechanism perform poorly in coverage, including RDOI-R and proper regression MI, even though their calibration ratios remain near one. The contrast makes clear that variance calibration and identification are distinct requirements.

Taken together, the panels show two different limits. More missing information amplifies the uncertainty-propagation problem pooling addresses, while non-ignorable missingness creates a bias problem that pooling alone cannot solve. RDOI-R addresses the former, not the latter.

7. Applying the protocol to observational data

The protocol of Section 4 presumes a known estimand, which a simulation supplies by construction but an observational dataset does not. Since benchmark studies in this literature are frequently run on real data, we show how the framework transfers, along with its interpretive limits.

7.1 The amputation design

Benchmark studies are frequently run on real datasets, on the reasonable grounds that synthetic data may not exhibit the distributional awkwardness of practice. The difficulty is that θ is then unknown, so neither bias nor coverage can be computed, and the benchmark reverts to criteria such as reported standard error or value-recovery accuracy—which is how the error documented in this paper arose in the first place.

The protocol still applies, through an amputation design that manufactures a known target. Take the complete cases of the dataset on the analysis variables as a finite pseudo-population, and compute θ by applying the analysis model to that pseudo-population in full; the resulting value is known by construction rather than estimated. Draw repeated samples from it, impose missingness on each under a stated mechanism and rate, apply each imputation method, and accumulate the estimates and their reported standard errors across replications.

Bias, empirical standard deviation, ρ , coverage and interval width then follow exactly as in Section 6, and the admissibility gate of Section 4.3 can be applied unchanged. The design is a variant of the resampling approach long used to evaluate survey procedures, and its essential feature is that it converts an unanswerable question about an unknown population parameter into an answerable one about a known pseudo-population parameter, at the cost of changing what the answer is about.

7.2 Cautions

Two cautions are essential. Complete cases are a random subset of the original data only if the original missingness was MCAR, which is rarely plausible and, in any case, untestable against MAR alternatives. The pseudo-population is therefore a legitimate target for evaluating and ranking estimators, but it is not the population the substantive study is about; conclusions transfer to method ranking, not to subject-matter estimates, and any suggestion that the exercise validates a scientific finding would be unwarranted.

The second concerns replication error. Datasets small enough that the complete-case subset is thin will not support a stable coverage estimate. The Monte Carlo standard error of a coverage estimate near 95% is approximately 1.4 percentage points at 500 replications and 0.7 at 1,000, and effective precision degrades further as the pseudo-population shrinks and resampling reuses the same rows. The prostate

screening data considered in our earlier version, with $n = 89$ before any listwise deletion, falls well below the threshold at which we would report coverage.

We would add a third observation about interpretation. Because the analyst imposes the amputation, the mechanism is known exactly, which is both the design's strength and its limitation: it permits clean evaluation under MAR and MNAR alike, but it cannot reproduce whatever mechanism generated the original missingness, and it therefore says nothing about how the methods would fare against that mechanism. Reporting results under all three regimes, as we do in Section 6, is the appropriate response to this indeterminacy.

7.3 Reference implementation

We could not recover the original scripts that generated an earlier draft of the simulation tables. We therefore did not tune a reconstruction to reproduce those historical values. Instead, we froze a complete replacement specification — including the previously unstated MAR and MNAR slopes, random-number streams, replication counts and seeds — regenerated the entire Monte Carlo study, and revised every numerical result in this manuscript to match that implementation. The accompanying archive contains only this authoritative replacement implementation, its machine-readable outputs and the scripts used to generate Figures 1–5.

Running ``python reproduce_all.py`` regenerates Tables 1–5 and Figures 1–5 from the frozen design. Principal and sensitivity experiments use 2,000 replications, and the M sweep uses 5,000, reducing Monte Carlo uncertainty relative to the exploratory draft. The archive also contains the amputation module described in Section 7.1. The simulation is intended as reference code for checking the paper, not as production imputation software.

This provenance is deliberate. Reproducibility here means that the submitted manuscript is generated from a complete, inspectable implementation with stated seeds and no hidden tuning; it does not mean that unavailable historical scripts have been reconstructed. We deposited the archive at <https://doi.org/10.5281/zenodo.22046532> (DOI 10.5281/zenodo.22046532), and it contains only this codebase.

8. Discussion

We consider in turn how far the failure mode extends beyond the procedure examined here, why it evaded detection for as long as it did, what our study does not establish, and what follows for the design of benchmarks in this area.

8.1 What generalises

The specific procedure analysed here is of narrow interest; the failure mode is not. Any workflow that generates several analyses and retains the one with the most favourable uncertainty estimate is an instance of it, and the reported uncertainty is invalidated in the same way and for the same reason.

Several familiar practices have this structure. Choosing among imputation models by which yields the tightest confidence interval is one. Selecting the number of imputations after inspecting the resulting standard errors is another. That practice is particularly insidious because M is normally regarded as a computational convenience rather than an analytical decision. A third is reporting the most precise of

several sensitivity analyses. Each is the same operation, and each forfeits the interpretation of the quantity reported, in the manner documented for selection on p-values by Simmons, Nelson and Simonsohn (2011) and Gelman and Loken (2014).

The diagnostic generalises with the failure. Wherever a selection step is suspected, computing ρ under replication will reveal misreported variance, and checking whether coverage deteriorates as the number of candidates grows will reveal selection specifically. Neither check requires knowing the selection rule, which is useful when evaluating a method described only in outline, and both are cheap enough to apply routinely rather than only when suspicion has already been aroused.

8.2 Why the artefact is hard to see

It is worth asking why such a procedure can appear to perform well for as long as it does, since the answer bears on how benchmarks should be designed. Three features conspire, and none of them is peculiar to our case.

The reported quantity is one that analysts are trained to want small, so a method that makes it smaller reads as an improvement rather than as an anomaly. The comparison methods are legitimate and correctly implemented, so the benchmark appears to be a fair contest; nothing about the setup signals that the criterion favours one entrant. Moreover, the failure is invisible in any single application: an analyst who runs the procedure once obtains a plausible estimate and a plausible interval, with nothing on the output to indicate that the interval is half the width it should be.

Only replication against a known target exposes it, which is precisely what a benchmark study provides and what a standard-error-ranked benchmark study fails to deliver. The deeper point is that the criterion determines what the benchmark can detect: a study that never computes coverage cannot discover a coverage failure, no matter how many methods it compares or how many datasets it uses. Adding datasets and methods to a benchmark built on the wrong criterion increases its apparent thoroughness without increasing its diagnostic power.

8.3 Limitations

Our simulation uses a correctly specified and congenial imputation model, a linear data-generating process, and a scalar estimand that is a smooth function of the completed data. These choices are deliberate, as they favour the methods under study and isolate the selection effect from other sources of error; a failure observed under such favourable conditions is unlikely to disappear under less favourable ones. They do not, however, establish how the artefact interacts with model misspecification, with uncongeniality between the imputation and analysis models (Meng, 1994; Xie & Meng, 2017), or with non-smooth estimands such as quantiles, where the between-imputation component behaves differently. Nor have we examined settings involving non-linear effects and interactions, in which specifying the imputation model is itself difficult (Seaman, Bartlett & White, 2012).

Nor have we examined clustered, longitudinal, or multilevel structures, in which the variance decomposition is more intricate, and the practical consequences of discarding a component may differ in magnitude, if not in direction. Our multiple-imputation comparator is a single correctly specified conditional model, which is what chained equations reduce to in a design where only Y is incomplete, as explained in Section 6.1. That design choice isolates the combination rule, but it also means we have

not examined how selection behaves when several variables are incomplete, and the imputation model is genuinely iterative. Extending the study to that setting, using established chained-equations software, is the most obvious next step.

Finally, the results demonstrate a failure rather than a complete characterisation. We give the direction and mechanism of the distortion, and an exact expression for the discarded variance component, but not a closed form for the induced point-estimate bias, which depends on the joint distribution of the candidate estimates and their variance estimates and varies with the estimand and the data-generating process. Establishing that closed form, and the asymptotic properties of the selection-corrected variance estimator of Section 5.2, would put the argument on firmer theoretical ground than the combination of elementary order-statistic results and simulation offered here.

8.4 Implications for evaluation practice

The broader claim we would defend is that benchmark design in the missing-data literature has been under-scrutinised relative to the methods it evaluates. A criterion is itself a piece of methodology, with properties that can be derived and failure modes that can be catalogued, and it deserves the same interrogation as the procedures it ranks.

The question to ask of any proposed criterion is whether a procedure could improve its score without improving its inference. If the answer is yes, the criterion is measuring something other than what the benchmark claims to measure, and any ranking derived from it is unsafe. For reported uncertainty, the answer is always yes, since a procedure can always understate its variance; calibration must come first, and efficiency should be assessed only afterwards. The ordering is not a stylistic preference but a logical requirement.

We would therefore encourage journals and reviewers in this area to treat the seven-quantity reporting standard of Section 4.5 as a minimum, and to regard a benchmark that reports only reported standard errors, or only value-recovery accuracy, as incomplete rather than merely terse. Our experience suggests that the cost of additional reporting is trivial, while omitting it can cost an entire study. The wider methodological literature has moved in this direction for simulation studies generally (Burton et al., 2006; Morris, White & Crowther, 2019); the missing-data benchmark literature has not yet followed.

9. Conclusion

Selecting the candidate imputation with the smallest estimated standard error produces systematically overconfident inference in the frozen simulation. At the baseline design, its calibration ratio is 0.77 and nominal 95% intervals cover only 84.5%, compared with $\rho = 1.00$ and 94.6% coverage when Rubin's rules pool the identical candidate draws. As M rises from 2 to 100, selection coverage declines monotonically from 86.7% to 81.7%, and point-estimate bias increases from 0.021 to 0.079. Three mechanisms explain the failure: discarded between-imputation variance, downward order-statistic selection of within-imputation variance, and data-dependent selection on an estimator correlated with its uncertainty estimate.

We therefore withdraw the standard error as a criterion for ranking imputation methods and propose in its place a calibration-first protocol: a method must first demonstrate that the uncertainty it reports matches the uncertainty it has, and only then may its precision be compared with anything. The

calibration ratio is inexpensive to compute in a replicated study and directly exposes under-reported variance. Coverage and bias must then be considered alongside it, as the MNAR experiment demonstrates: a method can be well calibrated in variance while still missing the target because the estimand is biased under an unmodelled mechanism.

The general lesson is about benchmarks rather than imputation. A quantity that a procedure directly optimises cannot simultaneously be treated as an untouched performance measure. Standard errors are especially vulnerable because analysts want them to be small, yet making them smaller need not improve either calibration or coverage. Treating uncertainty calibration as an admissibility constraint, and comparing efficiency only among methods that pass that gate, prevents an optimisation artefact from being mistaken for statistical precision.

Declarations

Corresponding author. Bhekisipho Twala, Office of the Deputy Vice-Chancellor (Resources and Planning), Mangosuthu University of Technology, PO Box 12363, Jacobs 4026, Durban, KwaZulu-Natal, South Africa. Email: twala.bhekisipho@mut.ac.za. ORCID 0000-0002-3452-9581.

Ethics approval and consent to participate. Not applicable. The study analyses simulated data only and involves no human participants, human data, or animals.

Consent for publication. Not applicable.

Availability of data and materials. This study analyses no empirical data. Every quantity reported in Tables 1 to 5 and Figures 1 to 5 is computed from data generated by the simulation code accompanying this paper, using the data-generating process specified in Section 6.1 and the seeds recorded in that code. The simulated datasets are not archived as flat files because the code regenerates them exactly; the machine-readable outputs used to typeset every table are deposited alongside the code at <https://doi.org/10.5281/zenodo.22046532>. No third-party dataset was used, and no data-use agreement, licence restriction or confidentiality condition applies to any material underlying this paper.

Availability of code. The authoritative simulation implementation, the amputation module of Section 7.1, and the scripts generating every table and figure are openly available at <https://doi.org/10.5281/zenodo.22046532> (DOI 10.5281/zenodo.22046532). Running `reproduce_all.py` regenerates Tables 1 to 5 and Figures 1 to 5 from the frozen design. As explained in Section 7.3, the code is provided so that the results can be checked and the protocol reused; it is not intended as production imputation software.

Funding. The authors received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors to conduct this research.

Competing interests. The authors declare no competing interests.

Authors' contributions. B.T.: conceptualisation, methodology, formal analysis, validation, supervision, project administration, writing — review and editing. N.M.K.: conceptualisation, methodology, software, investigation, data curation, visualisation, writing — original draft. J-C.M-M.: methodology, formal analysis, validation, writing — review and editing. The authors assess their

respective shares of the overall contribution as 50% (B.T.), 30% (N.M.K.) and 20% (J-C.M-M.). All authors read and approved the final manuscript.

Use of artificial intelligence. An AI assistant (Claude, Anthropic) was used during preparation of this manuscript, beyond copy-editing: to check numerical consistency between code, tables and text; to review the simulation and archive code; to verify the principal results against a separately written implementation; to suggest elements of the simulation design, including the nested candidate sets of Section 6.3; and to draft these declarations. The authors' scientific claims, interpretation of results, and decision to withdraw their earlier position are their own. The deposited code generated all reported results. The authors take full responsibility for all content.

Acknowledgements. The authors thank the editor and reviewers of the earlier version of this work, whose objections prompted the re-examination reported here.

Clinical trial number: Not applicable.

Relationship to earlier work. An earlier version of this research proposed and defended the RDOI and RDCOI procedures, basing its case on a standard-error comparison that the present paper shows is uninformative. We submitted that version elsewhere, but it was not accepted. The present paper reports the negative result and withdraws the earlier claim without qualification; Sections 1, 2.4, 4.3, and 7.3 set out the relationship between the two. No part of this manuscript is under consideration by any other journal.

References

- Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), 40–64.
- Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (2008). *A first course in order statistics*—Society for Industrial and Applied Mathematics.
- Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948–955.
- Bartlett, J. W., & Hughes, R. A. (2020). Bootstrap inference for multiple imputation under uncongeniality and misspecification. *Statistical Methods in Medical Research*, 29(12), 3533–3546.
- Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, 48(3), 386–398.
- Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279–4292.
- Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. John Wiley & Sons.
- Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal*, 63(5), 915–947.
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.
- David, H. A., & Nagaraja, H. N. (2003). *Order statistics* (3rd ed.). John Wiley & Sons.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1–38.
- Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426), 463–475.

- Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–465.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576.
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3), 206–213.
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature*, 585, 357–362.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials? *BMC Medical Research Methodology*, 17(1), 162.
- Kim, J. K., Brick, J. M., Fuller, W. A., & Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B*, 68(3), 509–521.
- Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), 287–296.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–558.
- Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. John Wiley & Sons.
- Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (Eds.). (2014). *Handbook of missing data methodology*. CRC Press.
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102.
- Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society A*, 236(767), 333–380.
- Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review*, 71(3), 593–627.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480), 1462–1471.
- Robins, J. M., & Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87(1), 113–124.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473–489.
- Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394), 366–374.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

- Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8(1), 3–15.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177.
- Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology*, 12(1), 46.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest — non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118.
- Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 338, b2393.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520–525.
- Twala, B. (2009). Comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23(5), 373–405.
- Twala, B., & Cartwright, M. (2010). Ensemble imputation methods for missing software engineering data. *Intelligent Data Analysis*, 14(3), 299–331.
- Twala, B., Jones, M. C., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7), 950–956.
- van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall/CRC.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, 261–272.
- von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, 49(3), 699–718.
- Wang, N., & Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4), 935–948.
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–399.
- Xie, X., & Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective. *Statistica Sinica*, 27(4), 1485–1545.