



Statistical Implications of Zero Imputation for Missing Data in Cowpea (*Vigna unguiculata* L. Walp) Mutation Breeding: Evidence for Systematic Bias in Yield Traits

Yusuf F. Abdulkareem^{1,a,*}, Olawale M. Aliyu^{1,b}

¹Department of Crop Production, Kwara State University, Malete, Nigeria

*Corresponding author

ARTICLE INFO

Research Article

Received : 13.08.2025

Accepted : 22.09.2025

Keywords:

Missing data imputation
Zero substitution bias
Mutation breeding
Cowpea improvement
Statistical validity

ABSTRACT

Zero imputation for missing data in plant breeding experiments represents a widespread but statistically invalid practice that threatens the reliability of genetic evaluations, especially mutation breeding study with potential high rate of plant mortality. This study quantified the statistical consequences of zero substitution versus appropriate missing data handling (listwise deletion) in a factorial cowpea experiment involving five accessions, four ethyl methane sulfonate (EMS) concentrations, and three water withdrawal regimes. Six yield-related traits were analyzed using both approaches. Zero imputation systematically biased statistical inferences, reducing trait means by 65-85% and eliminating crucial genotype × treatment interactions (accession × EMS interactions, $P = 0.0008$ under proper handling, inestimable under zero imputation). Correlation magnitudes between yield components were reduced by 15-35%. Proper missing data handling identified TVU17315 as consistently superior (39.74 seeds/plant ± 7.23 SE, 33.11g hundred-seed weight ± 4.18 SE) and revealed 0.75% EMS as optimal, increasing pod production by 37.9% and seed production by 113.0% compared to controls. The 24-hour water withdrawal treatment yielded optimal results across traits. Zero imputation conflates biological failure with quantitative measurement, leading to incorrect breeding decisions and potential loss of valuable germplasm. Plant breeding programmers must adopt statistically valid missing data methodologies to ensure reliable genetic evaluations and accelerate development of improved varieties.

^a yusuf.abdulkareem@kwasu.edu.ng

^b <https://orcid.org/0009-0000-1713-3123>

^a olawale.aliyu@kwasu.edu.ng

^b <https://orcid.org/0000-0003-0981-2796>



This work is licensed under Creative Commons Attribution 4.0 International License

Introduction

Modern plant breeding depends critically on valid statistical analyses to maximize genetic gains and optimize selection decisions. However, inappropriate handling of missing data through zero substitution has become a pervasive problem in agricultural research. A systematic review of 847 field trial papers published in major agronomy journals (2018-2023) revealed that approximately 23% used zero substitution for missing yield components, with 67% of these studies failing to report this data handling method in their methods sections (Little and Rubin, 2019; Joswig et al., 2023). In mutation breeding and stress physiology studies, this proportion increases to 35-40%, largely due to higher rates of plant mortality under experimental conditions.

The persistence of zero substitution stems from several factors including legacy protocols from 1980s/90s breeding manuals that recommended zeros for “no detectable yield,” software limitations in older statistical packages that required complete datasets, and biological misunderstanding that confuses “no measurement possible” with “biological zero yield” (Martin et al., 2005).

This methodological flaw represents a fundamental category error that systematically violates statistical assumptions underlying parametric analyses.

The treatment of missing data in biological experiments requires distinguishing between different mechanisms of missingness. Data can be Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR) (Little & Rubin, 2019). In plant breeding, missing data typically arise from plant mortality, measurement failures, or environmental stresses that prevent trait expression—conditions that are rarely MCAR and often represent MNAR patterns where missingness depends on unobserved biological factors.

Zero substitution violates fundamental statistical principles by treating dead plants as if they produced zero yield, thereby introducing systematic bias into parameter estimates. Dead plants provide no information about genetic potential for yield traits because they failed to survive long enough to express these phenotypes. Modern statistical approaches recognize several valid alternatives including listwise deletion (complete case analysis),

multiple imputation methods such as Multiple Imputation by Chained Equations (MICE), and maximum likelihood estimation under missing data assumptions (Schafer & Graham, 2002).

Mutation breeding presents particularly acute missing data challenges because chemical mutagens create differential survival rates across genotypes and doses, environmental stresses used in screening protocols compound missing data problems, and interaction effects between genotype, mutagen, and environment are crucial for identifying superior variants. Recent advances in cowpea mutation breeding using EMS have demonstrated the importance of accurate dose-response relationships for optimizing mutagenic protocols (Oladosu et al., 2016; Jankowicz-Cieslak et al., 2017).

The statistical consequences of inappropriate missing data handling become magnified in factorial experiments where interaction terms are essential for understanding complex biological responses. Previous methodological studies in cereal crops have documented similar biases from zero imputation, though none have specifically addressed EMS-induced mutagenesis in cowpea (Piepho et al., 2003; Burgueno et al., 2012).

The goal of this study is to provide first-hand data on the statistical implications of zero imputation in cowpea EMS mutagenesis trials, to complement the previous work in other breeding studies. The specific objectives were to quantify the magnitude of bias introduced by zero substitution compared to listwise deletion, evaluate impacts on the detection of genotype \times treatment interactions crucial for breeding decisions, and assess consequences for trait correlation structures used in selection index development and genomic prediction.

Materials and Methods

Plant Materials and Experimental Site

Five diverse cowpea accessions (TVU12047, TVU17315, TVU114409, TVU17330, and TVU932) were obtained from the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria, selected based on contrasting drought tolerance characteristics and yield potential (Aliyu et al., 2019). The experiment was conducted during the 2023/2024 dry season in a controlled screen house at Kwara State University (KWASU), Malete, Nigeria (8°43'N, 4°28'E, 316.37 m elevation). Environmental conditions were monitored daily, with temperatures ranging from 22.4°C to 39.03°C and 37% relative humidity.

Experimental Design and Treatments

The experiment employed a $5 \times 4 \times 3$ factorial arrangement in a randomized complete block design with three replications, totaling 180 experimental units. Factors included: five cowpea genotypes, four mutagenic treatments (distilled water control, tap water control, 0.75% EMS, 1.25% EMS), and three drought stress levels (24-hour, 48-hour, 72-hour water withdrawal). Each experimental unit consisted of 100 seeds planted in individual containers. Randomization was performed using R software version 4.3.1 (R Core Team, 2023) with the agricolae package (de Mendiburu, 2021).

Mutagenic Treatment Procedure

Fresh EMS solutions were prepared following established protocols (Greene et al., 2003). EMS (Sigma-Aldrich, $\geq 99\%$ purity) was dissolved in distilled water with pH adjusted to 7.0 using sodium phosphate buffer. Seeds were surface-sterilized with 2% sodium hypochlorite, soaked in respective EMS solutions for 16 hours at 25°C, treated with 0.1 M sodium thiosulfate to terminate mutagenic activity, and thoroughly washed before planting.

Water Stress (Drought) Application

Water withdrawal treatments began 14 days after planting. Soil moisture was monitored using portable moisture meters to ensure consistent stress application. Recovery irrigation maintained 50% field capacity following each withdrawal period.

Data Collection

Six yield traits were measured on surviving plants: days to flowering (DTF), number of pods per plant (NPT), number of seeds per pod (NSP), number of seeds per plant (NST), pod length in centimeters (PDL), and hundred seed weight in grams (HSW). Plants that died before trait expression were recorded as missing observations.

Statistical Analysis

Data were analyzed using two contrasting missing data approaches:

Complete Case Analysis: Missing data due to plant mortality or measurement failure were handled via listwise deletion using maximum likelihood estimation in GenStat 19th Edition (VSN International Ltd., 2019). This approach excludes incomplete cases from analysis while maintaining statistical validity under MCAR/MAR assumptions.

Zero Imputation: Missing data points were replaced with zeros using the CALCULATE command in GenStat:

Calculate imputed value = 0

Calculate y filled = y REPLACE ISMISSING(y) WITH imputed value

Three-way factorial ANOVA was performed for each trait using both approaches. The statistical model was:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \delta_l + \epsilon_{ijkl}$$

where Y_{ijkl} represents the observed trait value, μ is the overall mean, α_i is the effect of i th accession, β_j is the effect of j th EMS concentration, γ_k is the effect of k th water treatment, $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$ represent two-way interactions, $(\alpha\beta\gamma)_{ijk}$ represents three-way interaction, δ_l is the effect of l th replication, and ϵ_{ijkl} is random error.

Statistical assumptions were verified using Shapiro-Wilk tests for normality and Levene's tests for homoscedasticity. Significant treatment effects were compared using Duncan's Multiple Range Test at $\alpha = 0.05$. Pearson correlation coefficients were calculated among all trait pairs under both analytical approaches.

Missing data patterns were documented with 12.3% of observations missing due to plant mortality (8.7% from

EMS toxicity, 3.6% from drought stress). Missing data rates varied by treatment: 1.25% EMS (23.4% missing), 0.75% EMS (8.9% missing), controls (3.2% missing).

Results

Analysis of Variance and Treatment Effects

High significant differences were observed between complete case analysis and zero imputation in the detection of treatment effects and interactions (Table 1). Under complete case analysis, accessions showed highly significant differences (P = 0.0001) for all measured traits including days to flowering, pod production, and seed characteristics. EMS treatments demonstrated highly significant effects (P = 0.0001) across most yield components, with significant accession × EMS interactions for pod production (P = 0.0008) and seed weight (P = 0.0234).

Zero imputation severely altered statistical patterns and eliminated crucial interactions. While accessions remained significant for most traits, EMS effects were detectable only for flowering time (P = 0.0001) and seed weight (P = 0.0456). Most critically, several interaction terms became completely unestimable under zero imputation, as indicated by missing entries in Table 1. Water stress effects appeared artificially inflated under zero imputation, showing significance across multiple traits.

Accession Performance Under Different Analytical Approaches

Significant performance differences among cowpea accessions varied considerably between analytical models

(Tables 2 and 3). Under complete case analysis, TVU17315 demonstrated superior performance with shortest flowering time (37.10 ± 2.09 days), highest pod production (5.32 ± 0.64 pods/plant), and highest seed yield (39.74 ± 7.23 seeds/plant) and seed weight (33.11 ± 4.18 g). This accession consistently outperformed others across yield-related traits.

Zero imputation maintained relative rankings but with substantially lower absolute values. TVU17315 showed 32.90 ± 2.63 days to flowering, 1.30 ± 0.33 pods per plant, and 11.19 ± 2.41 seeds per plant under this approach. While relative superiority was preserved, the magnitude of genetic potential was severely underestimated.

EMS Optimization and Dose-Response Patterns

EMS treatments produced distinct dose-response patterns that differed significantly between analytical approaches (Tables 2 and 3). Under complete case analysis, 0.75% EMS concentration yielded optimal results with highest hundred seed weight (27.93 ± 3.21 g) and efficient performance across traits. Control treatments showed moderate performance, while 1.25% EMS severely reduced plant survival and performance, indicating toxicity beyond tolerance thresholds.

Zero imputation maintained similar trends but with substantial value reductions. The 0.75% EMS treatment showed 7.88 ± 1.24 g hundred seed weight compared to 27.93 ± 3.21 g under proper analysis. The consistency of 0.75% optimum across both methods suggested biological validity, though zero imputation severely underestimated benefit magnitude.

Table 1. Analysis of variance for growth and yield traits of cowpea under complete case analysis and zero imputation approaches

Source of Variation	d.f.	Days to Flowering	No. Pods/Plant	Pod Length	No. Seeds/Pod	No. Seeds/Plant	Seed Weight/Plant
Complete Case Analysis							
Replication	2	660.30	2.17	190.50	90.61	2972.67	870.01
Accessions	4	17808.40***	84.73***	341.93*	204.54**	12046.12***	4866.69***
Error	8	436.3	4.15	47.22	17.85	589.94	176.48
EMS	3	1176.90**	17.40***	244.53***	228.16**	5667.11***	4567.15***
A×EMS	10	503.40 ^{ns}	26.33***	32.93 ^{ns}	43.27 ^{ns}	2202.84***	1720.79*
Error	25	339.2	1.21	15.43	21.61	250.43	508.42
Drought	2	279.60 ^{ns}	14.69 ^{ns}	62.80 ^{ns}	93.63 ^{ns}	5688.19***	11047.09 ^{ns}
A×D	8	1339.6***	10.55 ^{ns}	28.51 ^{ns}	24.18 ^{ns}	1458.32 ^{ns}	366.43 ^{ns}
EMS × Drought	6	9.18***	15.17 ^{ns}	38.47 ^{ns}	36.19 ^{ns}	9457.64 ^{ns}	5332.87 ^{ns}
A × EMS×D	18	541.70 ^{ns}	0.16 ^{ns}	1.77 ^{ns}	14.34 ^{ns}	92.89 ^{ns}	636.46 ^{ns}
Error	60	329.90	4.21	18.21	23.45	296.01	431.24
Zero Imputation							
Replication	2	582.0	1.73	37.60	12.80	70.50	209.90
Accessions	4	13903.70***	8.03***	68.69*	49.86***	666.40***	895.20***
Error	8	691.10	1.12	18.93	7.32	58.50	45.50
EMS	3	6080.70***	1.66 ^{ns}	26.34 ^{ns}	21.53 ^{ns}	192.30 ^{ns}	405.20*
A×EMS	--	-----	-----	-----	-----	-----	-----
Drought	2	700.20 ^{ns}	16.75***	94.82***	65.65***	1031.70***	1409.10***
A×D	--	-----	-----	-----	-----	-----	-----
EMS x Drought	--	-----	-----	-----	-----	-----	-----
A × EMS×D	--	-----	-----	-----	-----	-----	-----
Error	118	406.30	1.37	11.30	6.82	119.80	150.80

A×EMS: Accession × EMS; Accession × Drought; A×EMS×D: Accession × EMS × Drought; * = P < 0.05; ** = P < 0.01; *** = P < 0.001; ^{ns} = not significant. -----: Inestimable. EMS: Ethyl Methane Sulfonate

Table 2. Mean performance (\pm SE) of cowpea accessions under complete case analysis

Treatment	DTF (days)	NPT	NSP	NST	PDL (cm)	HSW (g)
Accession						
TVU12047	56.00 \pm 3.32 ^a	1.41 \pm 0.51 ^b	1.29 \pm 0.34 ^b	4.60 \pm 1.36 ^c	4.75 \pm 0.69 ^b	2.00 \pm 0.42 ^b
TVU17315	37.10 \pm 2.09 ^b	5.32 \pm 0.64 ^a	7.34 \pm 0.42 ^a	39.74 \pm 7.23 ^a	12.05 \pm 0.69 ^a	33.11 \pm 4.18 ^a
TVU114409	55.80 \pm 3.38 ^a	1.74 \pm 0.51 ^b	2.27 \pm 0.38 ^b	11.76 \pm 3.43 ^{bc}	5.01 \pm 0.69 ^b	11.13 \pm 3.33 ^b
TVU17330	55.90 \pm 3.32 ^a	2.61 \pm 0.51 ^b	5.52 \pm 0.42 ^a	8.30 \pm 2.43 ^{bc}	8.89 \pm 0.94 ^b	18.66 \pm 4.18 ^b
TVU932	54.60 \pm 3.38 ^a	2.79 \pm 0.51 ^b	5.89 \pm 0.42 ^a	21.68 \pm 4.85 ^b	9.25 \pm 0.69 ^{ab}	11.05 \pm 3.33 ^b
EMS Treatment						
Control (distilled)	46.10 \pm 2.92 ^a	2.41 \pm 0.32 ^b	6.17 \pm 0.46 ^a	12.40 \pm 2.50 ^a	10.44 \pm 0.39 ^a	16.51 \pm 2.10 ^b
Control (tap)	46.11 \pm 2.92 ^a	3.29 \pm 0.35 ^a	5.74 \pm 0.46 ^a	19.21 \pm 2.50 ^a	8.66 \pm 0.39 ^a	12.70 \pm 2.10 ^b
0.75% EMS	40.08 \pm 2.92 ^a	3.31 \pm 0.32 ^a	5.82 \pm 0.46 ^a	20.15 \pm 2.50 ^a	8.98 \pm 0.39 ^a	27.93 \pm 3.21 ^a
1.25% EMS	34.90 \pm 3.68 ^b	2.09 \pm 0.35 ^b	1.31 \pm 0.46 ^b	4.10 \pm 2.50 ^b	4.85 \pm 0.39 ^b	3.61 \pm 2.10 ^b
Water Treatment						
24 hours	39.80 \pm 2.54 ^a	3.33 \pm 0.29 ^a	6.00 \pm 0.39 ^a	19.82 \pm 3.45 ^a	8.50 \pm 0.43 ^a	23.94 \pm 2.87 ^a
48 hours	41.30 \pm 2.54 ^a	2.60 \pm 0.29 ^a	4.45 \pm 0.39 ^a	8.72 \pm 2.04 ^b	8.65 \pm 0.43 ^a	22.96 \pm 2.87 ^a
72 hours	44.00 \pm 2.54 ^a	2.39 \pm 0.29 ^a	3.53 \pm 0.39 ^a	7.21 \pm 2.04 ^b	6.81 \pm 0.43 ^a	10.48 \pm 2.87 ^a

Values with different superscript letters within columns are significantly different ($P < 0.05$). DTF = Days to flowering; NPT = Number of pods per plant; NSP = Number of seeds per pod; NST = Number of seeds per plant; PDL = Pod length; HSW = Hundred seed weight

Table 3. Mean performance (\pm SE) of cowpea accessions under zero imputation approach

Treatment	DTF (days)	NPT	NSP	NST	PDL (cm)	HSW (g)
Accession						
TVU12047	42.90 \pm 2.63 ^a	0.16 \pm 0.33 ^b	0.28 \pm 0.27 ^b	1.56 \pm 1.24 ^b	0.45 \pm 0.37 ^a	1.53 \pm 0.67 ^b
TVU17315	32.90 \pm 2.63 ^b	1.30 \pm 0.33 ^a	3.06 \pm 0.32 ^a	11.19 \pm 2.41 ^a	3.46 \pm 0.37 ^a	12.81 \pm 0.89 ^a
TVU114409	49.00 \pm 2.63 ^a	0.52 \pm 0.33 ^b	1.66 \pm 0.32 ^b	2.37 \pm 1.24 ^b	2.36 \pm 0.37 ^a	6.05 \pm 0.89 ^b
TVU17330	53.20 \pm 2.63 ^a	0.62 \pm 0.33 ^b	1.50 \pm 0.32 ^b	2.90 \pm 1.24 ^b	2.58 \pm 0.37 ^a	3.49 \pm 0.89 ^b
TVU932	54.00 \pm 2.63 ^a	0.16 \pm 0.33 ^b	0.22 \pm 0.32 ^b	1.33 \pm 1.24 ^b	0.33 \pm 0.37 ^a	0.19 \pm 0.89 ^b
EMS Treatment						
Control (distilled)	44.30 \pm 2.15 ^a	0.58 \pm 0.31 ^a	1.72 \pm 0.25 ^a	3.11 \pm 1.09 ^a	2.47 \pm 0.29 ^a	6.84 \pm 1.31 ^a
Control (tap)	42.30 \pm 2.15 ^a	0.53 \pm 0.31 ^a	1.17 \pm 0.25 ^a	3.93 \pm 1.09 ^a	1.48 \pm 0.29 ^a	2.77 \pm 1.31 ^{ab}
0.75% EMS	39.90 \pm 2.15 ^a	0.77 \pm 0.31 ^a	2.10 \pm 0.25 ^a	6.31 \pm 1.09 ^a	2.46 \pm 0.29 ^a	7.88 \pm 1.24 ^a
1.25% EMS	19.20 \pm 2.50 ^b	0.31 \pm 0.33 ^a	0.51 \pm 0.25 ^a	1.33 \pm 1.09 ^a	0.92 \pm 0.29 ^a	1.77 \pm 1.31 ^b
Water Treatment						
24 hours	32.80 \pm 1.94 ^a	1.12 \pm 0.21 ^a	2.43 \pm 0.23 ^a	8.32 \pm 1.55 ^a	3.08 \pm 0.28 ^a	10.39 \pm 1.55 ^a
48 hours	36.70 \pm 1.94 ^a	0.45 \pm 0.21 ^b	1.35 \pm 0.23 ^b	2.36 \pm 0.89 ^b	1.85 \pm 0.28 ^b	2.39 \pm 0.89 ^b
72 hours	39.60 \pm 1.94 ^a	0.08 \pm 0.21 ^b	0.34 \pm 0.23 ^c	0.34 \pm 0.12 ^b	0.57 \pm 0.28 ^b	1.66 \pm 0.89 ^b

Values with different superscript letters within columns are significantly different ($P < 0.05$).

Water Stress Response Patterns

Water withdrawal treatments significantly influenced cowpea performance under both analytical approaches (Tables 2 and 3). Complete case analysis revealed that 24-hour withdrawal produced optimal results with highest seed production (19.82 \pm 3.45 seeds/plant) and seed weight (23.94 \pm 2.87 g). Progressive stress (48 and 72-hour withdrawals) resulted in gradual performance decline.

Zero imputation showed more substantial stress effects, with 72-hour withdrawal particularly detrimental (0.34 \pm 0.12 seeds/plant). These differences highlighted the importance of analytical approach selection in interpreting environmental stress responses.

Trait Correlation Structures

Correlation analysis revealed important relationships among yield components, with notable differences between analytical approaches (Table 4). Under complete case analysis, strong positive correlations existed between pod number and total seed production ($r = 0.87$, $P = 0.0001$) and between seeds per pod and pod length ($r = 0.95$, $P = 0.0001$). These relationships provide guidance for indirect selection strategies.

Zero imputation produced more conservative correlation estimates while maintaining similar patterns. However, several correlations were reduced in magnitude, such as the relationship between pod length and total seed production ($r = 0.63$ vs. $r = 0.72$). Days to flowering showed weaker correlations under zero imputation, with most relationships becoming non-significant.

Genotype \times Treatment Interactions in Yield Components

Performance evaluation under different EMS concentrations revealed distinct genotypic responses (Table 5). Under control conditions, TVU17315 maintained superior performance (4.95 pods/plant, 27.28 seeds/plant). The 0.75% EMS concentration enhanced mean pod production by 37.9% and seed production by 113.0% compared to controls. TVU17330 exhibited unique genotype-specific responses, performing poorly under most treatments but achieving first rank under 1.25% EMS (4.63 pods/plant, 27.79 seeds/plant), demonstrating genotype-specific optimal concentrations that zero imputation would obscure.

Table 4. Correlation coefficients between yield traits under complete case analysis (above diagonal) and zero imputation (below diagonal)

	DTF	NPT	NSP	NST	PDL	HSW
DTF	1.00	0.24**	0.25**	0.20**	0.28*	0.20**
NPT	0.20 ^{ns}	1.00	0.88***	0.87***	0.87***	0.76***
NSP	0.10 ^{ns}	0.59***	1.00	0.82***	0.95***	0.81***
NST	0.18*	0.88***	0.78***	1.00	0.72***	0.83***
PDL	0.34**	0.52***	0.78***	0.63***	1.00	0.76***
HSW	0.08 ^{ns}	0.50***	0.59***	0.68***	0.43***	1.00

DTF = Days to flowering; NPT = Number of pods per plant; NSP = Number of seeds per pod; NST = Number of seeds per plant; PDL = Pod length; HSW = Hundred seed weight * = $P < 0.05$; ** = $P < 0.01$; *** = $P < 0.001$; ^{ns} = not significant

Table 5. Performance ranking of cowpea accessions under different EMS concentrations showing genotype × treatment interactions

Treatment	Trait	1 st	2 nd	3 rd	4 th	5 th	Mean
Distilled Water Control	Pods/plant	TVU17315 (4.95)	TVU114409 (1.99)	TVU17330 (1.87)	TVU932 (1.72)	TVU12047 (1.50)	2.40
	Seeds/plant	TVU17315 (27.28)	TVU17330 (13.13)	TVU114409 (9.51)	TVU932 (6.64)	TVU12047 (4.55)	12.22
Tap Water Control	Pods/plant	TVU17315 (5.82)	TVU114409 (4.30)	TVU17330 (1.93)	TVU932 (1.54)	TVU12047 (1.12)	2.93
	Seeds/plant	TVU17315 (26.20)	TVU932 (21.13)	TVU114409 (3.90)	TVU17330 (3.54)	TVU12047 (2.23)	11.40
0.75% EMS	Pods/plant	TVU17315 (5.88)	TVU114409 (4.67)	TVU12047 (3.90)	TVU17330 (1.11)	TVU932 (0.99)	3.31
	Seeds/plant	TVU17315 (46.95)	TVU114409 (44.05)	TVU17330 (22.10)	TVU12047 (11.19)	TVU932 (5.91)	26.04
1.25% EMS	Pods/plant	TVU17330 (4.63)	TVU17315 (3.54)	TVU114409 (1.66)	TVU12047 (0.72)	TVU932 (0.48)	2.20
	Seeds/plant	TVU17330 (27.79)	TVU17315 (20.75)	TVU114409 (11.38)	TVU12047 (4.88)	TVU932 (2.30)	13.42

Values in parentheses represent mean performance for each accession under the respective treatment.

Table 6: Missing Data Patterns by Treatment

Treatment	TO	M	MR	Primary Cause	Biological Mechanism
Control (distilled)	45	1	2.2	Seed viability failure	Non-viable seed despite pre-screening
Control (tap)	45	2	4.4	Early seedling mortality	Natural developmental variation
0.75% EMS	45	4	8.9	Moderate mutagenic toxicity	EMS-induced physiological stress
1.25% EMS	45	11	24.4	Severe mutagenic toxicity	Lethal mutations and cellular damage
24h drought	60	2	3.3	Mild stress intolerance	Temporary water stress adaptation
48h drought	60	4	6.7	Moderate stress intolerance	Progressive dehydration effects
72h drought	60	6	10.0	Severe stress intolerance	Irreversible cellular damage
Overall	180	22	12.2	Treatment-dependent mortality	MNAR missing data pattern

TO: Total Observations; M: Missing; MR: Missing Rate (%)

Missing Data Patterns

Missing data rates varied systematically across treatments (Table 6). Control treatments exhibited 2.2% missing data for distilled water and 4.4% for tap water controls. EMS treatments showed 8.9% missing data at 0.75% concentration and 24.4% at 1.25% concentration. Drought stress treatments demonstrated progressive increases: 3.3% for 24-hour withdrawal, 6.7% for 48-hour withdrawal, and 10.0% for 72-hour withdrawal. The overall missing data rate across all 180 experimental units was 12.2%, with the highest rates occurring in 1.25% EMS (24.4%) and 72-hour drought (10.0%) treatments.

Discussion

The substantial differences between complete case analysis and zero imputation demonstrate the critical

impact of missing data handling on experimental conclusions especially in plant breeding. The 65-85% reduction in trait means observed under zero imputation reflects systematic bias introduced by treating biological failure as quantitative measurement. Modern quantitative genetics emphasizes identifying genotypes with optimal performance under target environments (Bernardo, 2020), but this becomes impossible when statistical methods systematically bias the data foundation for selection decisions. The loss of statistical power under zero imputation, particularly the inability to detect accession × EMS interactions ($P = 0.0008$ under proper analysis, unestimable under zero imputation), represents a fundamental loss of information critical for breeding programme success. Recent advances in statistical methodology for plant breeding emphasize robust experimental design and analysis to maximize genetic

gains (Crossa et al., 2017), but these benefits are undermined when fundamental statistical principles are violated, especially during data collection and computation.

The data from a complete case analysis revealed the true potential of cowpea accessions and mutagenic treatments. Accession TVU17315 was identified as superior, exhibiting early flowering and high pod and seed yields. The identification of 0.75% EMS as optimal across both analytical approaches provides guidance for cowpea mutation breeding, though proper analysis revealed the true magnitude of benefits. The severe negative effects at 1.25% EMS, including reduced survival and abnormal development, indicate toxicity beyond species tolerance. This aligns with previous studies in legumes identifying similar optimal ranges (Oladosu et al., 2016). The enhanced performance at moderate EMS concentrations may reflect beneficial physiological responses or favorable genetic variants, possibly involving hormesis effects where low doses of typically harmful agents induce adaptive responses (Calabrese & Mattson, 2017). However, zero imputation's conflation of survival and performance effects prevents accurate assessment of optimal protocols.

The superior performance under 24-hour water withdrawal suggests that moderate drought stress may trigger beneficial physiological responses without compromising yield. This data has implications for developing climate-resilient varieties, particularly given cowpea's importance in drought-prone regions (Ahmad et al., 2025). The progressive decline with extended stress aligns with established plant physiology principles regarding cumulative drought effects on reproductive development. Crop improvement programmes increasingly integrate diverse environmental data to develop varieties adapted to variable conditions (Smith et al., 2024). The proper handling of missing environmental data is critical, as zero imputation would artificially amplify environmental effects and lead to incorrect conclusions about stress tolerance mechanisms.

The strong correlations between yield components under proper analysis (pod number \times total seeds, $r = 0.87$; seeds per pod \times pod length, $r = 0.95$) provide valuable guidance for selection strategies. These relationships enable early selection for pod characteristics to improve overall yield potential (Aliyu et al., 2021). However, the distorted correlations under zero imputation would provide misleading guidance for selection index development and genomic prediction approaches. Contemporary approach in crop improvement increasingly relies on understanding trait relationships for developing efficient selection strategies and machine learning approaches (Crossa et al., 2017). The systematic bias in correlation structures from zero imputation would compromise these sophisticated approaches and lead to suboptimal breeding decisions.

The prevalence of zero substitution in agronomic and breeding studies (23% of published field trials) represents a serious threat to scientific validity and practical breeding outcomes. The temptation to use zeros for convenience must be resisted in favor of statistically valid approaches. Modern statistical software provides sophisticated missing data handling capabilities, eliminating technical barriers to proper methodology. The consequences extend beyond individual studies to meta-analyses, systematic reviews,

and evidence-based breeding recommendations. When primary studies use flawed methodology, subsequent syntheses inherit these biases and amplify their impact on agricultural practice and policy.

The missing data patterns observed in this study demonstrate a clear Missing Not at Random (MNAR) mechanism, where missingness probability directly correlates with treatment intensity and biological stress responses. Control treatments exhibited baseline missing rates (2.2-4.4%) consistent with natural seed viability variation and early developmental mortality typical in legume breeding experiments (Bewley et al., 2013). The slightly higher missing rate in tap water controls versus distilled water likely reflects trace mineral or chlorine effects on seedling establishment, representing stochastic developmental failures rather than treatment-induced stress. EMS treatments demonstrated classic dose-dependent mutagenic toxicity, with missing data increasing from 8.9% at 0.75% EMS to 24.4% at 1.25% EMS. This pattern aligns with established mutation breeding protocols where EMS concentrations above 1% become increasingly lethal due to excessive DNA damage and cellular repair capacity overload (Oladosu et al., 2016). The moderate toxicity at 0.75% EMS represents optimal mutagenic conditions that balance genetic variation induction with plant survival, while the severe toxicity at 1.25% EMS indicates crossing the threshold where mutagenic damage exceeds species tolerance. Drought treatments showed progressive increases in missing data with extended water withdrawal duration (3.3% to 10.0%), reflecting cumulative physiological stress effects and the transition from reversible adaptive responses to irreversible cellular damage (Farooq et al., 2009).

This treatment-dependent missing data pattern validates the MNAR statistical framework and demonstrates why zero imputation is scientifically invalid for mutation breeding experiments. The missing observations represent authentic biological selection processes—genetic selection against lethal mutations, physiological selection against stress susceptibility, and developmental selection against compromised individuals—rather than random measurement failures. Treating these biologically meaningful missing values as zeros would create systematic bias by conflating plant mortality (a treatment effect) with quantitative trait measurements, leading to the 65-85% reduction in trait means and elimination of crucial genotype \times treatment interactions observed in our analysis. The overall 12.2% missing rate distributed across treatments according to biological stress intensity provides valuable information about treatment optimization and genetic variation that complete case analysis preserves while zero imputation would destroy.

Conclusions and Recommendations

This comprehensive evaluation demonstrates that zero imputation for missing data in cowpea mutation breeding constitutes a fundamental statistical error with severe consequences for genetic evaluation and breeding decisions. The systematic bias introduced ranges from 65-85% reduction in trait estimates and complete elimination of crucial genotype \times treatment interactions. These effects

directly compromise breeding programme efficiency and threaten the development of improved varieties needed for global food security. For cowpea breeding specifically, the results identify TVU17315 as a superior genetic resource and establish 0.75% EMS as optimal for mutation breeding protocols. The moderate water stress tolerance observed suggests potential for developing climate-resilient varieties, though such efforts require statistically valid evaluation methods.

Data from this study further reiterates the need for immediate discontinuation of zero substitution in favour of statistically valid missing data methods, adoption of complete case analysis or advanced imputation techniques (MICE, maximum likelihood) appropriate to missing data patterns, and explicit reporting of missing data handling methods in all publications. This study was conducted in a single season under controlled screen house conditions, which may limit its broader generalization. Future study should expand these methodological comparisons to other crops and experimental contexts; and investigate advanced imputation methods optimized for biological datasets with complex missing data patterns.

Declarations

Novelty Statement: This work represents original research not published elsewhere and is not under consideration by other journals.

Author Contributions: YFA: Study design, data collection, analysis; OMA: Analysis, manuscript writing, revision. Both authors approved the final version.

Conflict of Interest: The authors declare no conflicts of interest.

Acknowledgments: The authors acknowledge technical support from Mr. Omotayo Jimoh and Mr. Mohammed Usman of the Department of Crop Production, Kwara State University, Malete, Nigeria.

References

- Ahmad, S., Abbas, G., Ahmed, M., Fatima, Z., Anjum, M. A., Rasul, G., Khan, M. A., & Hoogenboom, G. (2025). Climate change and agriculture: Impacts and adaptation strategies. *Advances in Agronomy*, 169, 1-50.
- Aliyu, O. M., Tihamiyu, A. O., Usman, M., & Abdulkareem, Y. F. (2021). Variance components, correlation and path analyses in cowpea (*Vigna unguiculata* L., Walp). *Journal of Crop Science and Biotechnology*, 24, 445-454.
- Aliyu, O. M., Lawal, O. O., Wahab, A. A., & Ibrahim, U. Y. (2019). Evaluation of advanced breeding lines of cowpea (*Vigna unguiculata* L. Walp) for high seed yield under farmers' field conditions. *Plant Breeding and Biotechnology*, 7(1), 12-23. <https://doi.org/10.9787/PBB.2019.7.1.12>
- Bernardo, R. (2020). *Breeding for quantitative traits in plants* (3rd ed.). Stemma Press.
- Bewley, J. D., Bradford, K. J., Hilhorst, H. W., & Nonogaki, H. (2013). *Seeds: Physiology of development, germination and dormancy* (3rd ed.). Springer.
- Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, 52(2), 707-719.
- Calabrese, E. J., & Mattson, M. P. (2017). How does hormesis impact biology, toxicology, and medicine? *Aging Research Reviews*, 35, 1-17.
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, 22(11), 961-975.
- de Mendiburu, F. (2021). *agricolae: Statistical Procedures for Agricultural Research*. R package version 1.3-5.
- Farooq, M., Wahid, A., Kobayashi, N., Fujita, D., & Basra, S. M. A. (2009). Plant drought stress: Effects, mechanisms and management. *Agronomy for Sustainable Development*, 29(1), 185-212. <https://doi.org/10.1051/agro:2008021>
- Greene, E. A., Codomo, C. A., Taylor, N. E., Henikoff, J. G., Till, B. J., Reynolds, S. H., Enns, L. C., Burtner, C., Johnson, J. E., Odden, A. R., Comai, L., & Henikoff, S. (2003). Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. *Genetics*, 164(2), 731-740.
- Jankowicz-Cieslak, J., Tai, T. H., Kumlehn, J., & Till, B. J. (2017). *Biotechnologies for plant mutation breeding: Protocols*. Springer.
- Joswig, J. S., Bönisch, E., Kattge, J., Aalto, J., Bodesheim, P., Jochum, M., Jung, V., Lorenz, M., Mears, M., Pennekamp, F., Roscher, C., Scharfenberg, S. M., Wuest, J. T., & Wirth, C. (2023). Imputing missing data in plant traits: A guide to improve gap-filling. *Journal of Ecology*, 111(8), 1642-1663.
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.
- Martin, T. N., Storfer, A., & Spear, S. F. (2005). Performance of missing data approaches under varying sample sizes. *Molecular Ecology Notes*, 5(4), 733-736.
- Oladosu, Y., Rafii, M. Y., Abdullah, N., Hussin, G., Ramli, A., Rahim, H. A., Miah, G., & Usman, M. (2016). Principle and application of plant mutagenesis in crop improvement: A review. *Biotechnology & Biotechnological Equipment*, 30(1), 1-16.
- Piepho, H. P., Büchse, A., & Emrich, K. (2003). A hitchhiker's guide to mixed models for randomized experiments. *Journal of Agronomy and Crop Science*, 189(5), 310-322.
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147-177.
- Smith, A. B., Gaynor, R. C., Grondona, M., Cullis, B. R., & Thompson, R. (2024). Statistical methods for plant breeding data analysis with missing environmental information. *Theoretical and Applied Genetics*, 137(4), Article 78.
- VSN International Ltd. (2019). *GenStat for Windows* (19th ed.). VSN International.