CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 2475–2493
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Bayesian Multiple Imputation Using Fully Conditional Specification for Handling Missing Values in Lung Cancer Data

Usman Pagalay[1], Ria Dhea L.N. Karisma*[1], Hawzah Saadati[1], Evawatialisah[2], Ratnaning Palupi[3], and Heni Widayani[1]

[1]*Program Studi Matematika Fakultas Sains dan Teknologi UIN Maulana Malik Ibrahim, Malang, Indonesia*
[2]*Program Studi Tadris Matematika Fakultas Ilmu Tarbiyah dan Keguruan UIN Maulana Malik Ibrahim, Malang, Indonesia*
[3]*State Polytechnic of Malang, Indonesia*

Abstract

Missing values in medical datasets can reduce statistical efficiency and introduce bias. This study applied Bayesian multiple imputation within the Fully Conditional Specification (FCS) framework and assessed its numerical stability. The data comprised 132 lung cancer patient records collected at Karsa Husada Hospital, Batu, between 2020 and 2025. The variables included age, sex, smoking status, comorbidity, family history, and lung cancer stage. Cancer stage was missing for 42.42% of patients (56/132). Associations between missingness indicators and observed variables were examined using χ^2 tests for categorical variables and independent t-tests for age. No statistically significant associations were detected ($p > 0.05$). These findings provided no evidence against MCAR for the variables examined but did not establish MCAR or distinguish MAR from MNAR. Imputation was conducted under an assumed MAR mechanism. All incomplete variables were binary and were imputed using Bayesian logistic regression with posterior parameter draws generated through the Metropolis–Hastings algorithm. The primary analysis generated 20 completed datasets using 20 FCS iterations. Trace plots, acceptance rates, and effective sample sizes showed no clear evidence of systematic non-convergence. Mean bias-reduced logistic regression was fitted to each completed dataset, and estimates were pooled using Rubin's rules. No predictor was significantly associated with Stage IV disease. Precision was limited, particularly for sex (95% CI: 0.012–22.40), smoking status (95% CI: 0.009–5.95), and family history (95% CI: 0.26–12.18). Sensitivity analyses using 50 and 100 completed datasets and a wider slope prior produced similar conclusions. The small sample size and high proportion of missing outcome data limited the precision of the estimates.

Keywords: Bayesian multiple imputation; Fully Conditional Specification (FCS); missing data; lung cancer

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (https://creativecommons.org/licenses/by-sa/4.0)

1. Introduction

The issue of missing values in a dataset is crucial to address. Missing values occur when values that should have been observed are unavailable due to reasons such as participant nonresponse,

*Corresponding author. E-mail: riadhea@uin-malang.ac.id

Received: July 24, 2026 Revised: August 20, 2026 Accepted: October 22, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.45218

data entry errors, loss of records, equipment malfunction, or incomplete data collection [1]. Imputation methods are required to accurately estimate missing values. Therefore, appropriate methods for handling missing data are needed to obtain valid and representative estimates. Some simple methods commonly used to handle missing data include mean imputation and mode imputation[2, 3]. These methods are commonly used because of their practical convenience. However, they have important limitations, including the potential to underestimate variability and produce biased estimates. Therefore, more principled approaches, such as multiple imputation, are needed.

Multiple imputation is a statistical technique that replaces each missing value with several plausible values [4, 5], thereby generating multiple complete datasets. Each completed dataset is analyzed separately. The results are later combined so that the uncertainty associated with missing data is reflected in the final inference [6]. Compared with single imputation, multiple imputation generally produces more valid statistical inference because it incorporates both within imputation and between imputation variability [7].

One approach using multiple imputation is Bayesian imputation. Bayesian imputation generates missing values using samples from the posterior distribution of the model parameters. Therefore, parameter uncertainty is reflected throughout the imputation process [8]. This approach is probabilistic. It preserves the natural variability of the data by producing multiple plausible values for each missing observation. In contrast, the K-Nearest Neighbor (KNN) imputation method is a deterministic approach. It estimates missing values based on the similarity between observations. This method does not explicitly reflect parameter uncertainty.

Selecting a missing value handling method requires considering the mechanism underlying the missing data. According to Rubin [6], missing data can be classified into three mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Because each missing data mechanism represents a different pattern of missingness. The performance of an imputation method varies according to the underlying mechanism. Each mechanism has distinct characteristics that can influence the choice of missing value handling method for further analysis. Several studies have suggested that the missing-data mechanism should be considered when selecting a method for handling missing values. This consideration helps align the imputation method with the characteristics of the data being analyzed [2, 3]. Therefore, missingness patterns should be examined before selecting an imputation strategy. The plausibility of alternative missing-data assumptions should also be assessed.

Several studies have demonstrated the effectiveness of model based multiple imputation approaches. Fully Conditional Specification (FCS) is one of the most widely used frameworks for handling missing data. It can be implemented through chained equations [5, 9]. FCS specifies a separate conditional imputation model for each incomplete variable. The missing values are then updated sequentially using the current values of the other variables. In a Bayesian implementation, plausible values are generated using posterior draws of the model parameters, thereby incorporating parameter uncertainty and preserving data variability [3, 10]. [11] reported that chained-equations imputation produced more accurate estimates and better preserved the correlation structure among variables than simple imputation methods. Several studies have also evaluated conventional imputation methods. Mukarromah et al. [12] reported that K-Means imputation achieved a lower Mean Squared Error (MSE) than mean imputation across different levels of missing data. Similarly, [13] found that the KNN Imputer, which was combined with Gradient Boosting, produced the lowest Root Mean Square Error (RMSE) compared with Mean Imputation and Interpolation methods.

Despite their advantages, existing imputation methods still have several limitations. For example, the K-Nearest Neighbor (KNN) method is a deterministic approach that estimates missing values based on similarities between observations, making it sensitive to data variability, sample size, and the choice of distance metric. Meanwhile, the model-based Bayesian approach

generates imputed values from posterior distributions. Parameter uncertainty is reflected in these distributions. The approach also uses relationships among variables represented in the probabilistic model [8]. These characteristics can enable the Bayesian approach to produce more representative imputations. This advantage may be particularly relevant when the variables have complex relationships or the sample size is relatively small.

Previous studies have demonstrated the effectiveness of Bayesian imputation and other missing-data methods. However, evidence regarding their application to medical datasets remains limited. This limitation is particularly relevant for datasets with relatively small sample sizes and substantial missingness. Therefore, this study aimed to apply Bayesian multiple imputation using the FCS framework to handle missing values in a lung cancer dataset. The study first examines the associations between missingness indicators and the available observed variables. Then, the numerical stability of the imputation procedure was evaluated. This evaluation used iteration diagnostics and posterior sampling diagnostics. The main contribution of this study is the integration of several analytical components within a multiple imputation framework. These components include missingness indicator analysis, iteration diagnostics, posterior sampling diagnostics, and mean bias-corrected logistic regression. The estimates obtained from the completed datasets are combined using Rubin's rules. This combination reflects the uncertainty arising from the imputation process. This study does not experimentally compare the performance of the Bayesian FCS procedure under simulated MCAR, MAR, and MNAR scenarios.

2. Methods

Missing values occur when some observations of a variable are missing from a dataset. Missing values can reduce the reliability of statistical analysis by reducing estimation efficiency, introducing bias, and affecting the validity of statistical inference [2]. Therefore, handling missing values is an important step in the data analysis process. The following are the most common missing data mechanisms.

2.1. Missing Value Mechanism

Based on [7], missing data can be classified into three mechanisms which are Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Consider a data matrix

$$Y = (Y_{\text{obs}}, Y_{\text{miss}}) \tag{1}$$

where

- Y_{obs} : observed data
- Y_{miss} : missing data

The missingness indicator is defined as

$$R = \begin{cases} 1, & \text{if } Y \text{ is observed,} \\ 0, & \text{if } Y \text{ is missing.} \end{cases} \tag{2}$$

The missingness mechanism is represented by

$$P(R \mid Y, \theta) \tag{3}$$

Eq. (3) represents the probability of the missingness pattern R conditional on the complete data Y and the parameter θ. MCAR occurs when the probability of missingness does not depend on either the observed or the missing data. Thus,

$$P(R \mid Y_{\text{obs}}, Y_{\text{miss}}) = P(R) \tag{4}$$

where Eq. (4) indicates that the probability of missingness does not depend on either the observed or the missing data. It indicates that the occurrence of missing values is unrelated to any characteristics of the data. Hence, the observed data represent a random sample of the population, allowing statistical analyses based on the observed data to produce unbiased estimates. Therefore, the missingness follows the MCAR mechanism.

Missing at Random (MAR) occurs when the probability of missingness depends only on the observed variables. Then, Missing Not at Random (MNAR) occurs when the probability of missingness depends on the missing values themselves. Therefore, identifying the missing data mechanism is required because it determines suitable method for handling missing data. Let

$$Y = (Y_{\text{obs}}, Y_{\text{miss}})$$

denote the data vector defined in Eq. (1). The MAR mechanism can be expressed as

$$P(R \mid Y_{\text{obs}}, Y_{\text{miss}}) = P(R \mid Y_{\text{obs}}) \tag{5}$$

Eq. (5) indicates that the probability of missingness depends only on the observed data. It is independent of the missing values after conditioning on the observed data. For MAR mechanism the observed variables contain enough information to model the distribution of the missing value. Therefore, model based methods, such as maximum likelihood estimation and multiple imputation, can be used to obtain unbiased parameter estimates.

However, data are classified as MNAR when the probability of missingness depends on the missing values themselves. The MNAR mechanism is defined as

$$P(R \mid Y_{\text{obs}}, Y_{\text{miss}}) \neq P(R \mid Y_{\text{obs}}) \tag{6}$$

Eq. (6) indicates that the probability of missingness cannot be explained using only the observed data because it depends on the missing values themselves. Accordingly, the observed variables do not provide enough information to accurately estimate the missing values. Therefore, handling MNAR data generally requires specialized models that consider the missingness mechanism.

2.2. Assessment of the Missing Data Mechanism

Before selecting an appropriate method for handling missing data, the missing data mechanism should be evaluated. Based on the classification proposed by [7], the missing data mechanism can be assessed through statistical tests of the missingness pattern. Most conventional imputation methods assume that the missing data mechanism follows either the MCAR or MAR assumption. For MCAR, missing values occur completely at random and are unrelated to any variables in the dataset. Under MCAR, the probability of missingness does not depend on either the observed or unobserved data. Since this study involves categorical variables, evidence against MCAR was assessed by testing the independence between the missingness indicator R and the observed variables Y_{obs} using Pearson's χ^2 test [14]. Failure to detect an association is consistent with MCAR for the observed variables examined, but does not establish MCAR because dependence on unobserved data cannot be assessed directly.

The null and alternative hypotheses are defined in Equations (7) and (8), respectively.

$$H_0 : \text{The missingness indicator } R \text{ is independent of } Y_{\text{obs}}, \tag{7}$$

$$H_1 : \text{The missingness indicator } R \text{ is not independent of } Y_{\text{obs}}. \tag{8}$$

The Pearson Chi-square statistic is given in Eq. (9)

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{9}$$

where O_{ij} denotes the observed frequency and E_{ij} denotes the expected frequency under the null hypothesis of independence.

The null hypothesis is rejected if

$$\chi^2 > \chi^2_{\alpha,(r-1)(c-1)} \tag{10}$$

Eq. (10) presents the decision rule for the Pearson chi-square test. The null hypothesis is rejected when the calculated chi-square statistic exceeds the corresponding critical value. Equivalently, the null hypothesis is rejected when the p-value is less than the significance level α.

If H_0 is not rejected, the result provides no evidence of an association between the missingness indicator and the observed variable examined. Therefore, the result is consistent with MCAR. However, failure to reject H_0 does not establish that MCAR holds. If H_0 is rejected, the result provides evidence against MCAR with respect to the observed variable examined. Such an association may be consistent with MAR. However, MAR and MNAR cannot generally be distinguished using the observed data alone.

2.3. Bayesian Imputation

After identifying the assumed missing-data mechanism, a method for handling the missing values was determined. In this study, Bayesian imputation was implemented within a multiple-imputation framework. The approach treats the missing values and the parameters of the imputation models as uncertain quantities. It also addresses the uncertainty associated with missing data in the subsequent statistical analyses [6, 9].

Under the Bayesian imputation approach, missing values are treated as random variables and sampled from their posterior predictive distributions conditional on the observed data. The MAR assumption was adopted in this study. The parameters of the data model were distinct from those of the missingness mechanism. Under these conditions, the missingness mechanism is ignorable and does not need to be modeled explicitly for Bayesian inference [2, 7].

Let Y_{obs} and Y_{miss} denote the observed and missing components of the data, respectively, and let $\boldsymbol{\theta}$ denote the parameters of the imputation model. Under the MAR assumption and an ignorable missingness mechanism, the joint posterior distribution of the missing values and model parameters can be expressed as [6]

$$p\left(Y_{\text{miss}}, \boldsymbol{\theta} \mid Y_{\text{obs}}\right) \propto p\left(Y_{\text{obs}}, Y_{\text{miss}} \mid \boldsymbol{\theta}\right) p(\boldsymbol{\theta}), \tag{11}$$

Eq. (11) represents the joint posterior distribution of the missing values and the imputation-model parameters. This equation applies under the MAR assumption and an ignorable missingness mechanism. In this equation, $p(\theta)$ denotes the prior distribution. Furthermore, $p(Y_{\text{obs}}, Y_{\text{miss}} \mid \theta)$ denotes the complete-data likelihood [7].

Within the Fully Conditional Specification (FCS) framework, a separate conditional imputation model is specified for each incomplete variable. FCS updates the missing values sequentially and repeatedly using the observed and most recently imputed values of the other variables as predictors. This framework is flexible because each variable can be modeled according to its distribution without requiring the explicit specification of a single joint distribution [5, 9].

Suppose that the dataset contains p variables, $\mathbf{Y} = (Y_1, \ldots, Y_p)$, of which some contain missing values. Let Y_{-j} denote all variables other than Y_j. For the j-th incomplete variable, FCS specifies the conditional model

$$p\left(Y_j \mid Y_{-j}, \boldsymbol{\beta}_j\right), \tag{12}$$

where $\boldsymbol{\beta}_j$ denotes the parameter vector of the j-th conditional imputation model. At FCS iteration c, the conditional posterior distribution is

$$p\left(\boldsymbol{\beta}_j \mid Y_{j,\text{obs}}, Y_{-j}^{(c)}\right) \propto p\left(Y_{j,\text{obs}} \mid Y_{-j}^{(c)}, \boldsymbol{\beta}_j\right) p(\boldsymbol{\beta}_j), \tag{13}$$

where $Y_{-j}^{(c)}$ contains the observed and currently imputed values of the remaining variables. A draw $\boldsymbol{\beta}_j^{*(c)}$ was obtained from the conditional posterior distribution. Conditional on this draw, the missing values were generated from

$$Y_{j,\text{miss}}^{(c)} \sim p\left(Y_{j,\text{miss}} \mid Y_{-j}^{(c)}, \boldsymbol{\beta}_j^{*(c)}\right). \tag{14}$$

FCS avoids the need to specify a complex joint distribution directly. However, its conditional models may not always correspond to a coherent joint probability distribution. Therefore, the conditional models should be specified carefully. They should also include substantively relevant predictors [9].

2.3.1. Regression Model for Imputation

Missing values are treated as random variables and estimated through the posterior distribution based on the observed data under the Bayesian Imputation approach. The obtained posterior distribution requires a statistical model that describes the relationship between the variable containing missing values and the other observed variables. The model commonly used is a regression model [5].

In this study, all variables requiring imputation were binary categorical variables. Therefore, the incomplete variables were modeled using Bayesian logistic regression within the chained equations framework. In this procedure, the incomplete variables were updated alternately and repeatedly using the current values of the other variables as predictors. The missing values were generated randomly based on their predictive probabilities [9].

For the j-th incomplete binary variable, the conditional probability of belonging to the event category was modeled as

$$\pi_{ij} = \Pr\left(Y_{ij} = 1 \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_j\right) = \frac{\exp\left(\mathbf{x}_{ij}^T \boldsymbol{\beta}_j\right)}{1 + \exp\left(\mathbf{x}_{ij}^T \boldsymbol{\beta}_j\right)}. \tag{15}$$

In Eq. (15), $Y_{ij} = 1$ denotes the internally coded event category for observation i, $\mathbf{x}_{ij}$ denotes the corresponding predictor vector, and $\boldsymbol{\beta}_j$ denotes the regression coefficient vector for the j-th imputation model. For each incomplete binary variable, the first category in the ordered observed categories was treated as the reference category, whereas the second category was treated as the event category and internally coded as $Y_{ij} = 1$.

The predictors, except the intercept, were standardized using the means and standard deviations calculated from observations for which Y_j was observed. This standardization was applied so that the prior distributions of the slope coefficients were defined on comparable scales. The intercept and slope coefficients were assigned normal prior distributions

$$\beta_{j0} \sim N(0, 5^2), \tag{16}$$

and

$$\beta_{jk} \sim N(0, 2.5^2), \qquad k = 1, \ldots, d_j - 1, \tag{17}$$

where d_j denotes the number of regression coefficients in the j-th imputation model.

For observations in which Y_j was observed, the likelihood of the binary logistic regression model was

$$L_j(\boldsymbol{\beta}_j) = \prod_{i \in \mathcal{O}_j} \pi_{ij}^{y_{ij}} \left(1 - \pi_{ij}\right)^{1 - y_{ij}}, \tag{18}$$

where $\mathcal{O}_j$ denotes the set of observations for which Y_j was observed, and π_{ij} is defined in Eq. (15).

The likelihood in Eq. (18) was combined with the prior distributions in Equations (16) and (17). This gives the conditional posterior distribution defined generally in Eq. (13).

The predictive probabilities depend on uncertainty in the regression coefficients. Therefore, the coefficients were not treated as fixed values. Draws from their posterior distribution were generated using the Metropolis–Hastings algorithm with a multivariate normal random-walk proposal [15, 16].

The initial coefficient vector and the proposal covariance matrix were obtained from a numerical approximation to the posterior distribution. The approximate posterior covariance matrix was based on the inverse Hessian of the negative log-posterior evaluated at its mode. The proposal covariance matrix was defined as

$$\boldsymbol{\Sigma}_{q,j} = \frac{2.38^2}{d_j}\widehat{\mathbf{V}}_{\boldsymbol{\beta}_j}, \tag{19}$$

where $\widehat{\mathbf{V}}_{\boldsymbol{\beta}_j}$ denotes the approximate posterior covariance matrix and d_j denotes the number of regression coefficients. The proposal covariance matrix was scaled by $2.38^2/d_j$, following a commonly used scaling rule for multivariate random walk Metropolis algorithms. The scaling factor controls the dispersion of the random-walk proposal [17].

The posterior distribution of the logistic regression coefficients does not have a standard closed form from which samples can be generated directly. Therefore, the Metropolis–Hastings algorithm was used to obtain draws from this posterior distribution. The algorithm constructs a Markov chain by repeatedly generating a candidate coefficient vector and deciding whether the candidate should replace the current coefficient vector. The proposal covariance matrix in Eq. (19) determines the size and direction of the changes proposed at each iteration. At Metropolis–Hastings iteration t, a candidate coefficient vector was generated from the multivariate normal random-walk proposal

$$\boldsymbol{\beta}_j^{\text{prop}} \mid \boldsymbol{\beta}_j^{(t-1)} \sim N\left(\boldsymbol{\beta}_j^{(t-1)}, \boldsymbol{\Sigma}_{q,j}\right). \tag{20}$$

The random-walk proposal in Eq. (20) was generated computationally using the eigen decomposition of the proposal covariance matrix. To generate candidate values while preserving the proposal covariance structure, eigen decomposition was applied [18, 19]

$$\boldsymbol{\Sigma}_{q,j} = \mathbf{Q}_{q,j}\boldsymbol{\Lambda}_{q,j}\mathbf{Q}_{q,j}^T. \tag{21}$$

Eigenvalues smaller than 10^{-10} were replaced by 10^{-10} to maintain numerical stability

$$\widetilde{\lambda}_{q,jk} = \max\left(\lambda_{q,jk}, 10^{-10}\right). \tag{22}$$

Let $\widetilde{\boldsymbol{\Lambda}}_{q,j}$ denote the diagonal matrix containing the adjusted eigenvalues and let $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$. The candidate coefficient vector was generated as

$$\boldsymbol{\beta}_j^{\text{prop}} = \boldsymbol{\beta}_j^{(t-1)} + \mathbf{Q}_{q,j}\widetilde{\boldsymbol{\Lambda}}_{q,j}^{1/2}\mathbf{z}. \tag{23}$$

The eigenvalue adjustments were used only to maintain numerical stability and were not interpreted as prior information. Since the normal random-walk proposal is symmetric then the proposal densities cancel from the Metropolis–Hastings ratio. The candidate was accepted with probability

$$\alpha_j = \min\left\{1, \frac{p\left(\boldsymbol{\beta}_j^{\text{prop}} \mid Y_{j,\text{obs}}, \mathbf{X}_{j,\text{obs}}\right)}{p\left(\boldsymbol{\beta}_j^{(t-1)} \mid Y_{j,\text{obs}}, \mathbf{X}_{j,\text{obs}}\right)}\right\}. \tag{24}$$

In the numerical implementation, the posterior ratio in Eq. (24) was evaluated on the logarithmic scale. Let

$$\Delta_j = \log p\left(\boldsymbol{\beta}_j^{\text{prop}} \mid Y_{j,\text{obs}}, \mathbf{X}_{j,\text{obs}}\right) - \log p\left(\boldsymbol{\beta}_j^{(t-1)} \mid Y_{j,\text{obs}}, \mathbf{X}_{j,\text{obs}}\right). \tag{25}$$

A value $u \sim \text{Uniform}(0,1)$ was generated. The coefficient vector was updated according to

$$\boldsymbol{\beta}_j^{(t)} = \begin{cases} \boldsymbol{\beta}_j^{\text{prop}}, & \text{if } \log(u) < \Delta_j, \\ \boldsymbol{\beta}_j^{(t-1)}, & \text{otherwise.} \end{cases} \tag{26}$$

The chain was initialized at $\widehat{\boldsymbol{\beta}}_{j,\text{MAP}}$. The first 1,000 iterations were excluded as initial iterations to reduce the influence of the starting value. Hence, 400 draws were retained by recording every third iteration. One of these draws was randomly selected for calculating the predictive probabilities and is denoted by $\boldsymbol{\beta}_j^*$.

The performance of the Metropolis–Hastings sampler was evaluated using the acceptance rate and effective sample size (ESS). The ESS was estimated using an autocorrelation-based formula [16, 20]

$$\text{ESS} = \frac{S}{1 + 2\sum_{\ell=1}^{L^*} \widehat{\rho}_\ell}, \tag{27}$$

where S denotes the number of posterior draws, $\widehat{\rho}_\ell$ denotes the estimated autocorrelation at lag ℓ, and L^* denotes the last lag included in the calculation. The ESS was calculated for each regression coefficient. The smallest ESS was recorded for each conditional imputation model.

$$L_{\max} = \min\left(S - 1, \left\lfloor 5\sqrt{S} \right\rfloor\right), \tag{28}$$

Eq. (28) defines the maximum lag $L_{\max}$ considered in estimating the effective sample size. The truncation lag L^* in Equation (27) was selected from the lags up to $L_{\max}$.

The selected posterior draw was used to calculate the predictive probabilities for observations with missing values in Y_j in Eq. (29):

$$p_{ij}^* = \frac{\exp\left(\mathbf{x}_{ij}^T \boldsymbol{\beta}_j^*\right)}{1 + \exp\left(\mathbf{x}_{ij}^T \boldsymbol{\beta}_j^*\right)}, \qquad i \in \mathcal{M}_j, \tag{29}$$

where $\mathcal{M}_j$ denotes the set of observations with missing values in Y_j. For the binary logistic imputation model, the general predictive draw in Eq. (14) was implemented as a Bernoulli draw as in Eq. (30).

$$Y_{ij,\text{miss}}^* \sim \text{Bernoulli}\left(p_{ij}^*\right), \qquad i \in \mathcal{M}_j. \tag{30}$$

Thus, the imputed values were random draws rather than fixed predictions. They reflected uncertainty in both the regression coefficients and the missing binary values. The procedure was repeated for each incomplete variable within the FCS framework to generate multiple completed datasets.

2.4. Chained Equations Process

The conditional model in Eq. (12) was applied sequentially to each incomplete variable. Let $\mathbf{Z}$ denote the fully observed variables. At FCS iteration c, the missing values were updated as follows

$$Y_{1,\text{miss}}^{(c)} \sim p\left(Y_{1,\text{miss}} \mid Y_{1,\text{obs}}, Y_2^{(c-1)}, \ldots, Y_p^{(c-1)}, \mathbf{Z}\right), \tag{31}$$

$$Y_{2,\text{miss}}^{(c)} \sim p\left(Y_{2,\text{miss}} \mid Y_{2,\text{obs}}, Y_1^{(c)}, Y_3^{(c-1)}, \ldots, Y_p^{(c-1)}, \mathbf{Z}\right), \tag{32}$$

$$\vdots \tag{33}$$

$$Y_{p,\text{miss}}^{(c)} \sim p\left(Y_{p,\text{miss}} \mid Y_{p,\text{obs}}, Y_1^{(c)}, \ldots, Y_{p-1}^{(c)}, \mathbf{Z}\right). \tag{34}$$

At each step, only the originally missing values of one variable are updated, while its observed values remain fixed. The most recently updated values of the other variables are used as predictors. One sequential update of all incomplete variables constitutes one iteration of the chained-equations procedure [5, 9].

2.5. Multiple Imputation

In multiple imputation, the imputation process is performed not only once but repeated m times, resulting in several completed datasets [6, 9].

$$D^{(1)}, D^{(2)}, \ldots, D^{(m)}. \tag{35}$$

For each completed dataset, the missing values were randomly initialized. They were then updated through the specified number of FCS iterations. Therefore, each completed dataset contained different plausible values for the originally missing observations.

Each completed dataset was analyzed separately, after which the resulting parameter estimates and their corresponding variances were combined using Rubin's rules. Multiple imputation was selected because it incorporates uncertainty associated with the missing values. This uncertainty is reflected in the resulting estimates and variances. Therefore, the imputed values are not treated as if they were observed without error [21]. The imputation procedure was conducted under the MAR assumption. The procedure was conditional on the observed variables included in the imputation models.

2.6. Mean Bias-Reduced Logistic Regression and Pooling

Mean bias-reduced logistic regression was fitted separately to each completed dataset in R using the `brglm2` package. The models were fitted using `glm()` with `method = brglm2::brglmFit` and `type = "AS_mean"`, that implements mean bias reduction through adjusted score equations. For the k-th completed dataset, let $\hat{\theta}_k$ denote the mean bias-reduced estimate of a regression coefficient. Its within-imputation variance, denoted by U_k, was obtained from the corresponding diagonal element of the model based variance–covariance matrix returned by `vcov()`. Thus, U_k represents the estimated sampling variance of $\hat{\theta}_k$ conditional on the k-th completed dataset. The standard model-based variance estimator was used.

The coefficient estimates and their within-imputation variances were pooled on the log-odds scale using Rubin's rules [6]. The pooled coefficient estimate was

$$\bar{\theta} = \frac{1}{m} \sum_{k=1}^{m} \hat{\theta}_k. \tag{36}$$

The average within-imputation variance was

$$\bar{U} = \frac{1}{m} \sum_{k=1}^{m} U_k, \tag{37}$$

and the between-imputation variance was

$$B = \frac{1}{m-1} \sum_{k=1}^{m} \left(\hat{\theta}_k - \bar{\theta} \right)^2. \tag{38}$$

The total variance was calculated as

$$T = \bar{U} + \left(1 + \frac{1}{m} \right) B. \tag{39}$$

The pooled standard error was therefore

$$\mathrm{SE}(\bar{\theta}) = \sqrt{T}. \tag{40}$$

Accordingly, the total variance incorporates two sources of uncertainty. The first is the average sampling uncertainty from the mean bias-reduced regression models fitted to the completed datasets, represented by $\bar{U}$. The second is the uncertainty arising from differences among the imputations, represented by B. The regression coefficients were first pooled on the log-odds scale. The pooled coefficients were then transformed into odds ratios using $\mathrm{OR}_j = \exp\left(\bar{\beta}_j\right)$.

2.7. Study Data and Implementation

The data used in this study were obtained from the medical records of Karsa Husada Hospital, Batu. The dataset consisted of patient diagnostic history reports collected from 2020 to 2025. A total of 132 observations were included in this study. The variables consisted of five independent variables (X) representing lung cancer risk factors according to [22] and one dependent variable (Y) representing the stage of lung cancer.

The independent variables consist of demographic and clinical characteristics, namely age, sex, smoking status, comorbidity, and family history. These variables were selected because previous studies have identified them as major risk factors contributing to the occurrence and progression of lung cancer [22]. Furthermore, factors such as age, smoking habits, and medical history have been reported to have significant associations with both the risk and severity of lung cancer [23]. In addition to clinical considerations, these factors are closely related to disease severity. Therefore, the variables used in this study were considered relevant for classifying lung cancer stages.

The dependent variable represented the stage of lung cancer and was categorized into two classes, namely Stage III and Stage IV. The classification was used to represent disease severity and to distinguish patients based on lung cancer progression. The description of variables presented in Table 1 was used throughout the imputation and regression analyses.

Table 1: Description of the study variables

Variable	Scale	Description
Age	Ratio	Patient age in years; analyzed as a continuous variable and not recoded.
Sex	Nominal	1 = Male; 2 = Female.
Smoking status	Nominal	1 = Active smoker; 2 = Non-smoker.
Comorbidity	Nominal	1 = Presence of comorbidity; 2 = Absence of comorbidity.
Family history	Nominal	1 = Presence of family history; 2 = Absence of family history.
Lung cancer stage	Nominal (binary)	Stage III was recoded as 0 and used as the reference outcome, whereas Stage IV was recoded as 1 and defined as the event.

Missing data were handled using multiple imputation based on Fully Conditional Specification (FCS), implemented manually in base R. All variables requiring imputation were binary categorical variables. Therefore, they were imputed sequentially using approximate Bayesian logistic regression models.

Age and sex were fully observed and were included as predictors in the conditional imputation models. The procedure generated $m = 20$ completed datasets, with 20 iterations performed for each imputation chain. The number of imputations was selected in consideration of the relatively high proportion of missing disease-stage values and the need to reduce Monte Carlo error [4, 24]. Each completed dataset was subsequently analyzed using bias-reduced logistic regression, and the resulting estimates and variances were combined using Rubin's rules.

The main steps of the data analysis and multiple imputation procedure are summarized in Algorithm 1.

3. Results and Discussion

Table 2 presents the descriptive statistics of the lung cancer dataset used in this study. The table summarizes the number of observations and provides information on the missing values for each variable.

The age variable is measured on a ratio scale, with a mean of 60.51 years and a standard deviation of 11.41 years, ranging from 25 to 85 years. In contrast, the variables sex, smoking status, comorbidity, family history, and lung cancer stage are categorical variables. As shown in Table 2, several variables have relatively high proportions of missing values, particularly lung cancer stage (42.42%), smoking status (28.03%), and family history (14.39%). Such a substantial

Algorithm 1 Bayesian multiple-imputation procedure

1. Clean and encode the variables in the original dataset D.
2. Calculate descriptive statistics and missing-value percentages.
3. Construct the missingness indicators as defined in Eq. (2).
4. Assess the associations between the missingness indicators and the observed variables using the Pearson χ^2 statistic in Eq. (9) and independent t-tests.
5. Identify the incomplete binary variables and specify the predictors for each conditional imputation model.
6. For each imputation $k = 1, \ldots, m$, perform the following steps:
   (a) Initialize the originally missing values by randomly sampling from the observed values of the corresponding variables.
   (b) For each FCS iteration $c = 1, \ldots, C$, update each incomplete binary variable Y_j sequentially according to Equations (31)–(34):
      i. Specify the Bayesian logistic regression model in Eq. (15) using the current values of the other variables as predictors.
      ii. Assign the prior distributions for the intercept and slope coefficients according to Equations (16) and (17).
      iii. Construct the likelihood according to Eq. (18) and combine it with the prior distributions to obtain the conditional posterior distribution in Eq. (13).
      iv. Construct the proposal covariance matrix using Eq. (19).
      v. Generate candidate coefficient vectors using the multivariate normal random-walk proposal in Eq. (20). The computational generation of the proposal is given in Equations (21)–(23).
      vi. Calculate the Metropolis–Hastings acceptance probability using Eq. (24), evaluate the posterior ratio on the logarithmic scale using Eq. (25), and update the coefficient vector according to Eq. (26).
      vii. Discard the first 1,000 Metropolis–Hastings iterations, retain 400 draws by recording every third iteration, and randomly select one retained posterior draw $\boldsymbol{\beta}_j^*$.
      viii. Calculate the acceptance rate and the effective sample size defined in Eq. (27).
      ix. Calculate the predictive probabilities for the missing observations using Eq. (29).
      x. Generate the missing binary values from the Bernoulli distribution in Eq. (30).
      xi. Keep the originally observed values unchanged and continue to the next incomplete variable.
   (c) After completing all C FCS iterations, store the completed dataset as $D^{(k)}$.
7. After all m imputations are completed, obtain the completed datasets defined in Eq. (35).
8. Evaluate the FCS trace plots, Metropolis–Hastings acceptance rates, and effective sample sizes.
9. Fit bias-reduced logistic regression separately to each completed dataset $D^{(k)}$.
10. Pool the parameter estimates using Eq. (36), calculate the within- and between-imputation variances using Equations (37) and (38), and calculate the total variance using Eq. (39).
11. Calculate the pooled standard errors using Eq. (40) and report the pooled odds ratios, 95% confidence intervals, and p-values.

Table 2: Descriptive statistics and missing data summary of the lung cancer dataset

Variable	Data Type	Descriptive Statistics	Valid n	Missing n	Missing (%)
Age	Numerical	Mean ± SD = 60.51 ± 11.41; Min–Max = 25–85	132	0	0.00
Sex	Categorical	Male: 95 (71.97%); Female: 37 (28.03%)	132	0	0.00
Smoking Status	Categorical	Smoker: 63 (66.32%); Non-smoker: 32 (33.68%)	95	37	28.03
Comorbidity	Categorical	Yes: 49 (37.69%); No: 81 (62.31%)	130	2	1.52
Family History	Categorical	Yes: 13 (11.50%); No: 100 (88.50%)	113	19	14.39
Lung Cancer Stage	Categorical (Binary)	Stage III: 62 (81.58%); Stage IV: 14 (18.42%)	76	56	42.42

proportion of missing values can reduce the effective sample size and increase estimation bias if observations with missing values are discarded through complete-case analysis. Therefore, Bayesian logistic-regression multiple imputation within the FCS framework was employed to preserve the available information. This approach also addresses the uncertainty arising from missing values.

The missing data mechanism in this study was analyzed by examining the relationships between the missing data indicators for each variable and the remaining observed variables. Chi-square tests were applied to categorical variables, whereas independent t-tests were used for the numerical variable. The results of the missingness analyses for the smoking status, family history, and lung cancer stage variables are presented in Table 3.

Table 3: Missingness analysis for variables with missing values

Missing Variable	Comparison Variable	Statistical Test	Test Statistic	p-value
Smoking Status	Lung Cancer Stage	Chi-Square	0.9442	0.3312
	Family History	Chi-Square	0.3828	0.5361
	Comorbidity	Chi-Square	0.0789	0.7776
	Sex	Chi-Square	0.2372	0.6262
	Age	Independent t-test	1.4708	0.1417
Family History	Lung Cancer Stage	Chi-Square	0.0000	1.0000
	Comorbidity	Chi-Square	0.2374	0.6261
	Sex	Chi-Square	0.4202	0.5168
	Smoking Status	Chi-Square	0.0690	0.7928
	Age	Independent t-test	1.1712	0.2541
Lung Cancer Stage	Family History	Chi-Square	0.0000	1.0000
	Sex	Chi-Square	0.7421	0.3890
	Smoking Status	Chi-Square	0.0785	0.7793
	Comorbidity	Chi-Square	1.0585	0.2946
	Age	Independent t-test	0.6834	0.5810

The analyses did not detect statistically significant associations between the missingness indicators and the examined observed variables ($p > 0.05$). These findings provide no evidence against an MCAR mechanism based on the variables examined. However, the absence of statistically significant associations does not establish MCAR and cannot distinguish between MAR and MNAR. Multiple imputation was subsequently conducted under an assumed MAR mechanism, conditional on the observed variables included in the imputation models [7]. Accordingly, Bayesian multiple imputation within the FCS framework was used to retain the available observations. This approach also addresses uncertainty associated with the missing values [1]. A separate missingness-indicator analysis was not performed for comorbidity because only two observations were missing, providing insufficient variation for a meaningful comparison.

The relatively high proportion of missing values, particularly for lung cancer stage (42.42%),

smoking status (28.03%), and family history (14.39%), suggests that removing observations with missing values would substantially reduce the effective sample size and may increase estimation bias. Therefore, the Bayesian logistic-regression imputation within the FCS framework was used in this study. This method preserves the available information. The uncertainty introduced by missing values is reflected through multiple imputations. In addition, the variability attributable to missing values is incorporated into the parameter estimation process.

In the FCS procedure, each incomplete variable was conditionally modeled using the remaining variables as predictors. All variables requiring imputation in this study were binary categorical variables. Therefore, they were imputed using Bayesian logistic regression. Age was fully observed and included as a continuous predictor. Sex was also fully observed and included as a binary predictor. A total of $m = 20$ completed datasets were generated through 20 iterations of the chained equations procedure [4, 24].

For each incomplete binary variable, the conditional probability of belonging to the event category was modeled as

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \tag{41}$$

Eq. (41) specifies the Bayesian logistic regression model used for each incomplete binary variable, where p_i denotes the probability that the i-th observation belongs to the event category, $\beta_0, \ldots, \beta_p$ denote the regression coefficients, and $X_{i1}, \ldots, X_{ip}$ denote the predictor variables. Posterior parameter draws were used to calculate the predictive probabilities, after which the missing binary values were generated from Equation (30).

This stochastic procedure reflects uncertainty in the estimated model parameters and imputed values. It also uses the multivariable relationships specified in the conditional imputation models. Iteration plots were examined to assess the stability of the chained equations procedure. For each incomplete binary variable, the event category was internally coded as 1. At each iteration, the proportion of originally missing observations assigned to this event category was recorded. This procedure was performed across all 20 iterations for each of the 20 imputation chains.

In Fig. 1, the thin blue lines represent the individual imputation chains, the thick red line represents the mean proportion across chains at each iteration, and the horizontal dotted line represents the overall mean across chains and iterations. The plots directly assess the variables that were imputed: smoking status, comorbidity, family history, and cancer stage. The event categories were non-smoker, absence of comorbidity, absence of family history, and Stage IV, respectively. Chains that remain intermixed and fluctuate without a persistent increasing or decreasing trend indicate a relatively stable iteration pattern.

The trace plots showed no persistent increasing or decreasing trends for smoking status, family history, or cancer stage. The mean proportions remained relatively stable across the 20 iterations. Comorbidity exhibited a discrete pattern, with chain-specific proportions taking values of 0, 0.5, or 1. This pattern occurred because only two comorbidity observations were missing. Changing the category assigned to one observation therefore changed the proportion by 0.5. Therefore, this discrete pattern was not by itself interpreted as evidence of non-convergence.

Numerical diagnostics for the posterior sampling procedure are presented in Table 4. As shown in Table 4, the mean posterior sampling acceptance rates ranged from 0.293 to 0.302. It remains within relatively narrow ranges across the four imputation models. All 400 updates for each incomplete variable were completed using posterior sampling. No empirical fallback was required in any of the 1,600 conditional-model updates in the primary analysis. The median minimum effective sample sizes ranged from 85.83 to 111.30. Although smaller minimum ESS values were observed in individual updates, the combined trace plot and numerical diagnostics did not provide clear evidence of systematic non-convergence. Nevertheless, the lower minimum ESS values and the discrete comorbidity pattern were interpreted cautiously.

Following multiple imputation, multivariable logistic regression was performed separately in each of the 20 completed datasets. Cancer stage was specified as the dependent variable,

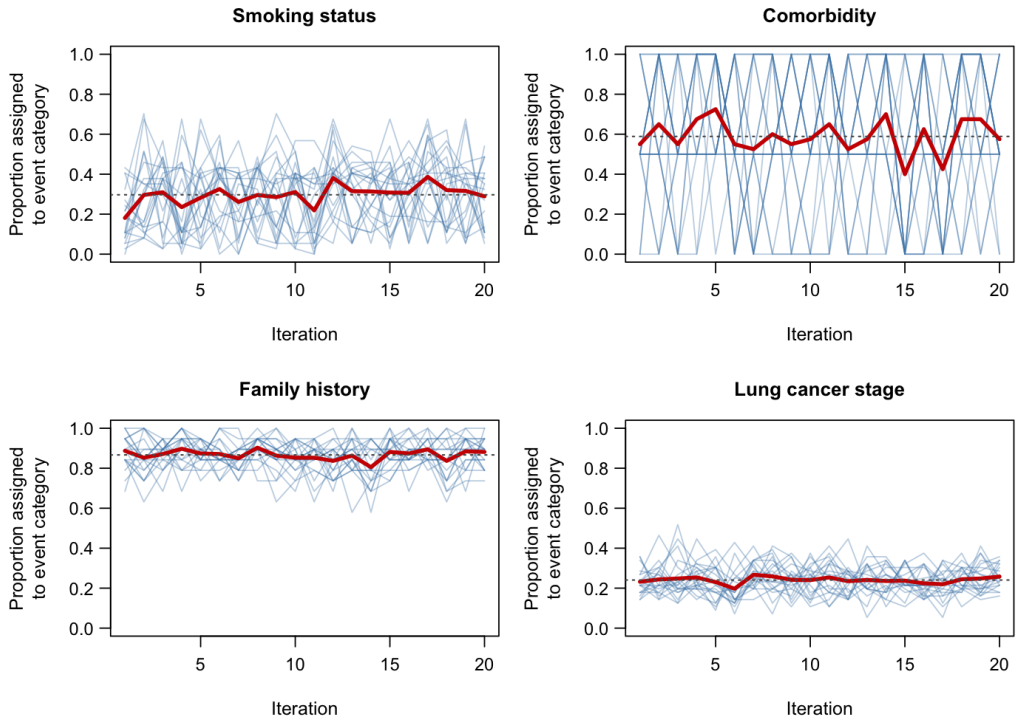


Fig. 1: Iteration plots of the proportions of originally missing observations assigned to the event category across 20 FCS iterations and 20 imputation chains. The event categories were non-smoker, absence of comorbidity, absence of family history, and Stage IV. Thin blue lines represent individual chains, thick red lines represent the mean across chains at each iteration, and horizontal dotted lines represent the overall means across chains and iterations.

with Stage III coded as 0 and used as the reference outcome, whereas Stage IV was coded as 1 and defined as the event. Age was included as a continuous predictor. Sex, smoking status, comorbidity, and family history were included as binary categorical predictors. For each binary predictor, category 1 was used as the reference category and category 2 was compared with category 1. The multivariable logistic regression model was specified as

$$\log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = \beta_0 + \beta_1 \mathrm{Age}_i + \beta_2 \mathrm{Sex}_{i,2} + \beta_3 \mathrm{Smoking}_{i,2} + \beta_4 \mathrm{Comorbidity}_{i,2} + \beta_5 \mathrm{History}_{i,2}, \tag{42}$$

Eq. (42) specifies the multivariable logistic regression model fitted separately to each completed dataset, where $P(Y_i = 1)$ denotes the probability that the i-th patient belongs to Stage IV. The coefficients $\beta_2, \ldots, \beta_5$ represent, respectively, the log-odds ratios for female versus male, non-smoker versus active smoker, absence versus presence of comorbidity, and absence versus presence of family history.

Because conventional maximum-likelihood logistic regression produced unstable estimates due to data separation in at least one completed dataset, mean bias-reduced logistic regression was used to obtain finite parameter estimates. A mean bias-reduced logistic regression model was fitted separately to each of the 20 completed datasets. The resulting regression coefficients and their corresponding variances were combined using Rubin's rules. The pooled effects were expressed as odds ratios (ORs), 95% confidence intervals (CIs), and p-values.

The stability of the pooled estimates with respect to the number of completed datasets was evaluated by repeating the analysis using $m = 50$ and $m = 100$. Both sensitivity analyses retained 20 FCS iterations and the primary prior specification. The primary prior specification consisted of an intercept prior standard deviation of 5 and a standardized-slope prior standard deviation of 2.5. Sensitivity to the slope-prior specification was evaluated separately by repeating

Table 4: Numerical diagnostics for posterior sampling within the imputation models

Variable	Mean AR	AR range	Median min. ESS	Minimum ESS	Fallback
Smoking status	0.301	0.275–0.323	85.83	26.92	0
Comorbidity	0.293	0.270–0.316	111.30	40.04	0
Family history	0.301	0.281–0.324	97.81	28.27	0
Cancer stage	0.302	0.276–0.325	92.47	25.94	0

Note: AR denotes acceptance rate, and ESS denotes effective sample size. Median minimum ESS is the median, across FCS updates, of the smallest coefficient-specific ESS within each conditional imputation model. Minimum ESS is the smallest value recorded across all updates. Each variable was updated 400 times. These updates consisted of 20 FCS iterations for each of the 20 imputations. In each update, the first 1,000 Metropolis–Hastings iterations were used as the initial sampling period. These iterations were not included in the posterior sample. Hence, 400 parameter values were recorded at intervals of three iterations. Fallback indicates that the conditional posterior model could not be updated. In this case, the missing values were sampled from the observed categories.

the analysis with $m = 20$ completed datasets. In the sensitivity analysis, the prior standard deviation for the standardized slope coefficients was increased from 2.5 to 5, while the intercept prior standard deviation was retained at 5. The pooled odds ratios were generally similar across the four configurations. The results are presented in Table 5. None of the examined predictors was statistically significant under any of the four configurations.

Table 5: Sensitivity analysis comparing pooled odds ratios across numbers of completed datasets and prior specifications

Predictor	Primary $m = 20$, SD 2.5	Prior SD 5 $m = 20$	$m = 50$ SD 2.5	$m = 100$ SD 2.5
Age, per one-year increase	0.97	0.97	0.97	0.98
Female vs. Male	0.52	0.57	0.55	0.52
Non-smoker vs. Smoker	0.23	0.21	0.22	0.23
No comorbidity vs. Comorbidity	1.00	0.94	1.07	1.03
No family history vs. Family history	1.77	1.65	1.83	1.83

Note: Values are pooled odds ratios from mean bias-reduced multivariable logistic regression models. The primary analysis used 20 completed datasets, an intercept-prior standard deviation of 5, and a standardized slope-prior standard deviation of 2.5. The wider slope-prior sensitivity analysis used a standardized slope-prior standard deviation of 5 while retaining the intercept-prior standard deviation of 5. The analyses using 50 and 100 completed datasets retained the primary prior specification. All configurations used 20 FCS iterations. The statistical conclusions remained unchanged across the four configurations, although the comorbidity estimate varied slightly around the null value of 1. OR = odds ratio; SD = standard deviation.

The posterior sampling diagnostics were generally consistent across the sensitivity analyses. Mean acceptance rates ranged from approximately 0.293 to 0.304. No empirical fallback was required in any configuration. Median minimum ESS values ranged from 74.92 to 111.30. The smoking status model under the wider slope prior had the lowest median minimum ESS and was therefore interpreted cautiously. For the j-th predictor, the pooled odds ratio was calculated as

$$\mathrm{OR}_j = \exp(\bar{\beta}_j), \tag{43}$$

Eq. (43) transforms the pooled regression coefficient $\bar{\beta}_j$ into the corresponding pooled odds ratio. The $\bar{\beta}_j$ denotes the regression coefficient obtained by combining the estimates from the 20 completed datasets in the primary analysis using Rubin's rules.

The regression coefficients and their corresponding variances from the 20 completed datasets were combined using Rubin's rules. The estimated mean bias-reduced multivariable logistic regression model for the primary analysis was

$$\begin{aligned}\log\left(\frac{\widehat{P}(Y_i=1)}{1-\widehat{P}(Y_i=1)}\right) &= 0.1068 - 0.0256\,\mathrm{Age}_i - 0.6579\,\mathrm{Sex}_{i,2} \\ &\quad - 1.4552\,\mathrm{Smoking}_{i,2} + 0.0041\,\mathrm{Comorbidity}_{i,2} + 0.5714\,\mathrm{History}_{i,2},\end{aligned} \tag{44}$$

Table 6: Mean bias-reduced multivariable logistic regression results after multiple imputation

Predictor	Odds ratio	95% CI	p-value
Age, per one-year increase	0.97	0.93–1.03	0.325
Female vs. Male	0.52	0.012–22.40	0.731
Non-smoker vs. Active-Smoker	0.23	0.009–5.95	0.377
No comorbidity vs. Comorbidity	1.00	0.30–3.31	0.995
No family history vs. Family history	1.77	0.26–12.18	0.560

Note: Odds ratios, confidence intervals, and p-values were obtained by pooling the regression coefficients and their variances across the completed datasets using Rubin's rules. Male, smoker, presence of comorbidity, and presence of family history were used as the reference categories. OR = odds ratio; CI = confidence interval.

Eq. (44) presents the estimated pooled multivariable logistic regression model for the primary analysis, where $\widehat{P}(Y_i = 1)$ denotes the estimated probability that the i-th patient belongs to Stage IV.

As shown in Table 6, age was not statistically significantly associated with Stage IV disease (OR = 0.97, 95% CI: 0.93–1.03; $p = 0.325$). Based on the unrounded coefficient, each one-year increase in age was associated with an estimated 2.5% decrease in the odds of Stage IV disease. However, the CI included 1, indicating that the association was not statistically significant. Female patients had lower estimated odds of Stage IV than male patients (OR = 0.52, 95% CI: 0.012–22.40; $p = 0.731$). Although the point estimate indicated lower odds, the CI was extremely wide and included 1. Therefore, the analysis did not provide sufficient statistical evidence of an association between sex and Stage IV disease.

Non-smokers had lower estimated odds of Stage IV than active smokers (OR = 0.23, 95% CI: 0.009–5.95; $p = 0.377$). However, the confidence interval was wide and included 1, indicating substantial uncertainty and no statistically significant association between smoking status and Stage IV disease. No statistically significant associations were observed for absence of comorbidity (OR = 1.00, 95% CI: 0.30–3.31; $p = 0.995$) or absence of family history (OR = 1.77, 95% CI: 0.26–12.18; $p = 0.560$). Overall, none of the examined predictors was statistically significantly associated with Stage IV disease at the 5% significance level.

The wide confidence intervals, particularly for sex and smoking status, indicated substantial uncertainty and limited precision. Although mean bias-reduced logistic regression produced finite estimates in the presence of data separation, the estimates remained limited in precision. This limitation may be attributed to the small sample size, sparse category combinations, and high proportion of missing outcome data [25]. Therefore, the point estimates should be interpreted cautiously.

The sensitivity analyses supported the stability of the overall statistical conclusion. Increasing the number of completed datasets from 20 to 50 and 100 produced only modest changes in the pooled odds ratios. Increasing the prior standard deviation for the standardized slope coefficients from 2.5 to 5 also did not change the conclusion regarding statistical significance. Nevertheless, these sensitivity analyses do not establish the absence of clinically important associations. This is because the confidence intervals remained wide across the examined model specifications.

An additional limitation is the high proportion of missing cancer stage data, which affected 42.42% of participants (56 of 132). Multiple imputation was performed under a MAR assumption, conditional on the observed variables included in the imputation models. Because this assumption cannot be verified using the observed data alone, bias due to potentially MNAR outcome data cannot be excluded. Analyses using 20, 50, and 100 completed datasets and alternative prior specifications produced generally similar pooled estimates and the same conclusions regarding statistical significance. However, this stability does not eliminate the uncertainty arising from the high proportion of missing outcome data. The findings should therefore be interpreted cautiously.

4. Conclusion

The missingness-indicator analyses detected no statistically significant associations with the observed variables. These findings provided no evidence against MCAR for the variables examined, but did not establish MCAR or distinguish MAR from MNAR. Bayesian multiple imputation was therefore performed under an assumed MAR mechanism. The iteration plots and numerical diagnostics showed no clear evidence of systematic non-convergence, although the diagnostics should be interpreted cautiously because of the limited data available for some variables.

The pooled mean bias-reduced logistic regression did not identify statistically significant associations between Stage IV disease and age, sex, smoking status, comorbidity, or family history. However, the wide confidence intervals indicated substantial uncertainty. The absence of statistically significant associations should therefore not be interpreted as evidence that clinically important associations were absent. Sensitivity analyses using 50 and 100 completed datasets and a wider slope prior produced materially similar estimates and did not change the overall statistical conclusion.

The small sample size, sparse category combinations, and 42.42% missingness in cancer stage substantially limited the precision. These limitations increased the dependence of the results on the imputation model and the assumed MAR mechanism. Therefore, the findings should be interpreted cautiously. They should be regarded as exploratory rather than confirmatory. Future studies should use larger samples with more complete outcome data. Sensitivity analyses under alternative missing-data assumptions, particularly plausible MNAR scenarios, should also be conducted to evaluate the robustness of the estimated associations.

CRediT Authorship Contribution Statement

Usman Pagalay: Supervision, Investigation, Project Administration, Conceptualization, Funding Acquisition. **Ria Dhea L.N. Karisma:** Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing–Original Draft Preparation. **Hawzah Saadati:** Data Curation, Software, Writing–Review & Editing, Project Administration, Methodology. **Evawatialisah:** Validation, Writing–Review & Editing. **Ratnaning Palupi:** Project Administration, Methodology. **Heni Widayani:** Methodology, Validation, Software, Visualization.

Declaration of Generative AI and AI-assisted technologies

Generative AI or AI-assisted technologies were used during the preparation of the manuscript, particularly Grammarly. Grammarly was used for grammar checking, proofreading, and improvements to language clarity. Data analysis, statistical modeling, interpretation of results, or generation of scientific findings were not performed using AI. The results and discussions were entirely developed based on the findings by the authors themselves.

Declaration of Competing Interest

The authors declare no competing interests.

Ethics Approval and Consent to Participate

This study was approved by the Health Research Ethics Committee of the Faculty of Medicine and Health Sciences, Universitas Islam Negeri Maulana Malik Ibrahim Malang, with ethical approval number 122/30/KEPK-FKIK/X/2025. This study used secondary data derived from patient medical records at Karsa Husada Hospital, Batu. Patient names and personally identifiable information were excluded from the dataset. Only relevant clinical characteristics were included for analysis, and the confidentiality of patient information was maintained throughout the study. The requirement for informed consent was waived by the Ethics Committee due to the

retrospective nature of the study and the use of secondary data without personally identifiable information.

Funding and Acknowledgments

This research was funded by Litapdimas through the Institute for Research and Community Service (LP2M), Universitas Islam Negeri Maulana Malik Ibrahim Malang.

The authors would like to thank Karsa Husada Hospital, Batu, for providing access to the clinical data and supporting this research. The authors also acknowledge the Faculty of Science and Technology and the Faculty of Medicine and Health Sciences, Universitas Islam Negeri Maulana Malik Ibrahim Malang, for providing academic support and a conducive research environment. The authors are grateful to the Head of the Ethical Clearance Unit and the laboratory staff for their support and technical assistance.

Data and Code Availability

The data used in this study cannot be shared publicly because they contain clinical information obtained from patient medical records. Data access is restricted to protect patient confidentiality and comply with ethical approval requirements. Access to the data may be considered upon reasonable request and with permission from Karsa Husada Hospital, Batu, and the relevant ethics committee. The code used for data analysis is available from the corresponding author upon reasonable request.

References

[1] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. 2nd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, 2002. DOI: 10.1002/9781119013563.

[2] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. 3rd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, 2019. DOI: 10.1002/9781119482260.

[3] Peter C. Austin, Ian R. White, Douglas S. Lee, and Stef van Buuren. "Missing Data in Clinical Research: A Tutorial on Multiple Imputation". In: *Canadian Journal of Cardiology* 37.9 (2021), pp. 1322–1331. DOI: 10.1016/j.cjca.2020.11.010.

[4] Ian R. White, Patrick Royston, and Angela M. Wood. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice". In: *Statistics in Medicine* 30.4 (2011), pp. 377–399. DOI: 10.1002/sim.4067.

[5] Stef van Buuren and Karin Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67. DOI: 10.18637/jss.v045.i03.

[6] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1987. DOI: 10.1002/9780470316696.

[7] Donald B. Rubin. "Inference and Missing Data". In: *Biometrika* 63.3 (1976), pp. 581–592. DOI: 10.2307/2335739.

[8] Hamish Campbell, Tim P. Morris, and Paul Gustafson. "A Fully-Integrated Bayesian Approach for the Imputation and Analysis of Derived Outcome Variables With Missingness". In: *Statistics in Medicine* 45.1–2 (2026), e70383. DOI: 10.1002/sim.70383.

[9] Stef van Buuren. *Flexible Imputation of Missing Data*. 2nd ed. Chapman and Hall/CRC, 2018. DOI: 10.1201/9780429492259.

[10] Irma Eldiyana, Elah Nurlaelah, and Nar Herrhyanto. "Estimasi Missing Data dengan Metode Multivariate Imputation by Chained Equations (MICE) untuk Membentuk Persamaan Regresi Linear Berganda". In: *Eureka Matika* 9.1 (2021). DOI: 10.17509/jem.v8i1.25750.

[11] Cindy Pramudita, Naura Aiswarya Asri Wicaksono, Anjar Kharisma Hartiyah, Annisa Maisaroh, and Iqbal Kharisudi. "Menangani Missing Value dengan Metode Imputasi Menggunakan Package MICE di R". In: *Prosiding Seminar Nasional Matematika PRISMA*. Vol. 8. Universitas Negeri Semarang, 2025, pp. 399–406.

[12] Mukarromah, Shantika Martha, and Ilhamsyah. "Perbandingan Imputasi Missing Data Menggunakan Metode Mean dan Metode Algoritma K-Means". In: *Buletin Ilmiah Matematika, Statistika dan Terapannya (Bimaster)* 4.3 (2015), pp. 305–312. DOI: 10.26418/bbimst.v4i03.12425.

[13] Anisa Widianti and Irfan Pratama. "Penanganan Missing Values dan Prediksi Data Timbunan Sampah Berbasis Machine Learning". In: *RABIT: Jurnal Teknologi dan Sistem Informasi Univrab* 9.2 (2024), pp. 242–251. DOI: 10.36341/rabit.v9i2.4789.

[14] Alan Agresti. *An Introduction to Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons, 2019.

[15] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97.

[16] Charles J. Geyer. "Introduction to Markov Chain Monte Carlo". In: *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011, pp. 3–48. DOI: 10.1201/b10905-2.

[17] Gareth O. Roberts and Jeffrey S. Rosenthal. "Optimal Scaling for Various Metropolis–Hastings Algorithms". In: *Statistical Science* 16.4 (2001), pp. 351–367. DOI: 10.1214/ss/1015346320.

[18] Brian D. Ripley. *Stochastic Simulation*. New York: John Wiley & Sons, 1987.

[19] James E. Gentle. *Computational Statistics*. 1st ed. New York: Springer, 2009. DOI: 10.1007/978-0-387-98144-4.

[20] Stan Development Team. *Stan Reference Manual: Effective Sample Size*. Version 2.19. Stan Development Team. 2019. https://mc-stan.org/docs/2_19/reference-manual/effective-sample-size-section.html.

[21] Ofer Harel, Emily M. Mitchell, Neil J. Perkins, Stephen R. Cole, Eric J. Tchetgen Tchetgen, BaoLuo Sun, and Enrique F. Schisterman. "Multiple Imputation for Incomplete Data in Epidemiologic Studies". In: *American Journal of Epidemiology* 187.3 (2018), pp. 576–584. DOI: 10.1093/aje/kwx349.

[22] S. Harahap, N. Sutandyo, C. Martin R., and H. Shatri. "Perbandingan Rejimen Kemoterapi Cisplatin Etoposide dengan Cisplatin-Docetaxel dalam Hal Kesintasan 2 Tahun dan Progression-Free Survival Pasien Kanker Paru Stadium Lanjut Jenis Non-Small Cell". In: *Jurnal Penyakit Dalam Indonesia* 3.2 (2016), pp. 67–73. DOI: 10.7454/jpdi.v3i2.11.

[23] American Lung Association. *The Global State of Lung Cancer: A Report from the Lung Cancer Research Foundation*. Available online. 2023. https://www.lung.org/getmedia/186786b6-18c3-46a9-a7e7-810f3ce4deda/SOLC-2023-Print-Report.pdf.

[24] John W Graham, Allison E Olchowski, and Tamika D Gilreath. "How many imputations are really needed? Some practical clarifications of multiple imputation theory". In: *Prevention Science* 8.3 (2007), pp. 206–213. DOI: 10.1007/s11121-007-0070-9.

[25] Masahiko Gosho, Tomohiro Ohigashi, Kengo Nagashima, Yuri Ito, and Kazushi Maruo. "Bias in Odds Ratios From Logistic Regression Methods With Sparse Data Sets". In: *Journal of Epidemiology* 33.6 (2023), pp. 265–275. DOI: 10.2188/jea.JE20210089.