Predictive Modeling of Chronic Kidney Disease and Analysis of Key Risk Factors

Stony Brook University, AMS 572- Fall 2025
Alan Rodriguez, alan.rodriguez@stonybrook.edu
Xiaoyan Lin, xiaoyan.lin@stonybrook.edu
Yeonbi Han, yeonbi.han@stonybrook.edu

Abstract

This project analyzes a chronic kidney disease (CKD) dataset of 400 patients and 25 variables from the UCI Machine Learning Repository. Two statistical hypotheses (H_1 and H_2) were evaluated, and the impact of missing-data mechanisms—MCAR and MNAR—was assessed to determine how inference and prediction behave under incomplete information.

Hypothesis 1 (H_1) examined whether mean blood pressure differs by appetite status. Using an F-test and Welch's t-test, we found that patients with poor appetite exhibit significantly higher blood pressure ($p < 0.01$). This association reflects correlation rather than causality but aligns with prior observations that appetite changes often accompany worsening clinical conditions. Under MCAR simulations, statistical significance disappeared once missingness exceeded 30%, demonstrating the sensitivity of univariate inference to information loss.

Hypothesis 2 (H_2) involved developing a multivariate logistic regression model to predict CKD diagnosis. After dummy encoding, KNN imputation, VIF screening, and backward stepwise selection, albumin (al) was removed due to quasi-complete separation. The final model retained three predictors—bgr, hemo, and sg—and achieved 96.3% accuracy with an AUC of 0.993. Under both MCAR and MNAR simulations, predictive accuracy and AUC remained consistently high even with 50% missingness, indicating that multivariate structure effectively mitigates data loss. Overall, univariate testing in H_1 proved fragile, while the multivariate model in H_2 remained robust and reliable.

Key words

Chronic Kidney Disease (CKD); Blood Pressure (bp); Appetite; t-test; F-test; Hypothesis Testing; Logistic Regression; VIF and Stepwise Selection; Risk Factor Analysis; KNN Imputation; MCAR/MNAR Simulation; Model Robustness Evaluation; Quasi-complete Separation; Odds Ratio (OR); Sensitivity and Specificity; Confusion Matrix; ROC Curve; AUC; Missing Data Mechanisms; Model Stability; Dummy Encoding; Reproducible Analysis

Table of Contents

1. Introduction

1.1 Background
1.2 Research Objectives

2. Data and Exploratory Analysis

2.1 Data Description and Variables
2.2 Data Preprocessing and Missingness Summary
2.3 Key EDA Results (Summary Statistics & Correlation Heatmap)
2.4 Classification of Missingness Mechanisms (MCAR vs MNAR)

3. Methodology

3.1 Hypothesis 1: Appetite vs. Blood Pressure
- Listwise deletion, F-test, Welch T-test

3.2 Hypothesis 2: Logistic Regression Model
- KNN imputation, VIF screening
- Stepwise AIC and removal of albumin(al)
- Final 3-variable model (bgr, hemo, sg)

3.3 Missingness Simulations (MCAR / MNAR)
- Simulation design and re-analysis plan

4. Results

4.1 Results of H_1: Appetite–Blood Pressure Comparison
4.2 Results of H_2: Logistic Regression Model for CKD Prediction
4.3 Robustness of H_1 Under MCAR and MNAR Simulations
4.4 Robustness of H_2 Under MCAR and MNAR Simulations

5. Discussion and Conclusion

5.1 Interpretation of Findings
5.2 Impact of Missingness
5.3 Limitations and Future Work

1. Introduction

1.1. Background and Motivation

Chronic Kidney Disease (CKD) is a multifactorial condition influenced by diverse clinical indicators. Real-world healthcare data typically include heterogeneous variables, irregular measurement scales, and substantial amounts of missingness. Although these characteristics make statistical analysis more

challenging, they also provide an ideal environment for evaluating the robustness of analytical methods under realistic data conditions.

Using a dataset of 400 patients, this project aims to examine CKD-related clinical markers and assess how well statistical inference and predictive modeling perform in the presence of complex and incomplete data. Through this process, we strengthen advanced analytical skills in managing noisy, imperfect datasets while generating statistically grounded insights that can support early detection and clinical decision-making for CKD.

1.2. Project Objectives

The core objectives of this project are structured into four major phases:

1. Basic Data Understanding and Missingness Identification – To understand the structure and variable types of the CKD dataset, perform data cleaning, and examine the fundamental mechanisms of missing data generation.
2. Hypothesis Formulation and Evaluation – To establish non-trivial hypotheses regarding the relationship between the major clinical variables, appetite (appet) and blood pressure (bp), and to statistically evaluate them using the T-test (unequal variance).
3. Advanced Predictive Model Development – To develop a Logistic Regression model to evaluate whether the selected variables exhibit strong associations with the target outcome (CKD diagnosis).
4. In-depth Missing Data Impact Analysis – To conduct MCAR and MNAR simulations and re-evaluate both analyses after KNN imputation in order to assess the robustness of the analytical conclusions under different missing-data mechanisms.

2. Data and Exploratory Analysis

2.1 Data Description and Variables

The dataset used in this project contains clinical information related to Chronic Kidney Disease (CKD). Originally compiled by L. Rubini, P. Soundarapandian, and P. Eswaran, it was later made publicly available through the UC Irvine Machine Learning Repository (UCI ML Repository) (“Chronic Kidney Disease”). The dataset consists of 400 patient records ($n = 400$) and 25 clinical variables ($p = 25$), satisfying the AMS 572 requirement of $n \geq 50$ and $p \geq 20$.

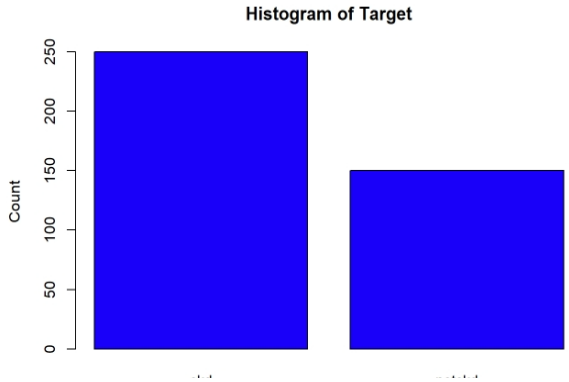


Figure 1 Histogram of Target

The target variable indicates CKD status and is coded as ckd or notckd. To examine the outcome distribution, a histogram was generated (Figure 1). The dataset contains 250 CKD cases and 150 non-CKD cases, indicating a moderate class imbalance with CKD representing approximately 62.5% of the total sample.

(1) Numerical Data: The dataset contains 14 numerical variables, including key physiological and biochemical measurements such as: *blood pressure (bp), specific gravity (sg), albumin (al), blood glucose*

random (bgr), blood urea (bu), serum creatinine (sc), hemoglobin (hemo), packed cell volume (pcv), white blood cell count (wbcc), red blood cell count (rbcc), sodium, and potassium. These numeric variables represent continuous clinical markers relevant to CKD diagnosis and will later be used in statistical testing and logistic regression modeling.

(2) Categorical Data: The dataset includes 11 categorical variables, all of which are binary (excluding missing values):*rbc, pc, pcc, ba, hypertension (htn), diabetes mellitus (dm), coronary artery disease (cad), appetite (appet), pedal edema (pe), and anemia (ane).* Their binary structure makes them straightforward to encode using dummy or binary encoding during preprocessing.

2.2 Data Preprocessing and Missingness Summary

The CKD dataset contains a mixture of numeric and categorical variables, with missing values present across most features. Because missingness can substantially influence both hypothesis testing (H_1) and logistic regression modeling (H_2), a structured preprocessing workflow was implemented prior to statistical analysis. This section summarizes the standardization of missing values, missingness patterns, and the rationale for different preprocessing strategies applied to H_1 and H_2.

(1) Standardization of Missing Values

The raw dataset included several irregular encodings of missingness, such as "?", "\t?", and whitespace-embedded strings. All such entries were uniformly converted to NA, ensuring:

- compatibility with R's KNN-based imputation functions,
- consistent handling of missingness across both numeric and categorical features, and
- a clean and analyzable data structure for downstream analysis.

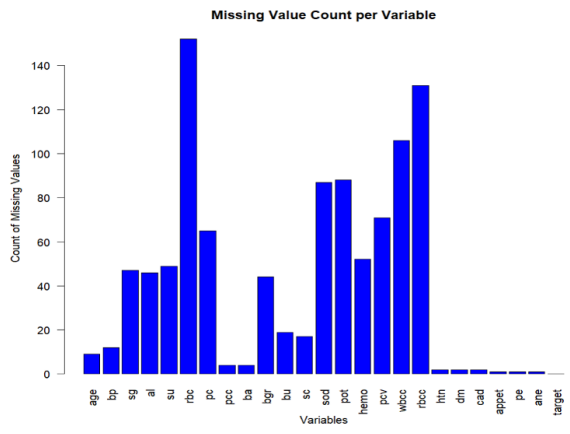
This standardization step was critical for maintaining consistency throughout the preprocessing pipeline.

(2) Missingness Pattern and Variable-Level Summary

A full missingness summary was generated for all 25 variables, documenting total counts, missing counts, and missing percentages. Key observations include:

- Most variables (e.g., age, bp, bu, sc) exhibited less than 10% missingness.
  Several features showed substantial missingness:
  rbc (38%), rbcc (32.75%), wbcc (26.5%), pcv (17.75%).
- The target variable contained no missing values, enabling supervised learning without label imputation.

These results indicate that listwise deletion would remove a large portion of the dataset, especially formultivariable modeling.



	Total_Count	Missing_Count	Unique_Count	Missing_Percentage
age	400	9	77	2.25
bp	400	12	11	3.00
sg	400	47	6	11.75
al	400	46	7	11.50
su	400	49	7	12.25
rbc	400	152	3	38.00
pc	400	65	3	16.25
pcc	400	4	3	1.00
ba	400	4	3	1.00
bgr	400	44	147	11.00
bu	400	19	119	4.75
sc	400	17	85	4.25
sod	400	87	35	21.75
pot	400	88	41	22.00
hemo	400	52	116	13.00
pcv	400	71	43	17.75
wbcc	400	106	90	26.50
rbcc	400	131	46	32.75
htn	400	2	3	0.50
dm	400	2	3	0.50
cad	400	2	3	0.50
appet	400	1	3	0.25
pe	400	1	3	0.25
ane	400	1	3	0.25
target	400	0	2	0.00

Figure 2 Missing Value Count per Variable

This visualization clearly highlights the heterogeneity of missingness across variables and provides motivation for differentiated preprocessing strategies. (Figure 2)

(3) Missingness Strategy by Analysis Objective: Because different analyses rely on different subsets of variables, missing-data handling was tailored to each step:

- H_1 (Appetite vs Blood Pressure)
  Since bp and appet had <5% missingness, → listwise deletion was appropriate and avoided unnecessary imputation.
- H_2 (Logistic Regression Model)
  Multivariable modeling required retaining sufficient observations.
  → KNN imputation (k = 5) was applied to numeric and categorical features.
  → Variables with >30% missingness (rbc, rbcc, wbcc) were excluded to prevent unstable imputations.

These high-level preprocessing decisions ensured that the cleaned dataset remained suitable for reliable hypothesis testing, logistic regression modeling, and later MCAR/MNAR robustness simulations.

2.3 Key EDA Results (Summary Statistics & Correlation Heatmap)

This section presents the exploratory data analysis (EDA) performed to understand the distributional properties of the CKD dataset and to examine structural relationships among the variables. Two main components were analyzed: (1) summary statistics for all variables and (2) correlation patterns among numeric features. These results provide important context for subsequent hypothesis testing and logistic regression modeling.

(1) Summary Statistics Figure 3 summarizes the numeric and categorical variables in the CKD dataset, including their ranges, quartiles, means, and missing counts. Numeric variables show broad ranges with clinically meaningful extremes (e.g., bp: 50–180, bgr: 22–490, bu: 1.5–391, sc: 0.4–76), indicating considerable heterogeneity and potential outliers. Several features exhibit high missingness—rbc (38%),

rbcc (32.75%), wbcc (26.5%), pcv (17.75%)—and were therefore excluded from the logistic regression model due to instability after imputation. All categorical variables (e.g., rbc, pc, pcc, ba, htn, dm, cad, appet, pe, ane) are binary, which simplifies encoding in subsequent modeling.

```
# Data structure
summary(df)
```

```
##          age               bp               sg               al
##   Min.   : 2.00    Min.   : 50.00    Min.   :1.005    Min.   :0.000
##   1st Qu.:42.00    1st Qu.: 70.00    1st Qu.:1.010    1st Qu.:0.000
##   Median :55.00    Median : 80.00    Median :1.020    Median :0.000
##   Mean   :51.48    Mean   : 76.47    Mean   :1.017    Mean   :1.017
##   3rd Qu.:64.50    3rd Qu.: 80.00    3rd Qu.:1.020    3rd Qu.:2.000
##   Max.   :90.00    Max.   :180.00    Max.   :1.025    Max.   :5.000
##   NA's   :9        NA's   :12        NA's   :47       NA's   :46
##          su              rbc                pc                pcc
##   Min.   :0.0000   Length:400         Length:400         Length:400
##   1st Qu.:0.0000   Class :character   Class :character   Class :character
##   Median :0.0000   Mode  :character   Mode  :character   Mode  :character
##   Mean   :0.4501
##   3rd Qu.:0.0000
##   Max.   :5.0000
##   NA's   :49
##          ba                 bgr               bu               sc
##   Length:400         Min.   :  22    Min.   :  1.50    Min.   : 0.400
##   Class :character   1st Qu.:  99    1st Qu.: 27.00    1st Qu.: 0.900
##   Mode  :character   Median : 121    Median : 42.00    Median : 1.300
##                      Mean   : 148    Mean   : 57.43    Mean   : 3.072
##                      3rd Qu.: 163    3rd Qu.: 66.00    3rd Qu.: 2.800
##                      Max.   : 490    Max.   :391.00    Max.   :76.000
##                      NA's   :44      NA's   :19        NA's   :17
##          sod               pot              hemo             pcv
##   Min.   :  4.5     Min.   : 2.500    Min.   : 3.10    Min.   : 9.00
##   1st Qu.:135.0     1st Qu.: 3.800    1st Qu.:10.30    1st Qu.:32.00
##   Median :138.0     Median : 4.400    Median :12.65    Median :40.00
##   Mean   :137.5     Mean   : 4.627    Mean   :12.53    Mean   :38.88
##   3rd Qu.:142.0     3rd Qu.: 4.900    3rd Qu.:15.00    3rd Qu.:45.00
##   Max.   :163.0     Max.   :47.000    Max.   :17.80    Max.   :54.00
##   NA's   :87        NA's   :88        NA's   :52       NA's   :71
##         wbcc              rbcc              htn              dm
##   Min.   : 2200     Min.   :2.100    Length:400         Length:400
##   1st Qu.: 6500     1st Qu.:3.900    Class :character   Class :character
##   Median : 8000     Median :4.800    Mode  :character   Mode  :character
##   Mean   : 8406     Mean   :4.707
##   3rd Qu.: 9800     3rd Qu.:5.400
##   Max.   :26400     Max.   :8.000
##   NA's   :106       NA's   :131
##         cad              appet               pe                ane
##   Length:400         Length:400         Length:400         Length:400
##   Class :character   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

Figure 3 Summary of Statistics

(2) Correlation Analysis (Correlation Heatmap)

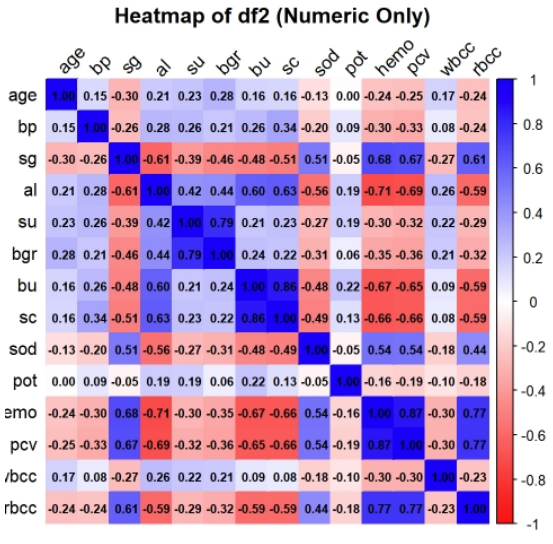


Figure 4 Numerical Variables Correlation Heatmap

To assess the relationships among numeric variables, a Pearson correlation matrix was computed and visualized using a heatmap (Figure 4). The heatmap identifies meaningful clinical patterns and potential multicollinearity issues.

a. Strong Positive Correlations: Two pairs of variables exhibit particularly strong correlations:

- hemoglobin (hemo) ↔ packed cell volume (pcv): $r \approx 0.87$
- blood urea (bu) ↔ serum creatinine (sc): $r \approx 0.86$

Both pairs align with known physiological relationships. For example, hemoglobin and PCV jointly reflect hematologic status, while blood urea and creatinine are standard kidney function indicators.

b. Implications for Multicollinearity: These strong associations suggest the potential for multicollinearity in predictive modeling. Consequently, VIF analysis was performed in later sections to ensure stable coefficient estimation.

c. General Correlation Structure: Most variables exhibit moderate or weak correlations, indicating that many predictors may contribute independent information to the logistic regression model.

2.4 Classification of Missingness Mechanisms (MCAR vs MNAR)

This section investigates whether missing values in the CKD dataset follow a Missing Completely at Random (MCAR) or Missing Not at Random (MNAR) mechanism. The classification is based on (1) variable-level missingness rates and (2) differences in missingness between CKD and non-CKD groups (Δ Difference).

(1) Variable-Level Missingness Patterns Missingness levels varied substantially across variables. Several laboratory variables showed high missingness (20–50%), such as:

- rbc: 38%, rbcc: 32.75%, wbcc: 26.5%, pcv: 17.75%

Meanwhile, variables such as age, bp, bu, and sc had less than 5–10% missingness. This heterogeneity indicates that the dataset is unlikely to follow a purely MCAR mechanism.

(2) Missingness by CKD Status → Evidence of MNAR

We compared missingness rates between CKD and non-CKD groups (Figure 5). Several variables displayed extremely large differences

Variable	Missing (CKD)	Missing (non-CKD)	Δ Difference
rbc	57.2%	6.0%	+51.2%
rbcc	49.6%	4.7%	+44.9%
wbcc	39.6%	4.7%	+34.9%
pcv	26.8%	2.7%	+24.1%
sod	32.8%	3.3%	+29.5%
pot	33.2%	3.3%	+29.9%

```
colnames(missing_compare) <- c(
  "Missing_in_CKD (%)",
  "Missing_in_notCKD (%)",
  "Δ Difference (%)"
)

missing_compare
```

```
##        Missing_in_CKD (%) Missing_in_notCKD (%) Δ Difference (%)
## age                  3.2                  0.67              2.53
## bp                   4.0                  1.33              2.67
## sg                  16.8                  3.33             13.47
## al                  16.4                  3.33             13.07
## su                  17.6                  3.33             14.27
## rbc                 57.2                  6.00             51.20
## pc                  22.4                  6.00             16.40
## pcc                  0.0                  2.67             -2.67
## ba                   0.0                  2.67             -2.67
## bgr                 15.2                  4.00             11.20
## bu                   5.2                  4.00              1.20
## sc                   4.8                  3.33              1.47
## sod                 32.8                  3.33             29.47
## pot                 33.2                  3.33             29.87
## hemo                18.4                  4.00             14.40
## pcv                 26.8                  2.67             24.13
## wbcc                39.6                  4.67             34.93
## rbcc                49.6                  4.67             44.93
## htn                  0.0                  1.33             -1.33
## dm                   0.0                  1.33             -1.33
## cad                  0.0                  1.33             -1.33
## appet                0.0                  0.67             -0.67
## pe                   0.0                  0.67             -0.67
## ane                  0.0                  0.67             -0.67
## target               0.0                  0.00              0.00
```

Figure 5 Missingnes rates between CKD and non-CKD groups

These large discrepancies strongly suggest that missingness is associated with disease severity or unobserved laboratory values, which is consistent with MNAR behavior.

(3) Variables Showing MCAR-Like Behavior

Several variables showed nearly identical missingness in CKD and non-CKD groups, suggesting MCAR-like randomness: age, bp, bu, sc, pcc, ba, htn, dm, cad, appet, pe, ane

These variables do not show systematic dependence on patient status.

(4) Final Conclusion: Hybrid Missingness Structure

Overall, the CKD dataset contains both MCAR and MNAR types of missingness:

- Some variables exhibit random (MCAR-like) missingness
- Key clinical variables show strong MNAR patterns linked to CKD severity

Thus, the CKD dataset presents a hybrid missingness structure, combining MCAR and MNAR behaviors.

This mixed structure motivated our subsequent strategy:

- using listwise deletion for H_1 (minimal missingness),
- applying KNN imputation for H_2, and
- conducting both MCAR and MNAR simulation scenarios to evaluate robustness.

3. Methodology

3.1 Hypothesis 1: Appetite vs. Blood Pressure (F-test and Welch T-test)

The objective of Hypothesis 1 (H_1) was to determine whether mean blood pressure (bp) differs between the two appetite categories (good vs. poor). Appetite (appet) is a binary categorical variable, while blood pressure is continuous; therefore, the appropriate statistical approach is an independent two-sample T-test comparing the means of a continuous outcome across two independent groups.

Because the Chi-square test requires both variables to be categorical, it is not suitable for this analysis. The continuous–categorical structure of (bp, appet) justifies the use of a T-test framework.

3.1.1 Missing Data Handling (Listwise Deletion)

Both bp and appet exhibited less than 5% missingness, and the missingness pattern did not indicate dependence on clinical characteristics. Since imputation was unnecessary and could introduce unwarranted assumptions, listwise deletion (dropna()) was applied. This resulted in minimal information loss and preserved the integrity of the statistical test.

3.1.2 Variance Testing and T-test Procedure

Before performing the T-test, an F-test was conducted to assess equality of variances between the two appetite groups.

(1) F-test for Equal Variances

- H_0: The variances of bp are equal between the “good” and “poor” appetite groups.
- H_a: The variances differ.

If the F-test indicates unequal variances ($p < 0.05$), the pooled-variance T-test is inappropriate. In such cases, the Welch T-test, which does not assume equal variances, provides a more reliable comparison.

(2) Welch T-test (Final Selected Test)

Due to unequal variances, the Welch T-test was used to evaluate the mean difference:

- H_0: μ_good = μ_poor
- H_a: μ_good ≠ μ_poor

This method is robust in settings where variances differ across groups, making it the correct choice for the bp–appet comparison.

3.2 Hypothesis 2 – Multivariate Logistic Regression Modeling

The goal of Hypothesis 2 is to identify key clinical predictors of Chronic Kidney Disease (CKD) and evaluate their combined effects using a multivariate logistic regression model. Unlike Hypothesis 1, CKD is influenced by multiple interacting renal, metabolic, and hematologic factors, making a multivariate modeling framework essential.

Step 1. Clinical Variable Selection

The original dataset contained 25 variables, but not all possessed clear clinical interpretability. Therefore, 12 clinically relevant predictors were selected based on prior nephrology literature and medical reasoning(Table 1). They were grouped into three functional domains:

Table 1 Selected relevant predictors

Category	Variables	Description
Renal Function	al, sg, bu, sc	Indicators of kidney filtration efficiency and urinary protein loss
Metabolic / Cardiovascular Load	age, bp, bgr, htn, dm	Reflect metabolic stress and comorbid conditions influencing CKD
Hematologic / Systemic Status	hemo, appet, ane	Capture anemia, appetite loss, and systemic disease burden

Variables with excessive missingness (> 30%) — such as rbc, rbcc, and wbcc — or redundant meaning (e.g., pcv, highly correlated with hemo, r ≈ 0.90) were excluded. Categorical predictors (appet, htn, dm, ane) were one-hot encoded, and the binary outcome target was coded as 1 for CKD and 0 for non-CKD.

Step 2. Missing Data Handling (KNN Imputation)

Because multivariate models depend on complete predictor matrices, listwise deletion would lead to serious sample loss and bias. To address this, missing values were imputed using K-Nearest Neighbors (KNN, k = 5) imputation, which estimates each missing entry from clinically similar patients.
This method preserves multivariate relationships among predictors and maintains the clinical covariance structure. After imputation, all missing cells were fully resolved, yielding a complete dataset for modeling.

Step 3. Multicollinearity Diagnosis (VIF Analysis)

Multicollinearity can inflate standard errors and destabilize coefficient estimates in logistic regression. To prevent this, a Variance Inflation Factor (VIF) analysis was conducted for all candidate predictors.

- A VIF threshold of > 10 was used to flag problematic variables.
- Although no variable exceeded this threshold, strong correlations identified in EDA (e.g., hemo–pcv, sc–bu) highlighted the necessity of this diagnostic step.
- After confirming acceptable VIF values, all 12 clinically selected variables were retained in the initial full model.

This screening ensured that the subsequent stepwise selection process would operate on a stable predictor set and avoid overfitting.

Step 4. Stepwise Variable Selection (AIC-Based Backward Selection)

To obtain a parsimonious and well-structured logistic regression model for CKD prediction, a Backward Stepwise Selection procedure was performed using the Akaike Information Criterion (AIC) as the optimization metric. The full model began with all twelve clinically selected variables. At each iteration, the algorithm evaluated whether removing one predictor would reduce the overall AIC value.

The guiding principles of this procedure were:

- Minimize AIC by balancing model fit and complexity
- Remove redundant or weak predictors
- Retain variables that meaningfully contribute to CKD discrimination

Through this iterative elimination, the algorithm identified a four-variable combination — bgr, hemo, sg, and al — as the lowest-AIC solution.

However, additional diagnostic checks revealed that the variable albumin (al) produced quasi-complete separation, which violates assumptions of maximum-likelihood estimation by inflating standard errors and destabilizing coefficient estimates. Because this violates the conditions required for reliable logistic regression inference, al was removed from the final model.

Step 5. Removal of Albumin (al) Due to Quasi-Complete Separation

Although albumin (al) was selected by AIC, further examination showed that al nearly perfectly separated CKD from non-CKD observations, producing:

- extremely large standard errors,
- unstable coefficient estimates,
- and repeated warnings from the glm algorithm.

This pattern is consistent with quasi-complete separation, where a predictor almost perfectly classifies the outcome category. Under this condition, the logistic regression likelihood fails to converge properly, and maximum-likelihood estimates become non-interpretable.

Because reliable coefficient estimation is central to the purpose of Hypothesis 2, al was excluded before final model estimation. Removing al ensured numerical stability and allowed standard inferential procedures to be applied correctly.

Step 6. Final Logistic Regression Model and Evaluation Framework

After addressing separation, the final model was constructed using the remaining three predictors: **bgr** (blood glucose random), **hemo** (hemoglobin) ,**sg** (specific gravity)

The model was estimated using standard logistic regression (glm, binomial family). To evaluate model performance, the dataset was divided into an 80/20 training–testing split. The evaluation relied on several widely accepted metrics, each chosen for its ability to assess a distinct dimension of predictive quality:

Table 2 Final Evaluation Metrics

Metric	What It Measures	Interpretation / Why It Matters
Accuracy	Proportion of correctly classified observations	Simple overall performance summary; may be sensitive to class imbalance
ROC Curve	Trade-off between Sensitivity (TPR) and 1–Specificity (FPR) across thresholds	Curves closer to the upper-left corner indicate stronger discriminative ability
AUC (Area Under ROC Curve)	Threshold-free summary of ROC performance	Higher AUC means better ability to distinguish CKD vs non-CKD patients
Pseudo-R^2	Model-fit measure analogous to linear R^2	Reflects how much variation in CKD status is explained (interpretation differs from linear R^2)

3.3 Missingness Simulation (MCAR / MNAR) – Incorporating the Hybrid Missingness Structure

Prior exploratory analysis revealed that the original CKD dataset exhibits a hybrid missingness pattern, containing both MCAR (Missing Completely at Random) and MNAR (Missing Not at Random) components.

Some variables (e.g., age, bp, bu, sc) showed near-random missingness, while others (e.g., rbc, rbcc, wbcc, sg, al) demonstrated substantial missing-rate differences between CKD and non-CKD groups, indicating strong MNAR behavior. Recognizing this mixed structure, the simulation was designed to explicitly incorporate both MCAR and MNAR mechanisms in order to assess the robustness of Hypothesis 1 (H_1) and Hypothesis 2 (H_2) under realistic missing-data conditions.

Step 1. Simulation Framework — Hybrid Missingness Awareness

The simulation followed two overarching goals:

1. Replicate the mixed MCAR/MNAR behavior observed in the original dataset
2. Evaluate the sensitivity of H_1 and H_2 to increasing levels of missingness

To accomplish this, we generated datasets under MCAR and MNAR conditions at multiple missingness levels (10%, 20%, 30%, 40%, 50%), where MNAR in 40%.

Step 2. MCAR Scenario — Completely Random Deletion

In the MCAR scenario:

- Missing values were introduced using uniform random sampling
- The probability of being missing was the same for all observations
- Missingness was applied uniformly across variables at fixed rates

This scenario serves as a "baseline," representing conditions where missingness is independent of patient characteristics or underlying variable values.

Step 3. MNAR Scenario — Replicating Original Dataset Patterns

For our first hypothesis, we assume that missingness is not at random (MNAR) and reflects realistic clinical behavior. Specifically, blood pressure values above 80 are assumed to be more likely missing, because in emergency situations physicians may not have enough time to record BP measurements. For the appetite variable, we assume that observations with good appetite are more likely to be missing, as patients without appetite problems may consider this question unimportant and choose not to answer it.

For our second hypothesis, The MNAR scenario was constructed to reflect the non-random missingness patterns observed in the original CKD data. In the real dataset:

- CKD patients had dramatically higher missingness in
  rbc, rbcc, wbcc, sg, and al
- Missingness depended strongly on the outcome label (CKD vs. non-CKD)

The MNAR simulation therefore:

- Assigned higher missingness probability to CKD patients
- Concentrated missingness in variables that showed the largest group differences
- Applied lower missingness probability to non-CKD patients

This approach models situations in which missingness is related to disease severity or unobserved clinical states — a common occurrence in medical datasets.

Step 4. Missing-Data Handling by Analysis Type

Each simulated dataset was processed using the same strategy applied to the original CKD dataset:

H_1: Appetite vs. Blood Pressure

- Only two variables are involved
- Low missingness → Listwise deletion (dropna)
  This preserves simplicity and avoids unnecessary imputation.

H_2: Logistic Regression

- Multivariable analysis requires maximal data retention → KNN imputation (k = 5)

- Variables with extremely high missingness remained excluded
  This ensures consistency with the preprocessing used in the primary analysis.

Step 5. Re-analysis Under MCAR and MNAR Missingness

For each scenario (MCAR and MNAR) and each missingness level (10–50%), the following steps were repeated:

1. Generate missingness according to the scenario
2. Use KNN imputation to fill in all the missing values
3. Re-assess H_1 using F-test → Welch t-test
4. Re-fit the final logistic model (bgr, hemo, sg) for H_2
5. Examine model stability, coefficient behavior, and performance metrics

The Methodology section describes the procedure only; numerical results and interpretations are presented in the Results Sections 4.3 and 4.4.

4. Results

4.1 H_1 Results (T-test/F-test)

For our first hypothesis, as we mentioned in the methodology before performing the independent samples T-test, it is essential to assess whether the two groups—patients with good appetite (bp_good) and those with poor appetite (bp_poor)—have equal variances in their blood pressure measurements. The F-test for equality of variances was conducted to evaluate this assumption.

The null hypothesis (H_0) states that the variances of the two groups are equal, while the alternative hypothesis (H_1) states that the variances are not equal. The F-test produced a test statistic of F = 0.4355 with a p-value = 4.228e-07. Since the p-value is far below the significance level ($\alpha = 0.05$), we reject the null hypothesis. Meaning that there is a statistically significant difference between the variances of blood pressure in the good appetite and poor appetite groups. This indicates that the assumption of equal variances is violated, and therefore, the Welch's T-test (set equal variances = False) should be applied in subsequent analysis.

```
## --- F-test for Equality of Variances ---
```

```
f_test_result <- var.test(bp_good, bp_poor)

cat("F-statistic:", round(f_test_result$statistic, 4), "\n")
```

```
## F-statistic: 0.4355
```

```
cat("P-value:", formatC(f_test_result$p.value, format = "e", digits = 3), "\n")
```

```
## P-value: 4.228e-07
```

```
equal_var <- f_test_result$p.value > alpha
cat("Variance equality assumed:", equal_var, "\n\n")
```

```
## Variance equality assumed: FALSE
```

Figure 6 H1: F-test Result

Following the F-test(Figure 6), which indicated unequal variances between groups, the Welch's T-test was performed to compare the mean blood pressure (bp) between patients with good appetite and those with poor appetite.

The null hypothesis (H_0) states that there is no significant difference in mean blood pressure between the two appetite groups, while the alternative hypothesis (H_1) proposes that a significant difference exists.

```
##
## H1 Conclusion (bp vs. appetite):
## The t-test indicates a statistically significant association between appetite status and blood
pressure.
## * T-statistic: -2.8087
## * P-value: 0.006
```

Figure 7 H1:T-test Result

The test produced a t-statistic = -2.8087 and a p-value = 0.0060. Since the p-value is less than the significance level ($\alpha = 0.05$), we reject the null hypothesis. Therefore, we can conclude that there is a statistically significant difference in mean blood pressure between patients with good appetite and those with poor appetite. This finding suggests a potential association between appetite and blood pressure, indicating that patients with poor appetite may exhibit different cardiovascular or renal profiles compared to those with good appetite.

4.2 H_2 Results (p-value, Coefficient, Odds Ratio, Accuracy, AUC)

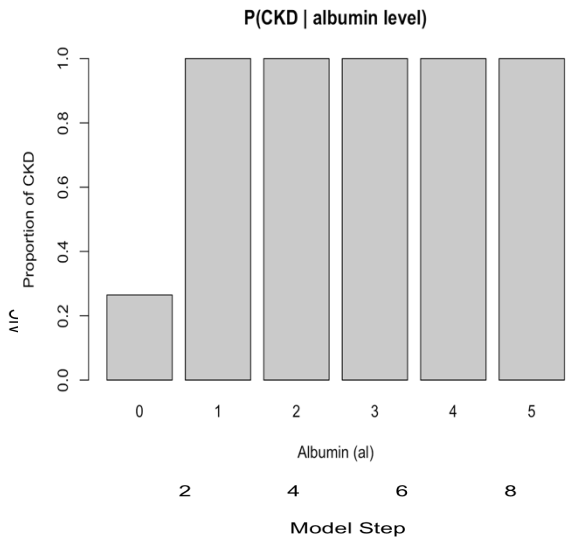
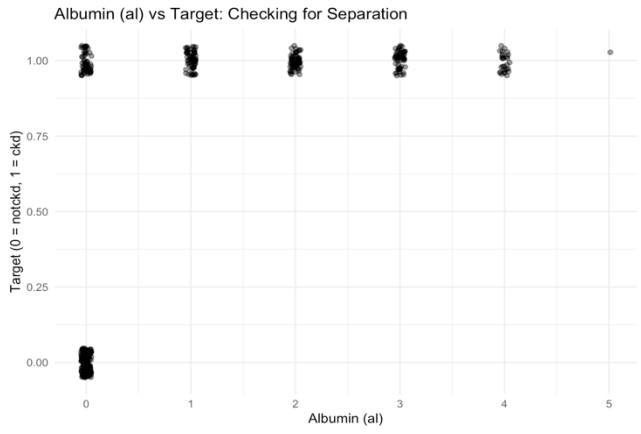


Figure 8 AIC Profile Across Candidate Models. The minimum AIC occurs at the 4-variable model (bgr, hemo, sg, al), selected by backward elimination.

For our second hypothesis, after conducting our variable selection procedures—including VIF analysis, stepwise AIC, and a diagnostic assessment for quasi-complete separation—we determined that three predictors (bgr, hemo, and sg) were appropriate for the final logistic regression model used to test Hypothesis 2. Although albumin (al) was initially selected by AIC, further investigation showed that it produced quasi-complete separation and unstable coefficient estimates, and therefore it was excluded from the final model. As shown in Figure 8, the proportion of CKD across albumin levels indicates that for al = 1–5, the CKD proportion is essentially 1.0, meaning every individual with albumin above zero is classified as CKD in the dataset. This produces a separation pattern that prevents stable maximum-likelihood estimation. The jitter plot in Figure 9 further confirms this pattern: albumin levels 1–5 contain only CKD cases (target = 1), whereas al = 0 includes both CKD and non-CKD observations. This is a clear example of quasi-complete separation, where a predictor nearly determines the outcome, leading to diverging coefficients, inflated standard errors, and unstable inference. These diagnostics justify removing albumin from the final logistic regression model to ensure numerical stability and reliable estimation..

Figure 9 Albumin vs Target (Jitter Plot)



Using the remaining three variables, we fitted a logistic regression model and obtained the following parameter estimates, shown in the Figure below. All three variables were statistically significant at the 0.01 level, as indicated by their very small p-values ($p < 0.001$). These p-values show strong evidence that each predictor is associated with CKD status after controlling for the others. The overall model fit also improved substantially compared to the null model: the null deviance decreased from 529.25 to 48.06, demonstrating that the predictors collectively explain a large proportion of the variability in disease outcome.

```
summary(glm_noal)
```

```
##
## Call:
## glm(formula = target ~ bgr + hemo + sg, family = binomial, data = df_reduced)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  738.18113  146.43290   5.041 4.63e-07 ***
## bgr            0.03992    0.01229   3.248  0.00116 **
## hemo          -2.31463    0.46461  -4.982 6.30e-07 ***
## sg          -698.07432  140.64585  -4.963 6.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 529.251  on 399  degrees of freedom
## Residual deviance:  48.056  on 396  degrees of freedom
## AIC: 56.056
##
## Number of Fisher Scoring iterations: 9
```

Figure 10 Logistic Regression Coefficient Estimates for bgr, hemo, and sg

Examining the estimated coefficients provides insight into how each predictor affects the likelihood of CKD. The coefficient for bgr (0.0399) is positive, indicating that higher blood glucose levels are associated with an increased probability of CKD, but not significantly. Although the coefficient appears numerically small, logistic regression coefficients are on the log-odds scale, so even small changes can translate into meaningful increases in risk. In contrast, hemo has a negative coefficient (−2.3146), suggesting that higher hemoglobin levels substantially reduce the odds of being classified as CKD, consistent with clinical expectations regarding anemia and kidney dysfunction. Finally, sg shows a strongly negative coefficient (−698.07). Because specific gravity is discretized and reflects urine concentration, lower sg values are associated with much higher odds of CKD, which explains why the coefficient magnitude is large—small differences in sg categories encode substantial differences in disease status. Since sg has a very large negative coefficient and never approaches zero in the dataset, the model compensates by producing a very large positive intercept (738.18) so that predicted probabilities remain within a plausible range.

The table 3 below summarizes the estimated odds ratios for the three predictors included in the final logistic regression model. The odds ratio for bgr is 1.0407 (95% CI: 1.0197–1.0705), indicating that each one-unit increase in random blood glucose is associated with approximately a 4% increase in the odds of

Table 3 odds ratio for the three predictor: bgr, hemo, sg

```
## --- Odds Ratios (with 95% CI) ---
```

```
print(or_table)
```

```
##       Variable    Coefficient     Odds_Ratio     OR_2.5        OR_97.5
## bgr        bgr     0.03992369  1.040731e+00 1.0196758  1.070470e+00
## hemo      hemo    -2.31463146  9.880259e-02 0.0319689  2.069911e-01
## sg          sg -698.07431647 6.763576e-304 0.0000000 3.111552e-203
```

CKD, holding other variables constant. In contrast, hemo shows a strong protective effect, with an odds ratio of 0.0988 (95% CI: 0.0320–0.2070). This means that higher hemoglobin levels are associated with substantially lower odds of CKD, consistent with the clinical relationship between anemia and kidney

dysfunction. Finally, sg exhibits an extremely small odds ratio (6.76×10^{-304}), with confidence intervals effectively indistinguishable from zero, reflecting its very strong negative association with CKD. Because lower specific gravity is a hallmark of impaired urinary concentration, even small changes in sg correspond to large shifts in CKD risk.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction notckd ckd
##      notckd      28   2
##      ckd          1  49
##
##                Accuracy : 0.9625
##                  95% CI : (0.8943, 0.9922)
##     No Information Rate : 0.6375
##     P-Value [Acc > NIR] : 3.692e-12
##
##                   Kappa : 0.9195
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9608
##             Specificity : 0.9655
##          Pos Pred Value : 0.9800
##          Neg Pred Value : 0.9333
##              Prevalence : 0.6375
##          Detection Rate : 0.6125
##    Detection Prevalence : 0.6250
##       Balanced Accuracy : 0.9632
##
##         'Positive' Class : ckd
##
```

Figure 11 Confusion Matrix and Statistics

The confusion matrix(Figure 11) shows that the logistic regression model performs very well in classifying CKD status. Out of 80 total observations, the model correctly identified 49 CKD cases and 28 non-CKD cases, misclassifying only 3 observations in total. This corresponds to an overall accuracy of 96.25%, with a 95% confidence interval of (0.8943, 0.9922). Sensitivity, which measures the ability to correctly identify CKD patients, is 96.08%, while specificity, the ability to correctly identify non-CKD individuals, is 96.55%. The positive predictive value (PPV) is 98.0%, meaning that nearly all individuals predicted as CKD truly have the disease, and the negative predictive value (NPV) is 93.33%, indicating that most predicted non-CKD cases are indeed negative. The balanced accuracy of 96.32% confirms that the model performs consistently well across both outcome classes.

The ROC curve for the final 3-variable logistic regression model(Figure 12) demonstrates excellent classification performance. The curve rises sharply toward the upper-left corner, indicating that the model achieves very high sensitivity while maintaining high specificity across a wide range of classification thresholds. The corresponding Area Under the Curve (AUC) is 0.9926. An AUC of this magnitude suggests that the model has an outstanding ability to discriminate between CKD and non-CKD patients: in practical terms, this means that more than 99% of the time, the model assigns a higher predicted probability to a true CKD case than to a non-CKD case. The near-perfect separation observed in the ROC curve additionally confirms that the combination of bgr, hemo, and sg provides very strong predictive power and yields a model with near-optimal diagnostic accuracy.

```
cat("AUC:", as.numeric(auc(roc_obj)), "\n")
```

```
## AUC: 0.9925625
```

```
plot(
  roc_obj,
  col = "blue",
  lwd = 2,
  main = "ROC Curve – Final 3-variable Logistic Model"
)
abline(a = 0, b = 1, lty = 2, col = "gray")
```

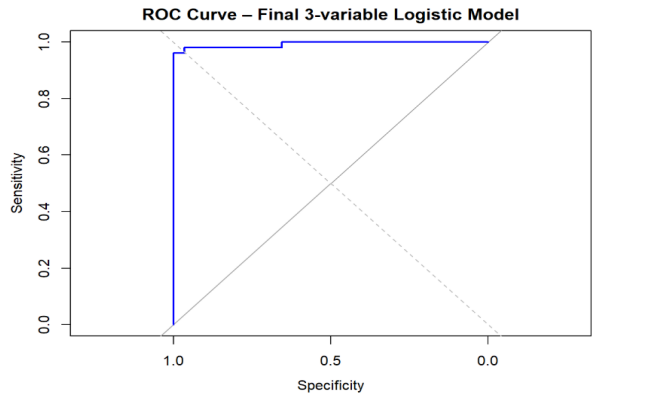


Figure 12 ROC curve for the final 3 variable Logistic Regression Model

Overall, the diagnostic statistics, confusion matrix results, and ROC–AUC analysis collectively indicate that the three selected predictors form a logistic regression model that performs exceptionally well in distinguishing CKD from non-CKD cases. The model demonstrates high accuracy, strong discriminative ability, and reliable predictive performance across multiple evaluation metrics.

4.3 Results of Hypothesis 1 – Revalidation under MCAR and MNAR Simulations

The updated results indicate that the conclusion of Hypothesis 1 is not consistent across all missing values. In the original dataset and with completely random missing data (MCAR), proportions of 10% and 20%, the t-test showed a statistically significant difference in mean blood pressure between patients with good and poor appetite, thus rejecting the null hypothesis (H_0). However, when the MCAR proportion increased to 30%, 40%, and 50%, the p-values exceeded the significance threshold, and the null hypothesis was no longer rejected. This suggests that a higher proportion of completely random missing data reduces the statistical power of the test and may mask potential associations. In the non-random missing data (MNAR) case, the t-test again showed a significant difference, indicating that the association between appetite and blood pressure can still be detected even with a non-random pattern of missing values. Overall, these findings suggest that the relationship between appetite and blood pressure has a certain degree of stability, but its detectability is sensitive to the amount and mechanism of missing data. KNN imputation appears to be effective with small amounts of missing data, but its ability to fully recover the signal decreases as data loss becomes more severe.

```
##
## === Step 4-1 Result: H1 under MCAR/MNAR ===
```

```
print(h1_missing_summary)
```

```
##       Type           t           p Mean_good Mean_poor           Decision
## 1 Original -2.6999458 0.008150774  75.42587  81.09756          Reject H0
## 2 MCAR 10% -2.6011517 0.010766985  75.67398  81.12500          Reject H0
## 3 MCAR 20% -2.7409969 0.007666162  75.47904  81.53846          Reject H0
## 4 MCAR 30% -1.6262891 0.109037069  76.56977  80.36364 Fail to Reject H0
## 5 MCAR 40% -1.2905412 0.201848182  74.41860  77.45455 Fail to Reject H0
## 6 MCAR 50% -0.7224023 0.473544720  74.78754  76.73913 Fail to Reject H0
## 7 MNAR 40% -2.1556365 0.033620137  73.28076  77.68293          Reject H0
```

Figure 13 H1: Results of Revalidation under Missing Types (MCAR, MNAR)

4.4 Results of Hypothesis 2 – Revalidation under MCAR and MNAR Simulations

Robustness analysis of Hypothesis 2 shows that the final three-variable logistic regression model remains highly stable under both missing completely at random (MCAR) and missing not at random (MNAR) conditions. Under all missing data conditions (MCAR from 10% to 50%, MNAR is 40%), the model consistently maintains extremely high accuracy (97.0%–98.7%), excellent AUC values (0.995–0.999) the plot below also provide a better visulation of the AUC under different missing data conditions, and high sensitivity and specificity, all of which are comparable to the performance observed on the original dataset. The pseudo R^2 values also remain high and within a narrow range (0.90–0.93), indicating that the model retains consistent explanatory power even after a significant portion of the data has been removed and imputed by KNN. Importantly, the consistency of the classification metrics across all missing levels demonstrates that the relationships captured by the logistic regression model (linking bgr, hemo, and sg

with CKD status) are highly robust and effectively address data loss. Even with 50% MCAR, the model maintains almost the same predictive performance, demonstrating that the KNN interpolation method effectively preserves the underlying data structure and enables the model to recover accurate and reliable predictions.

```
##
## === Step 4-2 Result: H2 Robustness Check (Final 3-variable model) ===
```

```
print(h2_summary)
```

```
##        Type MissingRate Accuracy       AUC Sensitivity Specificity  PseudoR2
## 1 Original         0.0   0.9825 0.9964533       0.984   0.9800000 0.9091994
## 2 MCAR 10%         0.1   0.9800 0.9980533       0.980   0.9800000 0.9299750
## 3 MCAR 20%         0.2   0.9825 0.9980267       0.980   0.9866667 0.9258238
## 4 MCAR 30%         0.3   0.9700 0.9970667       0.972   0.9666667 0.9081709
## 5 MCAR 40%         0.4   0.9975 0.9998933       1.000   0.9933333 0.9846892
## 6 MCAR 50%         0.5   0.9875 0.9950133       0.984   0.9933333 0.9218753
## 7 MNAR 40%         0.4   0.9825 0.9985200       0.988   0.9733333 0.9303903
##       LogLik
## 1 -24.028135
## 2 -18.530377
## 3 -19.628906
## 4 -24.300290
## 5  -4.051617
## 6 -20.673766
## 7 -18.420498
```

Figure 15 H2: Results of Revalidation under Missing Types (MCAR, MNAR)

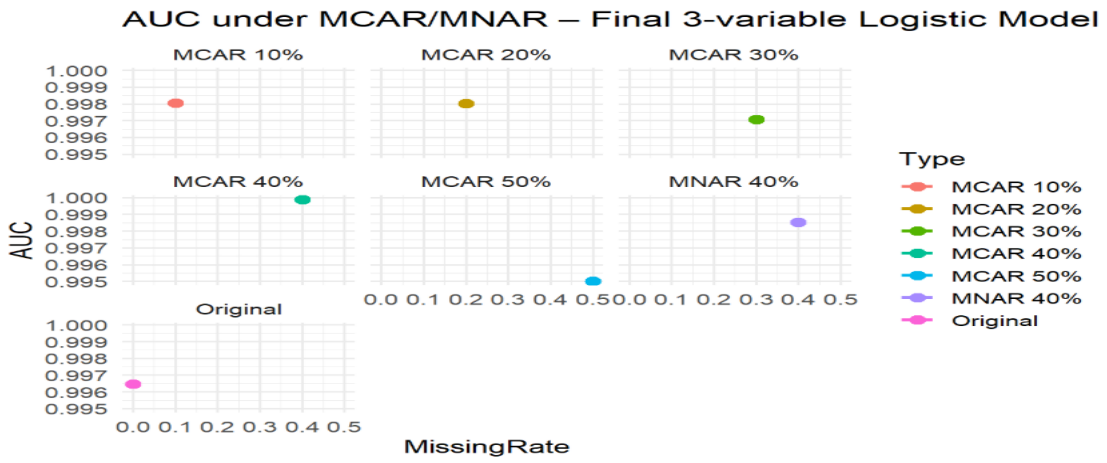


Figure 14 H2 re- analysis: AUC under MCAR, MNAR

5. Discussion and Conclusion

5.1 Interpretation of Findings

This project evaluated two statistical hypotheses (H_1 and H_2) and analyzed the stability of those conclusions under both MCAR and MNAR missing-data scenarios. Taken together, the results offer several insights into CKD-related clinical markers and their diagnostic relevance.

First, the analysis for H_1 found a statistically significant difference in mean blood pressure between patients with good and poor appetite after confirming unequal variances using the F-test. Although this reflects correlation rather than causation, the association aligns with clinical understanding: patients reporting poor appetite frequently present with more advanced CKD symptoms, including blood-pressure

abnormalities. Second, the multivariate logistic regression model developed for H_2 identified three predictors—bgr, hemo, and sg—as the most informative variables for distinguishing CKD from non-CKD patients. After addressing quasi-complete separation caused by albumin (al), the final logistic model demonstrated excellent discriminative performance across accuracy, AUC, and pseudo-R^2.

Third, the contrasting behaviors of H_1 and H_2 under simulated missingness emphasize the differing robustness of univariate versus multivariate analysis. H_1 showed reduced statistical power when missingness exceeded 30% MCAR, whereas the logistic model in H_2 maintained stable performance even with substantial MCAR and MNAR missingness. This indicates that multivariate structures can buffer information loss much more effectively than simple two-group comparisons. Overall, the results highlight the value of multivariable modeling in clinical research and demonstrate that conclusions regarding CKD risk factors remain reliable even under imperfect data conditions.

5.2 Impact of Missingness

Missing data proved to be a central theme throughout this project, influencing both methodology and interpretation. The CKD dataset exhibited a hybrid missingness structure consisting of MCAR-like patterns (e.g., age, bp, bu) and strong MNAR patterns (e.g., rbc, rbcc, wbcc, sg, al), which required tailored strategies for analysis.

For H_1, the effect of missingness was pronounced. Under MCAR scenarios with 10–20% missingness, the t-test result remained significant, but statistical significance disappeared once MCAR reached 30–50%. This demonstrates that small univariate comparisons are highly sensitive to random information loss and that listwise deletion becomes unreliable when missingness increases.

For H_2, missingness had a very different impact. Because logistic regression leverages multiple correlated predictors, the final model remained highly stable across all MCAR and MNAR simulations. Even with 50% MCAR and 40% MNAR, the accuracy, AUC, and pseudo-R^2 remained nearly unchanged. This suggests that multivariate structure and KNN imputation jointly preserve the underlying clinical relationships, allowing the model to recover predictive performance even when substantial portions of the dataset are missing.

In summary, missingness did not compromise the reliability of the multivariable CKD prediction model, but it substantially affected the univariate hypothesis test. This distinction underscores the importance of selecting analysis methods that align with the complexity of the underlying clinical relationships.

5.3 Limitations and Future Work

5.3.1 Limitations

A significant limitation of this project is the dataset's modest size of 400 total patients. This limits how generalizable the findings of our project is, depending on how well the dataset we used represents the entire CKD population. Additionally, the dataset did not disclose the specific cause or mechanism behind the missing data points. While MCAR and MNAR were simulated, these operate on assumptions that may not perfectly align with the full scope of why the total CKD population may regularly have missing data points.

Regarding the first hypothesis test H1, only a single CKD predictor was utilized, while relying on an assumption of approximately normal sampling distribution. This may affect the reliability of the

simulations that were performed beyond 30%. Next, while KNN performed well with this dataset, it still was subject to the typical limitations of KNN. This is because KNN imputation takes the neighboring row's averages to fill in gaps, which both: shrinks variability and risks being unstable if many variables are missing or highly correlated. Another limitation of this project is the quasi-complete separation of albumin. This effectively implies that albumin may be too strong of a predictor of CKD, and limits the logistic regression. This is problematic because we lost an important clinical predictor, and thus effectively biased the remaining predicting variables. Following this, a risk of our project is that: KNN imputation, dummy encoding, VIF screening, and backwards AIC selection may have an issue with overfitting. Effectively, our project finding a high degree of accuracy (96.3%) amongst bgr, hemoglobin, and sg suggests that these predictors may not actually be as strong and simply reflect overfitting, as opposed to accurately predicting CKD outside of the dataset that we used.

5.3.2 Future Work

As for ways to further improve future research into CKD, analysing larger, multisource datasets is the most significant way to expand upon our project's insight. This is because it would allow us to address potential overfitting, as well as further refine our understanding of the relationship between predictors and CKD while simultaneously reducing bias and improving modeling with missingness. Following this, performing multiple imputation methods would enhance our project's accuracy because it would allow for seeing which imputation method preserved the original dataset's data structure best, better examine the logistic model's performance, as well see how sensitive H1 is to any method of imputation. Another avenue for improving the accuracy of our project would be to use missingness indicators as predictors, as our usage of MCAR and MNAR operate on simplified mechanisms. This implies that these simulations may not properly capture the real reason why data may be missing in actual clinical settings. This improves both interpretability and reduces bias.

We also used logistic regression, which is best for examining phenomena that follow a linear relationship. This is somewhat problematic for CKD diagnostics because many predictors may not function linearly, and instead have non-linear critical thresholds for when they become problematic. Additionally, albumin had to be separated due to it almost perfectly predicting CKD versus non-CKD. Instead, a penalized logistic regression may be better as it would allow albumin, an important predictor of CKD to have remained. Also, while our project used test splits and cross-validation, this is inferior to training across multiple datasets.

Our project effectively examined how missing data mechanisms influence the reliability of multivariable analysis of a CKD dataset. Despite its success, it leaves room for improvement by using multiple datasets, as well as refining the methods that were used to analyse the dataset to further enhance accuracy as well as reduce bias and overfitting concerns.

References:

Bobbitt, Zach. "Understanding the Null Hypothesis for Logistic Regression." Statology, 29 September 2021, https://www.statology.org/null-hypothesis-of-logistic-regression/. Accessed 24 11 2025.

"CHAPTER 25 Missing-data imputation." Columbia University, https://sites.stat.columbia.edu/gelman/arm/missing.pdf. Accessed 24 11 2025.

"Chronic Kidney Disease." UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease. Accessed 24 November 2025.

"Stepwise Regression in Python." GeeksforGeeks, 23 July 2025, https://www.geeksforgeeks.org/machine-learning/stepwise-regression-in-python/. Accessed 24 November 2025.

Tamhane, Ajit C., and Dorothy D. Dunlop. Statistics and Data Analysis: From Elementary to Intermediate. Prentice Hall, 2000. Accessed 24 November 2025.

Multicollinearity & VIF James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). Springer.