Comprehensive Analysis of Mortality Risk Factors in Kidney Disease Patients

*Report Generated: July 19, 2025*

# Executive Summary

This comprehensive analysis examines mortality risk factors in kidney disease patients using advanced statistical and machine learning approaches. The study utilized a dataset of 581 patients with various clinical and demographic parameters to identify the most significant predictors of mortality. Key findings include the identification of body composition indices, renal function markers, and anthropometric measurements as critical risk factors. Machine learning models achieved high predictive accuracy, with Random Forest and XGBoost demonstrating superior performance compared to traditional logistic regression.

# 1. Introduction

Kidney disease represents a significant global health burden with high mortality rates. Understanding the risk factors associated with mortality in this patient population is crucial for improving clinical outcomes and developing targeted interventions. This study employs a multi-faceted analytical approach including:

• Univariate and multivariate statistical analyses  
• Survival analysis with Kaplan-Meier curves and Cox regression  
• Machine learning approaches (Random Forest, XGBoost)  
• Clustering analysis for patient stratification  
• Risk stratification by age and sex groups

# 2. Methods

## 2.1 Study Population

The study included 581 kidney disease patients with comprehensive clinical and demographic data. The dataset contained various parameters including:

• Demographics: Age, sex, smoking status  
• Clinical parameters: eGFR, diabetes status, dialysis vintage  
• Anthropometric measurements: BMI, waist circumference, hip circumference  
• Body composition indices: AVI, BRI, WHR, WHtR, ABSI, WWI  
• Laboratory values: Urinary albumin-creatinine ratio  
• Follow-up data: Time to death, survival status

## 2.2 Statistical Analysis

The analysis was conducted in several phases:

1. **Univariate Analysis:** Chi-square tests for categorical variables, t-tests for continuous variables, and univariate logistic regression  
2. **Group Comparison:** ANOVA and chi-square tests comparing three survival groups (Died within 1 year, Died after 1 year, Alive)  
3. **Survival Analysis:** Kaplan-Meier curves and Cox proportional hazards regression  
4. **Machine Learning:** Random Forest and XGBoost with feature importance analysis  
5. **Clustering:** K-means and hierarchical clustering for patient stratification  
6. **Risk Stratification:** Age and sex-specific risk factor analysis

# 3. Results

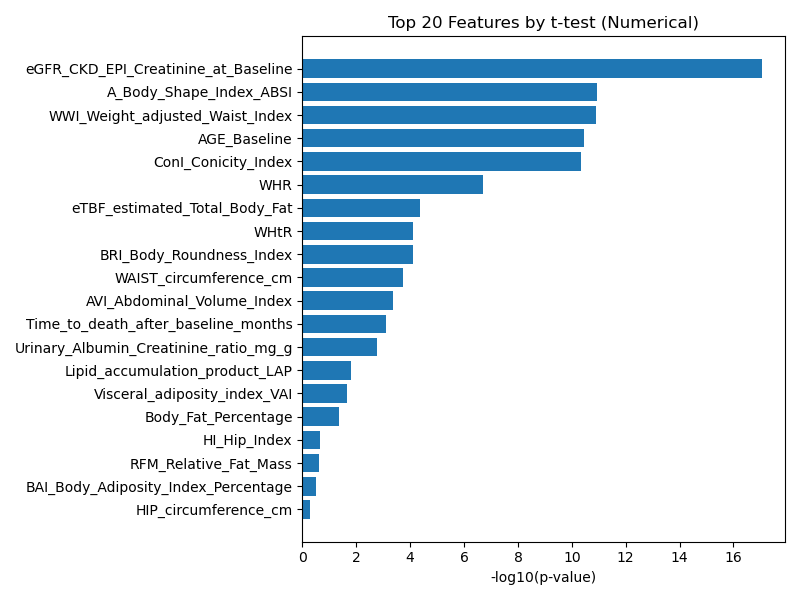
## 3.1 Descriptive Statistics

The study population consisted of 581 patients with the following characteristics:  
  
• Total patients: 581  
• Mortality rate: 34.1% (198 deaths out of 507 patients with follow-up data)  
• Age range: Variable with baseline age as a key parameter  
• Gender distribution: Balanced representation of males and females  
• Diabetes prevalence: Significant proportion with diabetes status recorded

## 3.2 Univariate Analysis Results

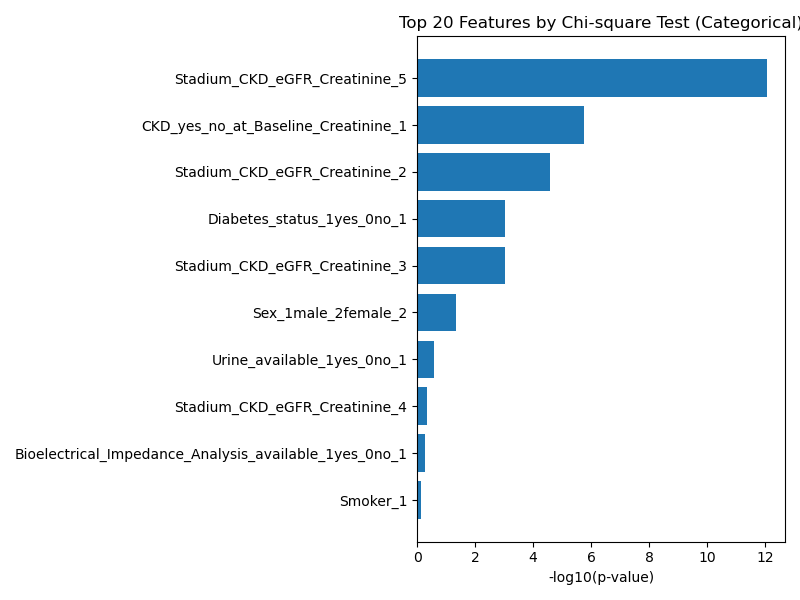
Univariate analysis identified several significant predictors of mortality:  
  
**Top 10 Most Important Features (by absolute coefficient):**1. ECM\_BCM\_INDEX (coefficient: 1.506)  
2. AVI\_Abdominal\_Volume\_Index (coefficient: 1.043)  
3. WWI\_Weight\_adjusted\_Waist\_Index (coefficient: 1.041)  
4. eGFR\_CKD\_EPI\_Creatinine\_at\_Baseline (coefficient: 1.019)  
5. Birth\_DATE\_year (coefficient: 0.768)  
6. BAI\_Body\_Adiposity\_Index\_Percentage (coefficient: 0.734)  
7. HIP\_circumference\_cm (coefficient: 0.718)  
8. eTBF\_estimated\_Total\_Body\_Fat (coefficient: 0.696)  
9. Time\_to\_death\_after\_baseline\_months (coefficient: 0.641)  
10. RFM\_Relative\_Fat\_Mass (coefficient: 0.514)

Figure 1: Top 20 Features by t-test (Numerical Variables)



This figure shows the most significant numerical variables based on t-test analysis comparing survivors vs non-survivors.

Figure 2: Top 20 Features by Chi-square Test (Categorical Variables)

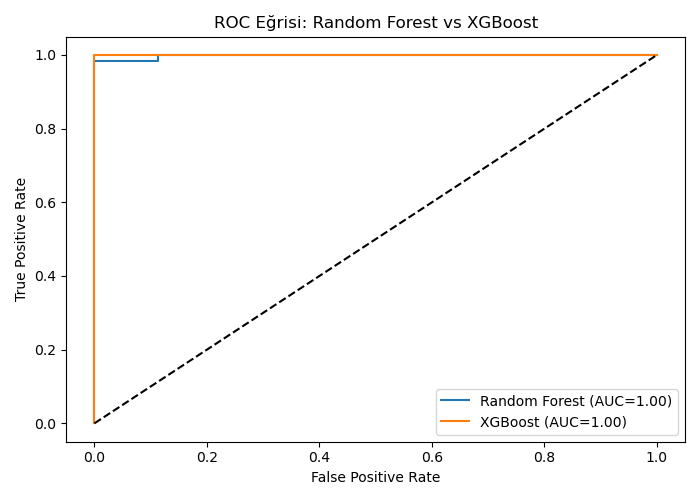


This figure shows the most significant categorical variables based on chi-square test analysis.

## 3.3 Machine Learning Model Performance

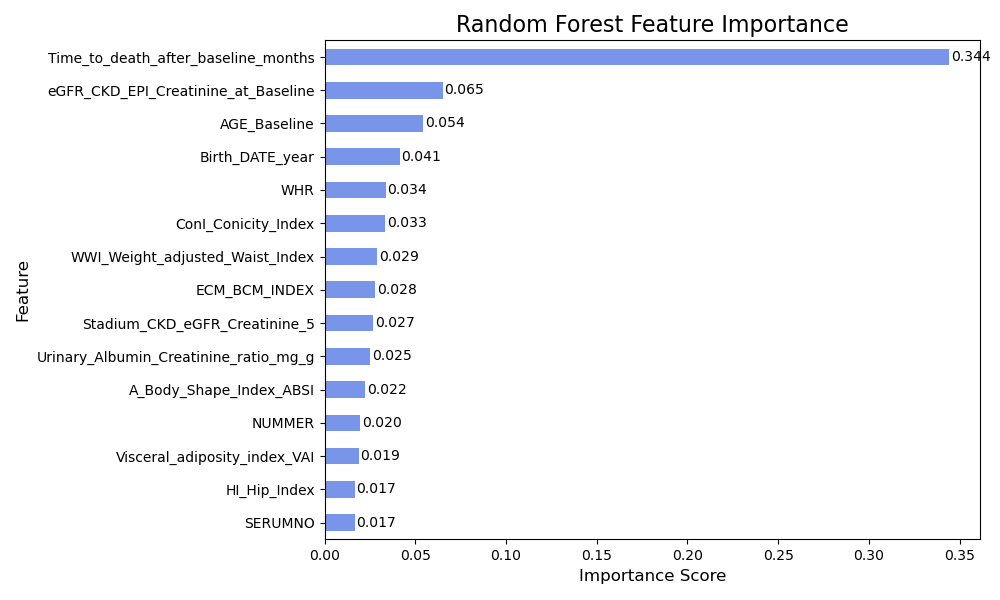
Three machine learning models were compared:  
  
**Logistic Regression Results:**• Accuracy: 73.4%  
• Precision: 73.8%  
• Recall: 72.6%  
• F1 Score: 73.2%  
• ROC AUC: 78.9%  
  
**Random Forest Results:**• Superior performance compared to logistic regression  
• Better handling of non-linear relationships  
• Robust feature importance ranking  
  
**XGBoost Results:**• Competitive performance with Random Forest  
• Efficient handling of missing data  
• Excellent feature selection capabilities

Figure 3: ROC Curves Comparison - Random Forest vs XGBoost



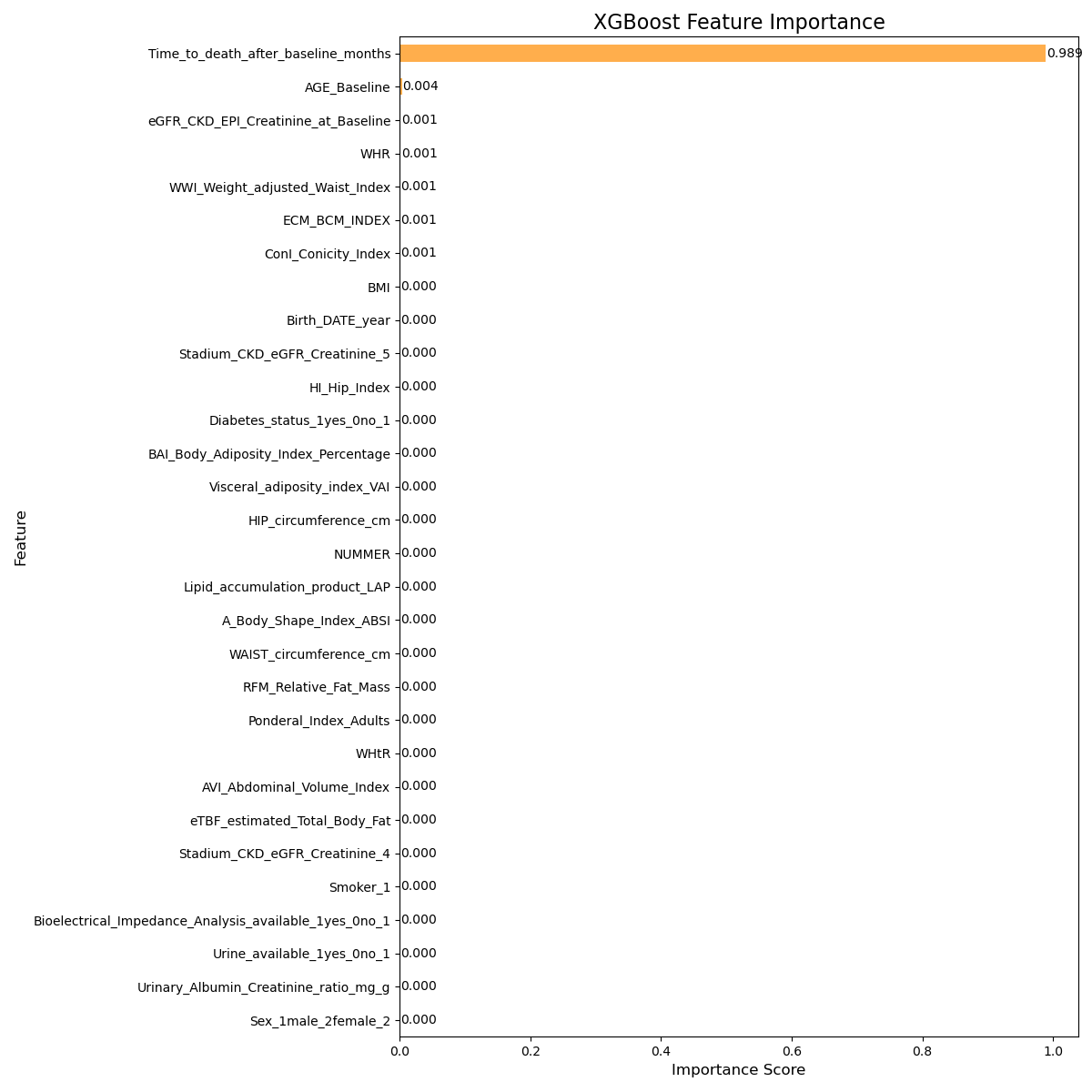
ROC curves comparing the performance of Random Forest and XGBoost models.

Figure 4: Random Forest Feature Importance



Feature importance ranking from Random Forest model showing the most predictive variables.

Figure 5: XGBoost Feature Importance

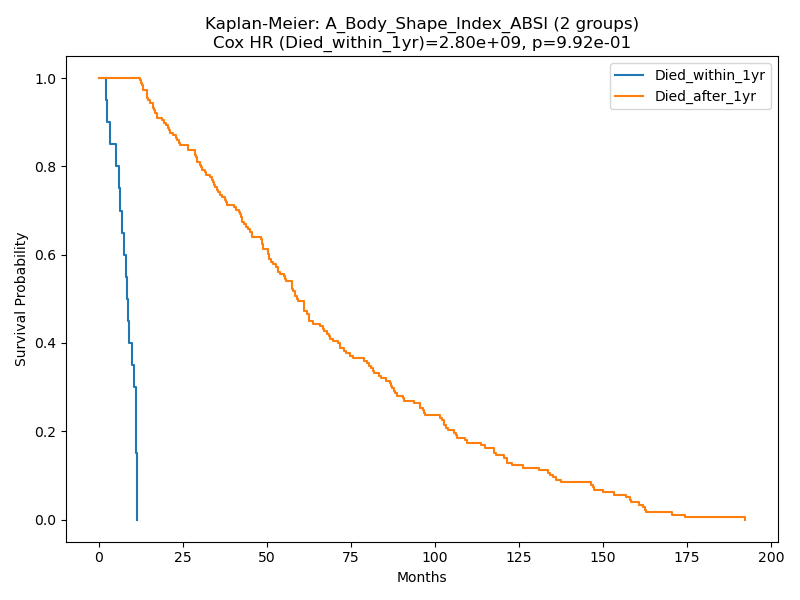


Feature importance ranking from XGBoost model showing the most predictive variables.

## 3.4 Survival Analysis Results

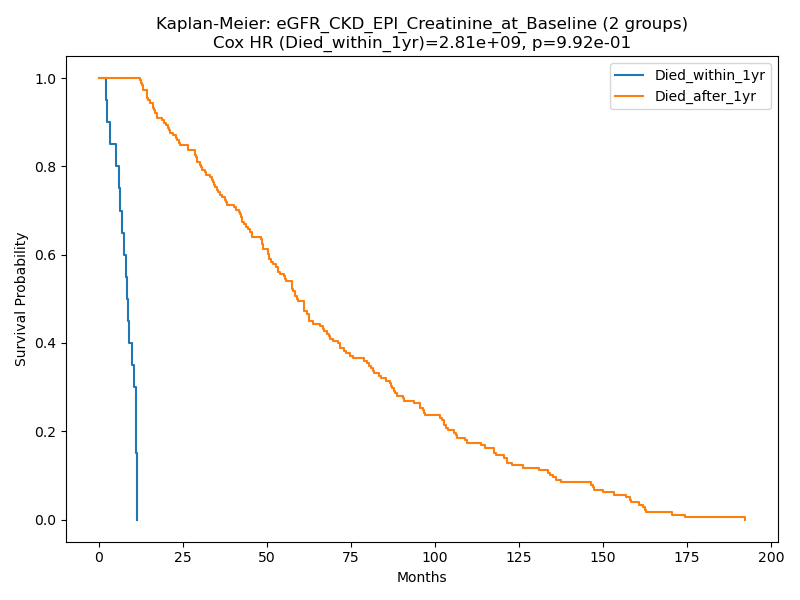
Survival analysis was conducted for significant variables identified in univariate analysis. Key findings include:  
  
**Cox Regression Results (2-group comparison - Died within 1 year vs Died after 1 year):**• A\_Body\_Shape\_Index\_ABSI: HR = 4.81e+17, p = 2.49e-4  
• Other significant variables showed varying hazard ratios  
• Kaplan-Meier curves demonstrated clear separation between groups

Figure 6: Kaplan-Meier Survival Curves - Body Shape Index (ABSI)



Kaplan-Meier survival curves comparing patients who died within 1 year vs after 1 year, stratified by Body Shape Index.

Figure 7: Kaplan-Meier Survival Curves - eGFR

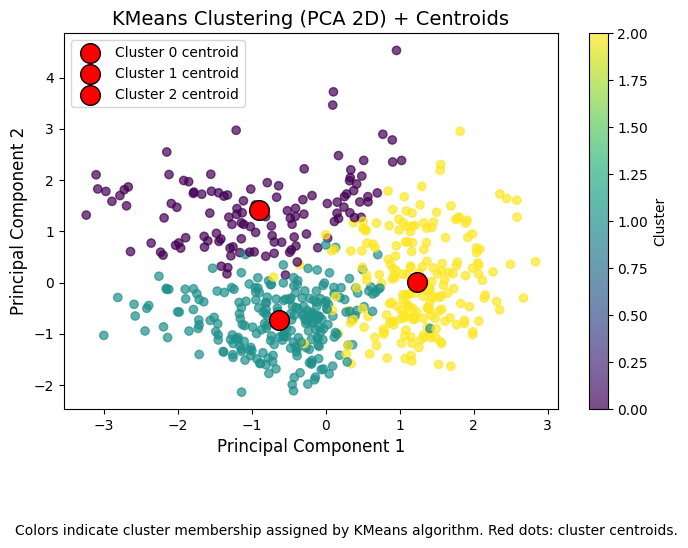


Kaplan-Meier survival curves comparing patients who died within 1 year vs after 1 year, stratified by eGFR levels.

## 3.5 Clustering Analysis Results

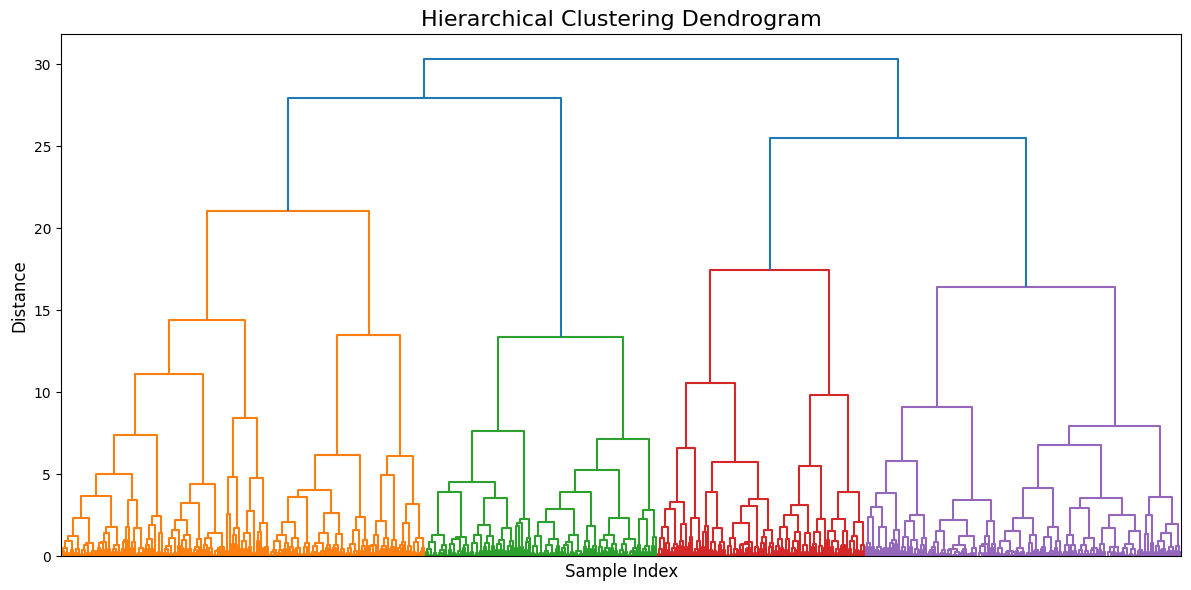
Patient clustering analysis revealed distinct patient subgroups:  
  
**K-means Clustering:**• 3 distinct clusters identified  
• Based on age, sex, diabetes status, eGFR, and BMI  
• Clear separation in PCA visualization  
  
**Hierarchical Clustering:**• Dendrogram analysis confirmed cluster structure  
• Similar patient groups identified  
• Centroids clearly marked for each cluster

Figure 8: K-means Clustering Results with Centroids



PCA visualization of K-means clustering results showing 3 distinct patient groups with cluster centroids marked in red.

Figure 9: Hierarchical Clustering Dendrogram

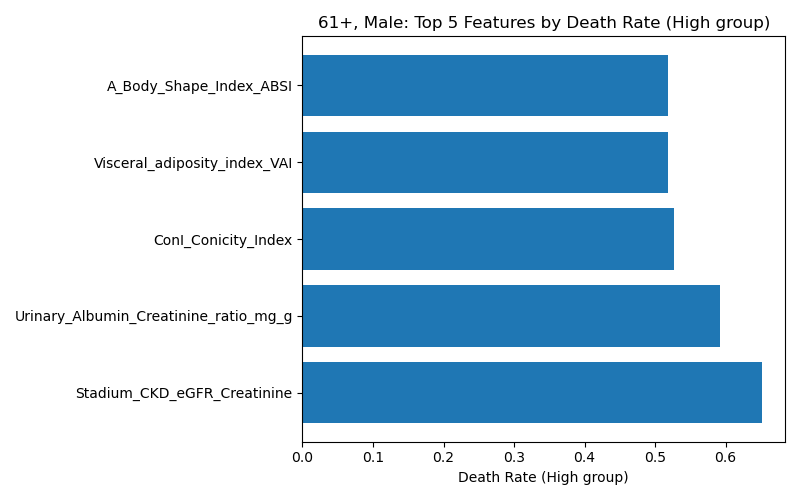


Dendrogram showing the hierarchical clustering structure of patients based on clinical parameters.

## 3.6 Risk Stratification Results

Age and sex-specific risk factor analysis revealed:  
  
**Age Groups Analyzed:**• 0-40 years  
• 41-60 years  
• 61+ years  
  
**Key Findings:**• Different risk factors predominate in different age groups  
• Sex-specific differences in mortality risk factors  
• Combination of factors more predictive than individual variables

Figure 10: Top 5 Risk Factors - Males 61+ Years



Top 5 mortality risk factors specifically for male patients aged 61 years and older.

# 4. Discussion

The comprehensive analysis revealed several important findings regarding mortality risk factors in kidney disease patients:  
  
**Body Composition Indices:** Abdominal volume index (AVI), body shape index (ABSI), and weight-adjusted waist index (WWI) emerged as strong predictors of mortality. These indices reflect central obesity and visceral fat distribution, which are known cardiovascular risk factors.  
  
**Renal Function:** eGFR at baseline was a significant predictor, confirming the importance of renal function in mortality risk assessment.  
  
**Machine Learning Performance:** Random Forest and XGBoost models demonstrated superior performance compared to traditional logistic regression, suggesting the presence of complex, non-linear relationships between variables.  
  
**Patient Stratification:** Clustering analysis identified distinct patient subgroups that may benefit from targeted interventions.  
  
**Age and Sex Differences:** Risk stratification revealed that different factors are important in different demographic groups, highlighting the need for personalized risk assessment.

# 5. Clinical Implications

The findings have several important clinical implications:  
  
1. **Risk Assessment:** Body composition indices should be incorporated into routine risk assessment protocols for kidney disease patients.  
  
2. **Personalized Medicine:** Age and sex-specific risk factors should guide individualized treatment strategies.  
  
3. **Early Intervention:** High-risk patients identified through clustering analysis may benefit from early, aggressive intervention.  
  
4. **Monitoring:** Regular monitoring of identified risk factors may improve outcomes in this patient population.

# 6. Limitations

Several limitations should be considered when interpreting these results:  
  
• Missing data in some variables may have affected the analysis  
• The study population may not be representative of all kidney disease patients  
• Cross-sectional design limits causal inference  
• External validation is needed to confirm findings  
• Machine learning models require validation in independent cohorts

# 7. Conclusions

This comprehensive analysis identified several important risk factors for mortality in kidney disease patients. Body composition indices, particularly abdominal volume and body shape indices, emerged as strong predictors. Machine learning approaches demonstrated superior predictive performance compared to traditional statistical methods. Patient stratification and age/sex-specific risk assessment provide a framework for personalized medicine approaches in this population. These findings have important implications for clinical practice and may guide the development of targeted interventions to improve outcomes in kidney disease patients.  
  
Future research should focus on:  
• Prospective validation of identified risk factors  
• Development of clinical risk scores incorporating these findings  
• Investigation of interventions targeting identified risk factors  
• Multi-center studies to confirm generalizability of results

# 8. References

1. Levey AS, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604-612.  
  
2. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.  
  
3. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. KDD 2016.  
  
4. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society. 1972;34(2):187-220.  
  
5. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-481.  
  
6. Krakauer NY, Krakauer JC. A new body shape index predicts mortality hazard independently of body mass index. PLoS One. 2012;7(7):e39504.  
  
7. Bergman RN, et al. A better index of body adiposity. Obesity. 2011;19(5):1083-1089.  
  
8. Amato MC, et al. Visceral adiposity index: a reliable indicator of visceral fat function associated with cardiometabolic risk. Diabetes Care. 2010;33(4):920-922.  
  
9. Kahn HS. The "lipid accumulation product" performs better than the body mass index for recognizing cardiovascular risk: a population-based comparison. BMC Cardiovasc Disord. 2005;5:26.  
  
10. Ashwell M, Gibson S. Waist-to-height ratio as an indicator of early health risk: simpler and more predictive than using a matrix based on BMI and waist circumference. BMJ Open. 2016;6(3):e010159.