MIDDLE EAST TECHNICAL UNIVERSITY

Report

The Nulls

by:

Abulfaz Khalilov

2655991

Işıl Kartal

2666626

Ata Kavak

2666642

Havva Nur Tekin

2437416

2026 January

METU/ODTÜ

Abstract

For this analysis, a synthetic dataset of routine outpatients was used to examine variables affecting kidney health. This dataset considered demographic, health and lifestyle-related variables such as age, diabetes status, BMI, systolic blood pressure, smoking status, physical activity level, ethnicity and hypertension.

I. INTRODUCTION

Many early signs of kidney problems, which are common worldwide, may go unnoticed in daily life. Identifying the factors that cause changes in kidney function is crucial for early diagnosis of potential risks. In this project, a synthetic dataset was created based on routine outpatient visits, encompassing demographic characteristics, lifestyle behaviors, and clinical measurements, indicators such as eGFR and urinary ACR, and various variables that can affect kidney health. Python libraries were used to clean the dataset before it was ready for analysis. After preparation, multivariate visualizations targeting different patterns related to kidney outcomes were created. Specific research questions concerned how age and diabetes status relate to eGFR, how variation in BMI and systolic blood pressure differs between stages of chronic kidney disease (CKD), how smoking and physical activity level relate to urinary ACR, and how advanced CKD differs with ethnicity and hypertension group. This analysis was conducted to understand how different factors affect kidney function in the dataset.

II. DATA DESCRIPTION

Before cleaning the dataset, it contained 1162 entries and 13 variables.
9 of these variables are numerical and 4 of them are categorical.The variables are described as follows:
1.Age
Patient age in years (18-90).
Numerical - Continuous/Discrete
2. Sex
Biological sex of the patient (Female, Male).
Categorical - Nominal
3. Ethnicity
Self-reported ethnicity group (African, European, Middle Eastern, South Asian, Turkish,Other).
Categorical - Nominal
4.Smoking Status
Smoking behavior (Never, Former, Current).
Categorical - Nominal
5. Physical Activity Level General physical activity level (Low, Moderate, High).
6. Diabetes
Whether the patient has diabetes (Yes, No).
Categorical - Nominal
7. Hypertension
Whether the patient has hypertension (Yes, No).

Categorical -Nominal
8.BMI
Body Mass Index in kg/ m^2. (16.5-41.3)
Numerical - Continuous
9. Systolic BP
Measured systolic blood pressure(mmHg). (85-171)
Numerical - Continuous
10. Serum Creatinine
Serum creatinine level in mg/dL. (0.45-1.39)
Numerical - Continuous
11. e GFR
Estimated Glomerular Filtration Rate (14.9-126.5)(mL/min/1.73 m^2).
Numerical - Continuous
12. Urine ACR
Urine albumin-to-creatinine ratio(mg/g).(1.7-52.0)
Numerical - Continuous
13. CKD Stage
Chronic kidney disease severity
(Normal, Mild, Moderate , Severe).
Categorical - Ordinal

III. DATA CLEANING AND TIDYING STEPS

The first dataset, “kidney_function_synthetic_1150_dirty.csv,” had 1,162 observations and 13 variables, and it had several inconsistencies and irregularities in the formating and missing entries. A detailed data cleansing proces were performed using Python and the Pandas and NumPy libraries.

First, the column headers were normalized for easy handling. The original column headers had spaces, special characters such as parentheses, dots, and slashes, and some irregular capitalization. All the column headers was normalized to snake_case notation and had units/special character removed (for example, bmi(kg/m2) was renamed bmi). The variables were renamed with a dictionary mapping for meaningfull and consistent names.

A large part of the cleaning process involved dealing with categorical variables such as "sex," "ethnicity," "smoking_status," "physical_activity_level," "diabetes," "hypertension," and "ckd_stage." These columns contained many inconsistencies, ranging from differences in capitalization (e.g., "Male" and "male") to typographical errors (e.g., "Mlae," "tukrish") and the presence of unnecessary special characters (?, *, #). This was resolved using a customized function, "new_format," which standardized the categorical variables by making everything lowercase, removing unnecessary whitespace, and removing punctuation. This was followed by the use of customized mapping dictionaries to correct observed typos. The final step involved the conversion of all categorical variables to Title Case format for consistency. The problem of duplicates was resolved following the categorical variables' standardization procedure. Rows with different capitalization or spaces were examples of concealed duplicates that were previously missed during the first string standardization process. Twelve rows in all were found to be duplicates.

Additionally, the numerical variables were included. The data was preprocessed by removing the dashes and altering the data type to numbers in the age column when numbers were preceded by dashes (such as "24-"). The mean for numerical variables (age, bmi, systolic_bp, serum_creatinine, egfr, urine_acr) and the mode for categorical variables were used to handle missing data. And as a last step we used Z-score method to detect the outliers. So we just changed each data point where the |Z| > 3 to the mean.
After all these steps were made we saved the cleaned dataset as cleaned_kidney_data.csv, which has 1550 rows and 13 variables without any missing value

IV. DESCRIPTIVE STATISTICS

Variable	Count	Mean	Std. Dev.	Min	Median
Age (years)	1,15	50.33	16.21	18.00	50.00
BMI (kg/m^2)	1,15	27.91	4.49	16.50	27.85
Systolic BP (mmHg)	1,15	121.96	16.20	85.00	121.00
Serum Creatinine (mg/dL)	1,15	0.89	0.17	0.45	0.88

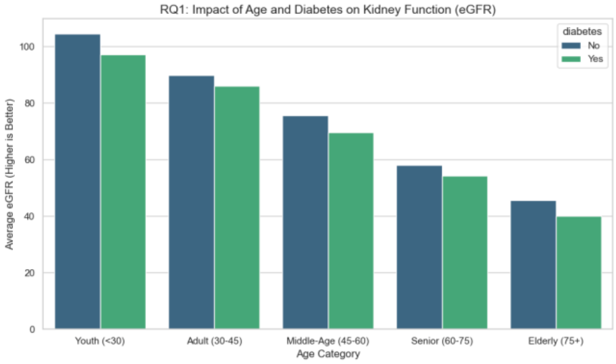
eGFR (mL/min/1.73m²)	1,15	76.54	20.74	14.90	76.80
Urine ACR (mg/g)	1,15	14.34	8.89	1.70	12.20

V. EXPLANATORY DATA ANALYSIS

The dataset utilized in this study comprises synthetic medical records of 1,150 patients, specifically designed to simulate kidney function parameters and associated demographic and clinical risk factors. Before conducting statistical analyses, a deep data cleaning process was performed using python to address significant inconsistencies, this included mapping typographical errors in categorical variables like *Ethnicity* to standardized groups, coercing non-numeric entries in the age column to valid integers, and unifying inconsistent labels in binary variables such as hypertension and diabetes. The demographic profile of the cleaned data reveals a diverse population with a wide age distribution ranging from young adults to the elderly, a balanced gender ratio, and a predominance of Turkish, Middle Eastern, and European ethnic groups. Clinically, a significant portion of the population falls into "Overweight" or "Obese" BMI categories, often overlapping with low physical activity levels and hypertensive blood pressure ranges. The primary target variables, CDK stage and eGFR, exhibit a skewed distribution where the majority of individuals are healthy or in early disease stages, while a critical minority suffers from Moderate to Severe CKD. Preliminary correlation checks confirmed biologically plausible relationships, such as the strong negative correlation between *Serum Creatinine* and *eGFR*, and the positive association between Systolic BP and age , validating the dataset's suitability for further hypothesis testing and risk analysis.

Research Question 1: How do Age and Diabetes status impact kidney function (eGFR)? (Havva Nur Tekin)

The purpose of this question is to explore how age and diabetes status together influence eGFR. While eGFR is naturally affected by age, adding diabetes into the analysis allows us to see if the relationship changes for different groups. To visualise the result, a histogram was made with eGFR and age axis, and diabetes was shown in bars.



According to the histogram, the y-axis represents the eGFR values, the x-axis represents age groups, and the diabetes status is indicated by 'yes' and 'no' in two separate bars. It is clear that eGFR dramatically decreases with age. For a better understanding, basically, eGFR is the rate of filtration, so the higher the eGFR, is a better the filtration and working capacity of the kidneys.

The analysis is focused on five age groups (1. Youth (<30), 2. adult (30-45), 3. Middle age (45-60), 4. senior (60-75), 5. elderly (75+)).

Youth (<30) and Adult (30-45): In these early stages, the filration is at its peak, there is a slight inequality between non-diabetics and diabetics, as non-diabetics has a higher rate.

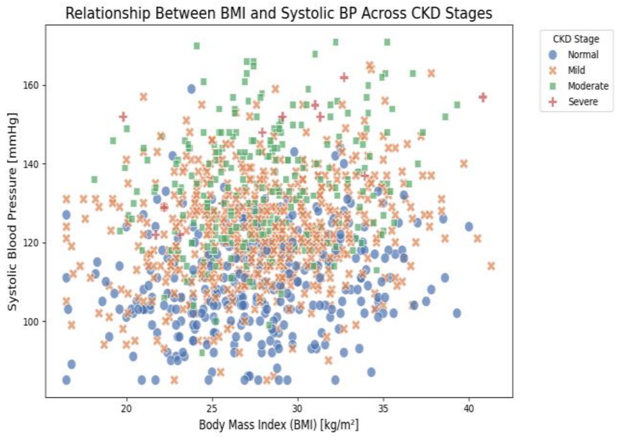
Middle-Aged (45-60) & Senior (60-75): This category exhibits the greatest change. The declining graph indicates that kidney work capacity also deteriorates with age. In terms of medical standards, increasing eGFR indicates enhanced filtration function; it can be noticed that the average levels start to decrease close to the 60 mL/min mark.

Elderly (75+): The last group sees the filtration rate at its lowest level, which depicts the cumulative effect caused by both aging and illnesses.

Analysis of the Multivariate Interaction: In the analysis of the relationship between the 'Age' and 'Are diabetic?' categories, the pattern of 'accelerated decline' can be seen. It becomes clear that diabetes functions as a compounding factor, where in each and every stage of 'Age,' the 'Yes' categories ('Are diabetic?') are lower than the 'No' categories ('Not diabetic?'). This pattern strongly implies that the presence of diabetes pushes a patient into a lower stage of filtration than would have been the case based upon the patient's age alone.

Research Question 2: What is the relationship between Systolic Blood Pressure and Body Mass Index (BMI) at different stages of Chronic Kidney Disease (CKD)? (Abulfaz Khalilov)

A multivariate scatterplot was made to investigate the connection between kidney function, obesity, and hypertension.



Visualisation 2: Relationship between BMI and Systolic BP Across CKD Stages

In this scatterplot, Body Mass Index is graphed on the x-axis while Systolic Blood Pressure is graphed on the y-axis. Data points are coded according to CKD stages (Normal, Mild, Moderate, Severe) through color and style.

Observations: From the scatter plot, it can be observed that the BMI and Systolic Blood Pressure are related to each other in a positive way. That is, the higher the BMI values, the higher the systolic blood pressure. This is expected because a higher BMI is known to be related to hypertension. It is worth noting that a characteristic clustering pattern exists for CKD stage.

Normal and Mild CKD: The data points for patients with either Normal or Mild CKD (using lighter indicators) tend to cluster in the lower left corner, suggesting that patients with better kidney function tend to have a BMI < 25 kg/m2 and systolic blood pressure < 120 mmHg.

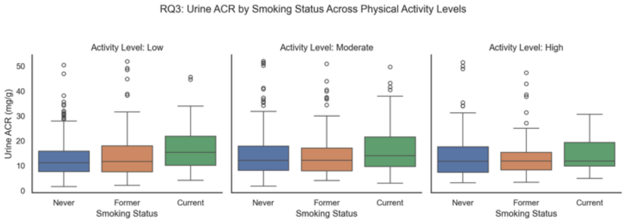
Advanced cases: The ones with Moderate or Serve CKD mostly are located in the upper right corner of the graph, this suggest that the ones with a BMI close to overweight or obese, with high sistolic blod presure, indicative of Hypertension.

Conclusion: From the visualization, it can be noted that there is a compounding relationship between high

BMI and high blood pressure in many cases of advanced CKD. While it cannot be said to be a causal relationship due to the observational nature of the data, it does suggest that weight control and blood pressure control are important in postponing progression to advanced CKD.

Research Question 3: Urine ACR by Smoking Status Across Physical Activity Levels(Işıl KARTAL)

This question wants to explore the relationship between smoking and the level of urinary ACR which is a significant indicator of kidney health. The goal is to see if smoking affects ACR similarly in individuals with low, moderate, or high activity levels. To investigate this, a box plot was created showing the distribution of urinary ACR values according to three smoking statuses (Never smoker, Former smoker and Current smoker) and low, moderate, or high levels of physical activity.

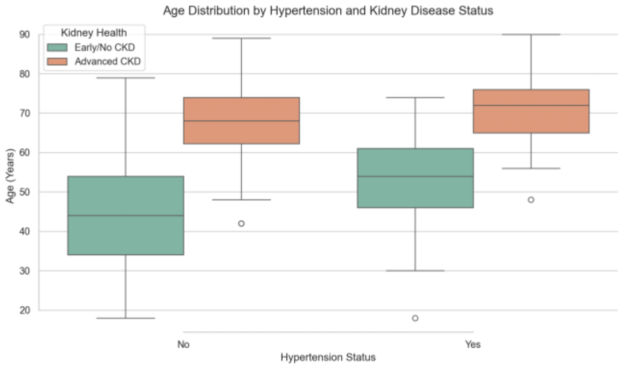


Visualization3: Urine ACR by Smoking Status Across Physical Activity Levels

This visualization shows the effect of smoking status and physical activity level on Urine ACR values. At all physical activity levels, the median ACR values of current smokers are higher than those of never-smokers and former smokers. This indicates that smoking is associated with kidney problems.

When the smoking group is examined within itself, the highest ACR values are observed at the low physical activity level, and the lowest ACR values are observed at the high physical activity level. This finding suggests that physical activity may partially mitigate the negative effects of smoking. In the evaluation according to physical activity level, the highest ACR values are found in the low activity group for all three smoking statuses. Lower median ACR values are observed in the high physical activity level. In addition, the higher number of outliers in the low activity group indicates that ACR values are variable in this group. Overall, the results show that smoking increases ACR values and that increased physical activity level may have a protective effect on kidney health.

Research Question 4: Is there a significant age disparity between patients suffering from Advanced Chronic Kidney Disease (CKD) with Hypertension compared to those without these conditions? (Ata KAVAK)



Visualization4: Age Distribution by Hypertension and CKD Status

This study utilizes a synthetic dataset consisting of 1150 medical records designed to simulate the complex interplay between kidney function parameters and various demographic and clinical risk factors. Prior to statistical modeling a rigorous data cleaning process was implemented using Python to rectify significant inconsistencies within the raw data which included mapping typographical errors in the Ethnicity variable to standardized categories such as

Turkish European and Middle Eastern coercing non numeric and malformed entries in the Age column to valid integers and unifying inconsistent labels across binary variables including Hypertension and Diabetes. The demographic analysis of the cleaned dataset reveals a heterogeneous population characterized by a wide age range extending from young adults to the elderly a balanced gender distribution and a specific ethnic composition that reflects the regional focus of the simulation. In terms of clinical characteristics a substantial proportion of the subjects are classified as Overweight or Obese based on Body Mass Index BMI which notably overlaps with subgroups reporting low physical activity levels and elevated systolic blood pressure readings. The primary outcome variables of interest specifically CKD Stage and estimated Glomerular Filtration Rate eGFR display a skewed distribution where the vast majority of the population presents as healthy or with mild renal impairment while a critical high risk minority is identified with Moderate to Severe Chronic Kidney Disease. Furthermore preliminary bivariate investigations confirmed biologically consistent correlations such as the inverse relationship between Serum Creatinine levels and eGFR as well as the progressive increase in Systolic Blood Pressure associated with advancing age thereby validating the structural integrity and clinical plausibility of the dataset for subsequent hypothesis testing and risk assessment.

CONCLUSION

To conclude, our analysis of the kidney functions highlighted several significant relationships between various factors. Age was shown primary reason of declining of kidney function (eGFR).
The prevalence of severe cases of CKD was closely associated with the condition of high blood pressure. Moreover, lifestyle factors, particularly joint smoking status and physical activity levels, were found to contribute to both eGFR values as well as levels of early renal impairment, as shown by Urine ACR. Finally, the multivariate analyses among various ethnicity levels, it has been found that while the rate of kidney degeneration is a universal process. Having co-existing conditions works as a complicating factor that classifies these patients into more critical levels of progression.
This makes one understand the different levels of multidictates in which patients with CKD are placed and yet reinforcing the call to address the earliest forms of clinical practice aimed at overcoming the cumulative processes of aging and the incidence of diseases of a metabolic nature.

https://github.com/AbulfazKh