King Saud University
College of Computer Science and Information
Department of Information Technology
Data Science & Artificial Intelligence Track

Research paper- 2024-2025/1446-1447

Causal Analysis of Osteoporosis Dataset

Author: Sirin Alhumud

Abstract

This study investigates the causal factors contributing to osteoporosis using a structured dataset. Data exploration, statistical testing, and causal diagramming were utilized to analyze relationships between age, physical activity, calcium intake, vitamin D intake, body weight, smoking, prior fractures, and other demographic factors. The findings underscore the central role of age, along with significant age-dependent interactions with key variables. These insights provide a foundation for preventive strategies targeting high-risk groups, particularly older adults.

Introduction

Osteoporosis, a condition characterized by reduced bone density and increased fracture risk, progresses silently until fractures occur (National Institutes of Health, 2023). Understanding its causal factors is critical for effective prevention. This study examines a dataset of 1,958 individuals using statistical methods and causal diagrams to explore direct effects and age-dependent interactions with lifestyle and demographic factors contributing to osteoporosis.

Methods

1. **Data Description**

   The dataset comprises 16 variables, including demographic, medical, and lifestyle factors. The target variable, Osteoporosis, is binary (1 = Diagnosed, 0 = Not Diagnosed).

2. **Preprocessing and Cleaning**

   No duplicates were found, and missing values were imputed using mode. Inconsistent categorical values (e.g., "Modre" for Alcohol Consumption) were corrected using regular expressions to ensure data consistency. Interaction variables (e.g., Age_PhysicalActivity) were created to capture combined effects.

3. **Exploratory Data Analysis (EDA)**

   Numerical Variables:

   Boxplot and histogram indicated a higher prevalence of osteoporosis among older individuals and the heatmap showed a strong correlation (r = 0.69) between age and osteoporosis *[Figures 1-3]*.

   Categorical Variables:

   The analysis of categorical variables involved bar charts to identify potential patterns and Chi-Square tests to determine statistical significance. The plots suggested that sedentary individuals, underweight individuals, those with insufficient calcium or vitamin D intake, and smokers might exhibit a higher prevalence of osteoporosis *[Figure 4]*. However, Chi-Square tests indicated no statistically significant associations for these variables when

considered independently (p-values > 0.05). For instance, the p-values for Physical Activity (0.4290), Calcium Intake (0.9104), and others show no strong direct statistical link to osteoporosis risk. This highlights the importance of **Age** and its interactions with these variables, as visual trends alone do not fully capture the complexity of their relationships with osteoporosis.

Age-Related Chi-Square Test Results:

The effect of interactions between age and other variables on osteoporosis was analyzed by creating interaction variables that combine age with each factor (e.g., Physical Activity, Vitamin D Intake, and Calcium Intake). The results revealed statistically significant interactions between age and Physical Activity, Vitamin D Intake, Calcium Intake, Body Weight, Smoking, Prior Fractures, Gender, Hormonal Changes, and Family History ($p < 0.05$). Conversely, no significant interactions were observed for Alcohol Consumption, Medical Conditions, or Medications ($p \geq 0.05$). These findings highlight that the influence of various lifestyle and biological factors on osteoporosis risk is significantly mediated by age.

Causal Diagram (DAG) Analysis

The **Simplified Diagram** establishes age as the central direct factor influencing osteoporosis, reflecting its predominant role *[Figure 5]*. The **Enhanced Diagram** integrates significant interactions between age and other factors such as physical activity, body weight, smoking, vitamin D intake, calcium intake, and prior fractures. These interactions illustrate how age amplifies the influence of lifestyle and biological factors on osteoporosis risk *[Figure 6]*. The inclusion of these interaction effects is supported by significant chi-square results ($p < 0.05$).

Results

The analysis identified age as the central predictor of osteoporosis, influencing risk directly and through interactions with key factors. Significant interactions included physical activity, calcium intake, vitamin D intake, body weight, smoking, race/ethnicity, hormonal changes, family history, and prior fractures. Sedentary lifestyles, insufficient nutrition, underweight status, and smoking were particularly impactful, while prior fractures strongly indicated bone fragility. Variables such as alcohol consumption, medical conditions, and medications showed no significant association. The causal diagram highlights the age-dependent relationships influencing osteoporosis risk *[Figure 6]*.

Discussion

Age plays a critical role in osteoporosis risk, directly and by amplifying the effects of modifiable and demographic factors. Interactions with physical activity, nutritional intake, smoking, and prior fractures emphasize the importance of addressing these risks. The study also highlights the role of genetic and hormonal predispositions in shaping osteoporosis susceptibility, reinforcing the need for targeted interventions.

Recommendations

1. **Encourage Physical Activity**: Promote regular exercise, especially among older adults, to strengthen bones.

2. **Improve Nutrition**: Advocate for adequate calcium and vitamin D intake through diet or supplements.
3. **Smoking Cessation**: Develop programs to reduce smoking-related bone health risks.
4. **Address Low Body Weight**: Implement strategies to manage underweight individuals.
5. **Screen for Prior Fractures**: Target individuals with prior fractures for early osteoporosis management.

Limitations

1. **Imputation Bias**: Mode imputation may have masked true variability.
2. **Missing Confounders**: Genetic factors and detailed dietary habits were not included.
3. **Limited Generalizability**: Results may not fully apply to all populations.

Conclusion

This study highlights the significant role of age and its interactions with lifestyle and biological factors in osteoporosis risk. By integrating causal diagrams and statistical tests, the findings provide practical insights for prevention strategies and set the stage for more comprehensive future research.

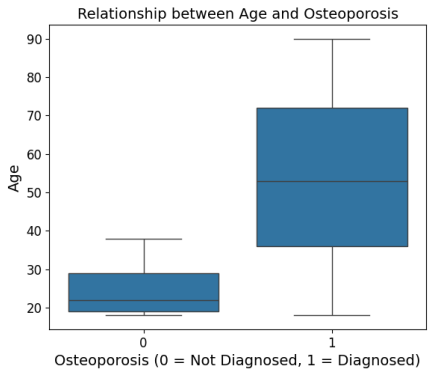
References

1. Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Wiley. https://www.wiley.com/en-us/Causal+Inference+in+Statistics%3A+A+Primer-p-9781119186847
2. National Institutes of Health (2023). *Osteoporosis Overview*. https://www.niams.nih.gov/health-topics/osteoporosis

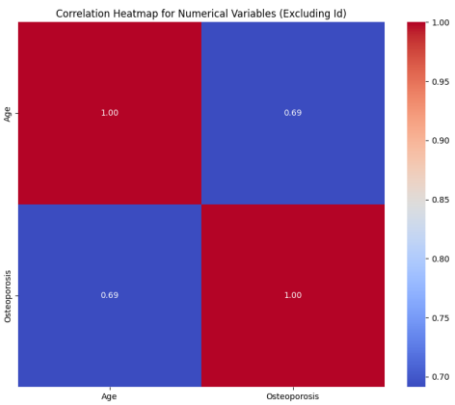
Appendix

1. **Full Analysis Code**: The full Jupyter Notebook containing the all steps, can be accessed at the following link: [Link to Jupyter Notebook].

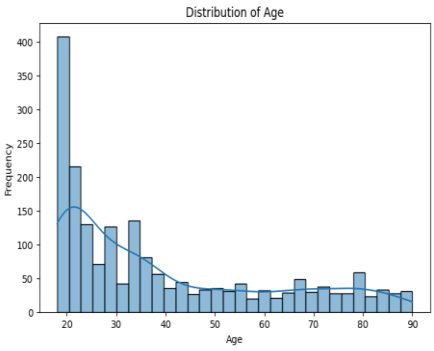
2. **Figures**:



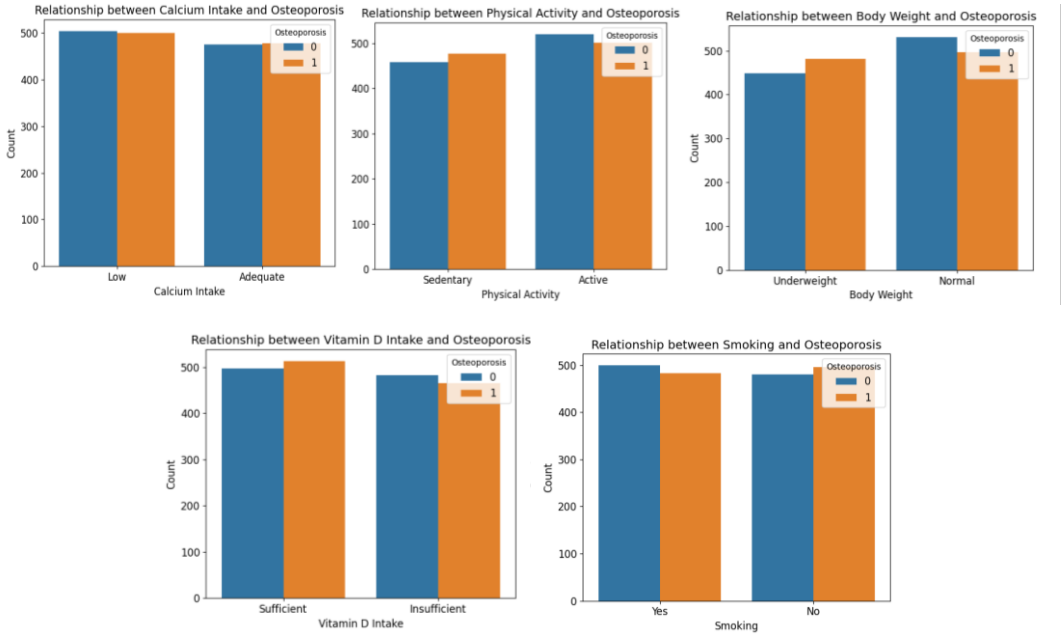
[Figure 1: Boxplot showing the relationship between age and osteoporosis]



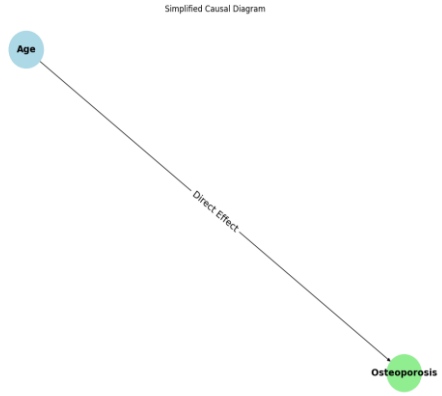
[Figure 2: Correlation heatmap for numerical variables]



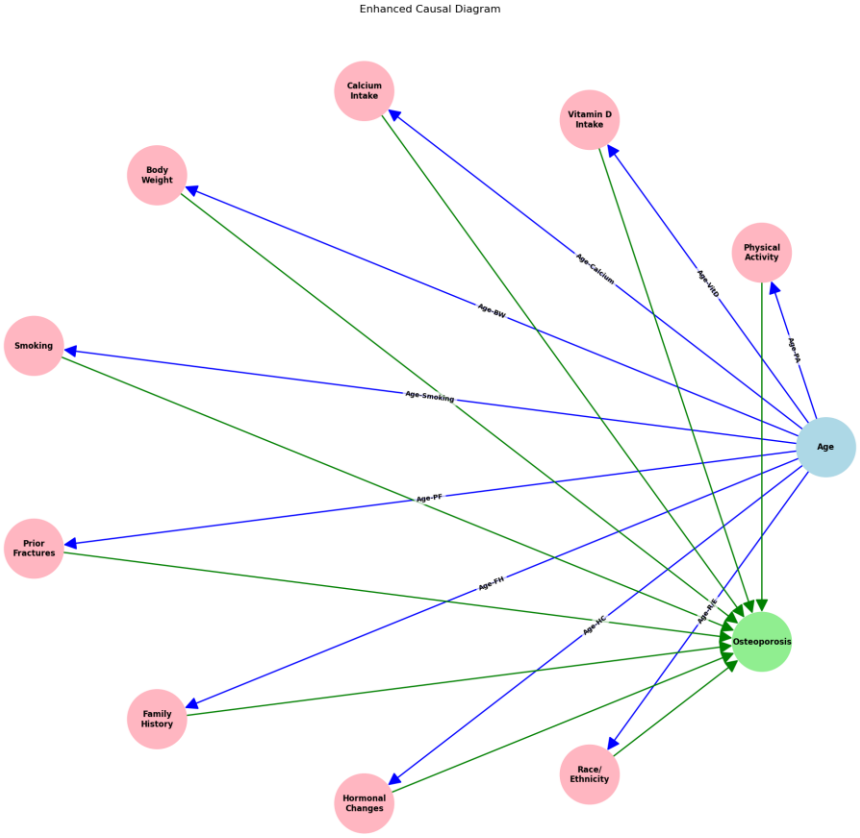
[Figure 3: Histogram showing the distribution of age]



[Figure 4: BarChars for categorical variables]



[Figure 5: Initial causal diagram (DAG) highlighting primary relationships]



[figure 6: Enhanced causal diagram (DAG) with interaction effects]