Capstone Two Final Project Report

A. Karadsheh

Exploring NHANES 2003-2006 Health and Nutrition Data in Relation to Mortality: Focus on Morphometrics, Activity, and Metabolic Proxies

Introduction

Objective: This project aims to predict all-cause mortality using morphometric measurements, physical activity, and metabolic markers derived from the National Health and Nutrition Examination Survey (NHANES) data for 2003-2004 and 2005-2006 cycles. The central hypothesis is that energy balance, influenced by respiration (via morphometrics, activity, and sleep-related factors influencing breathing and therefore energy balance), drives health outcomes and mortality risk.

Dataset: The analysis builds on data from NHANES, a biennial survey conducted by the CDC, combined with mortality data from Leroux et al(Leroux et al., 2019; Smirnova et al., 2020). The primary dataset, `NHANES_analysis_data.csv` was merged with other data sets from the NHANES site (https://www.cdc.gov/nchs/nhanes/index.html), to finally have 3,198 participants with 121 columns, including demographic, morphometric, activity, laboratory, and mortality data. Datasets used include body measurements (BMX), blood pressure (BPX), demographics (DEMO), and laboratory results, merged and cleaned across multiple notebooks.

Background: Building on Leroux et al.'s work, which demonstrated that accelerometry-derived activity features predict 5-year mortality better than traditional risk factors, this project extends the analysis by incorporating morphometric (e.g., BMI, waist-to-height ratio) and laboratory data (e.g., crp, creatinine, triglycerides) to explore their relationship with mortality. The focus on respiration and metabolism proxies (e.g., activity levels, hypertension) aims to elucidate how energy balance impacts longevity.

Methodology

Data Acquisition and Preprocessing

1. Data Sources (`1_Initial_data_setup.ipynb`, `2_DataAcquisitionandMerging.ipynb`):

- NHANES data for 2003-2004 and 2005-2006 were downloaded in XPT format from the CDC and converted to CSV using SAS Universal Viewer.

- Key files included:

- Mortality and Activity Data: Derived from Leroux et al.'s(Leroux et al., 2019; Smirnova et al., 2020) dataset (`NHANES_analysis_data.csv`), containing 3,198 eligible participants with accelerometer-derived features (e.g., Total Activity Counts [TAC], Moderate-to-Vigorous Physical Activity [MVPA]).

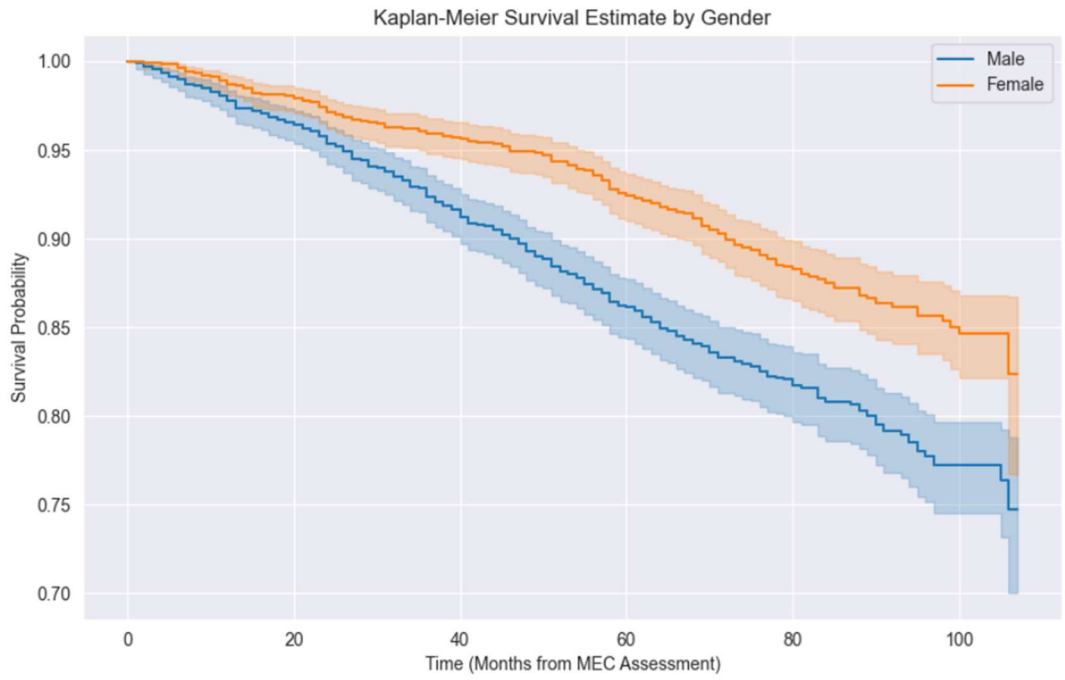

Figure 1 Kaplan-Meier survival graph of the study population. Females trending to survive longer than males in the study participants.

- Body Measurements (BMX): Combined from `BMX_C.csv` (2003-2004) and `BMX_D.csv` (2005-2006), with 19,593 rows and 34 columns after cleaning (`8_MorphometricsDataFileWranglingEDA.ipynb`).

- Demographics (DEMO): Combined from `DEMO_C.csv` and `DEMO_D.csv`, with 20,470 rows and 46 columns (`4_demographics_Files.ipynb`).

- Blood Pressure (BPX, BPQ): Combined from respective files for both cycles (`2_DataAcquisitionandMerging.ipynb`).

- Laboratory Data: Merged from files like `CRP_D.csv` (C-reactive protein) and others, with missing values filled as 'NA' (`9_Mortality_Activity_Morphometrics_Labs_Analysis_Models.ipynb`).

- The unique participant IDs (SEQN) from the mortality dataset were used to filter relevant records from other NHANES files, ensuring consistency (`2_DataAcquisitionandMerging.ipynb`).

2. Data Wrangling (`8_MorphometricsDataFileWranglingEDA.ipynb`, 4_demographics_Files.ipynb`):

- Body Measurements:

- Common columns between 2003-2004 (33 columns) and 2005-2006 (27 columns) were identified, with missing flag columns in 2005-2006 filled with NaN.

- Columns were renamed for clarity (e.g., `BMXWT` to `Weight_kg`, `BMXBMI` to `BMI_kgm2`).

- Flag and comment columns (e.g., `WeightComment`, `HeightLengthDifferenceFlag`) were dropped, reducing to 15 columns, including `ParticipantID`, `Weight_kg`, `Height_cm`, `BMI_kgm2`, etc.

- Missing values were prevalent (e.g., 17,007 missing for `RecumbentLength_cm`, 19,032 for `HeadCircumference_cm`), reflecting age-specific measurements.

- Demographics:

- Combined DEMO files retained 46 columns, with missing values in `RIDEXMON` (468), `RIDAGEMN` (393), and `RIDAGEEX` (831) noted for potential dropping.

- Key variables included `RIDAGEYR` (age in years), `RIAGENDR` (gender), `RIDRETH1` (race/ethnicity), and survey weights (`WTINT2YR`, `WTMEC2YR`).

- Merging:

- The final dataset (`NHANES_mort_bmx.csv`) integrated mortality, activity, morphometric, and demographic data, with missing values in laboratory data (131–165 'NA' per column) imputed using median values.

- The Waist-to-Height Ratio (WHtR) was calculated as a derived feature.

3. Exploratory Data Analysis (EDA) (`8_MorphometricsDataFileWranglingEDA.ipynb`):

- Descriptive statistics for morphometric variables showed:

- `Weight_kg`: Mean 60.31 kg, range 2.4–371 kg.

- `Height_cm`: Mean 148.85 cm, range 82.2–202.7 cm.

- `BMI_kgm2`: Mean 25.37 kg/m², range 12.1–81.1 kg/m².

- Missing values were significant in age-specific measurements (e.g., `HeadCircumference_cm`), likely due to protocol differences for children vs. adults.

- Demographic analysis (`1_Initial_data_setup.ipynb`) revealed:

- Gender: 1,611 males, 1,587 females.

- Race: White (1,862), Black (601), Mexican American (572), Other (102), Other Hispanic (61).

- Correlation analysis was performed on numeric variables to identify relationships.

4. Feature Engineering (`9_Mortality_Activity_Morphometrics_Labs_Analysis_Models.ipynb`):

- Selected 13 features related to breathing and metabolism: `Age`, `TAC`, `WHtR_calculated`, `OSA_Probability`, `LogCRP (mg/dL)`, `Lactate dehydrogenase LDH (U/L)`, `Triglycerides (mg/dL)`, `Uric acid (mg/dL)`, `Creatinine (mg/dL)`, `Age_Group`, `BMI_cat`, `Race`, `HTN_Category_Measured`.

- Categorical variables (e.g., `BMI_cat`, `Race`, `HTN_Category_Measured`) were one-hot encoded.

- Numerical features were scaled using `StandardScaler`.

- Missing laboratory values were imputed with median values to preserve data integrity.

Model Development

- Model: A RandomForestClassifier was used with Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in mortality outcomes (15.6% deceased, `mortstat` mean 0.156).

- Parameters: 200 trees, max depth 20, max features 'sqrt', and balanced class weighting.

- Evaluation: 80/20 train-test split, 5-fold cross-validation, and F1-score as the primary metric.

Results

1. Data Characteristics:

- The final dataset (`nhanes_mort_act_morph_lab_df`) contained 3,198 rows and 121 columns, with 3,067 non-missing values for laboratory markers like `Creatinine (mg/dL)` (mean 0.972, range 0.375–17.39) and `Osmolality (mmol/Kg)` (mean 279.38, range 248–305).

- Mortality rate: ~15.6% (500 deceased out of 3,198), with follow-up time (`permth_exm`) averaging 77.27 months (range 1–107).

- Demographic distribution: Balanced gender (50.4% male), predominantly White (58.2%), with mean age 65.45 years (range 49–84).

2. Model Performance (`9_Mortality_Activity_Morphometrics_Labs_Analysis_Models.ipynb`):

- Accuracy: 0.80.

- F1-Scores:

  - Class 0 (survived): 0.88.

  - Class 1 (deceased): 0.38, improved from 0.14 with SMOTE, reflecting better detection of deceased cases (39 true positives vs. 14 without SMOTE).

- Precision for Class 1: 0.38 (down from 0.39), indicating trade-offs in minority class prediction.

- Cross-Validation: Mean F1-score of 0.20 (±0.06), suggesting potential underfitting.

3. Feature Importance:

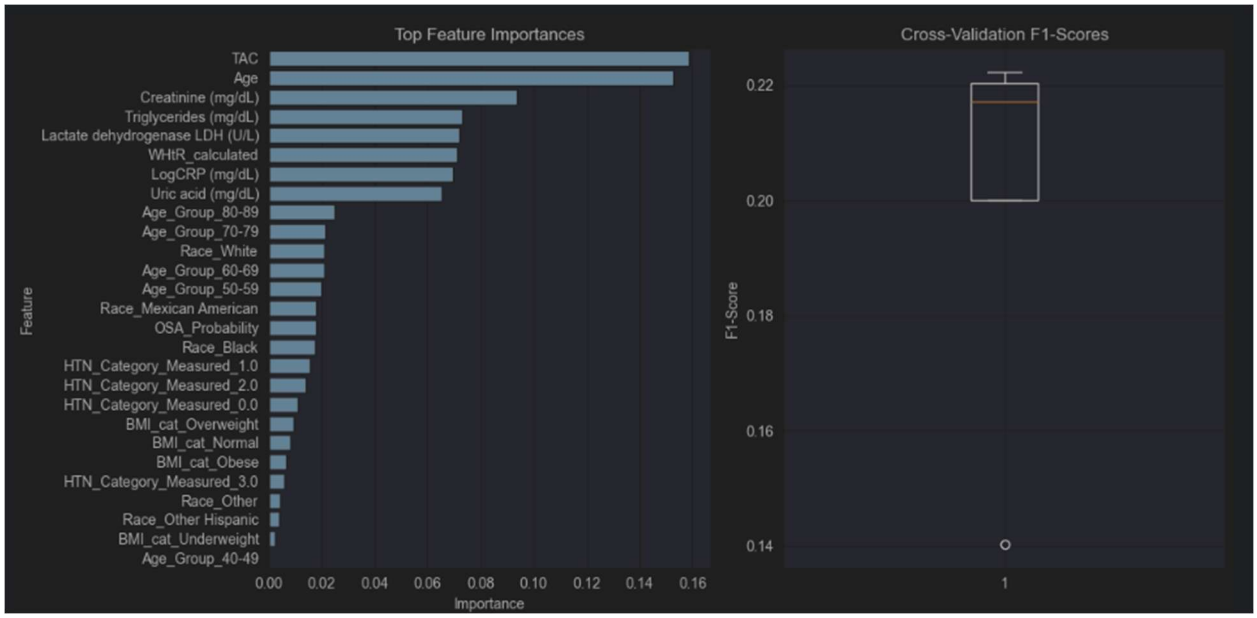

Figure 2Top feature importances represented in the model. Total activity had the highest correlation and impact on mortality, followed by age, and Waist to height ratio. Notable lab measures were triglycerides, lactate dehydrogenase, CRP (logCRP) and Creatinine.

- Top features influencing mortality prediction:

  - `TAC` (Total Activity Counts): 0.1588 (highest, indicating physical activity's strong role).

  - `Age`: 0.1526 (expected, as age is a known mortality risk factor).

  - `Creatinine (mg/dL)`: 0.0938 (renal function marker).

  - `Triglycerides (mg/dL)`: 0.0730 (metabolic stress indicator).

  - `Lactate dehydrogenase LDH (U/L)`: 0.0721 (tissue damage marker).

  - `WHtR_calculated`: 0.0714 (obesity-related breathing insufficiency).

  - `LogCRP (mg/dL)`: 0.0696 (inflammation marker).

  - `Uric acid (mg/dL)`: 0.0653 (metabolic stress).

- Other notable features: `Age_Group_80-89` (0.0247), `Race_White` (0.0210), and hypertension categories (`HTN_Category_Measured_0.0` to `3.0`).

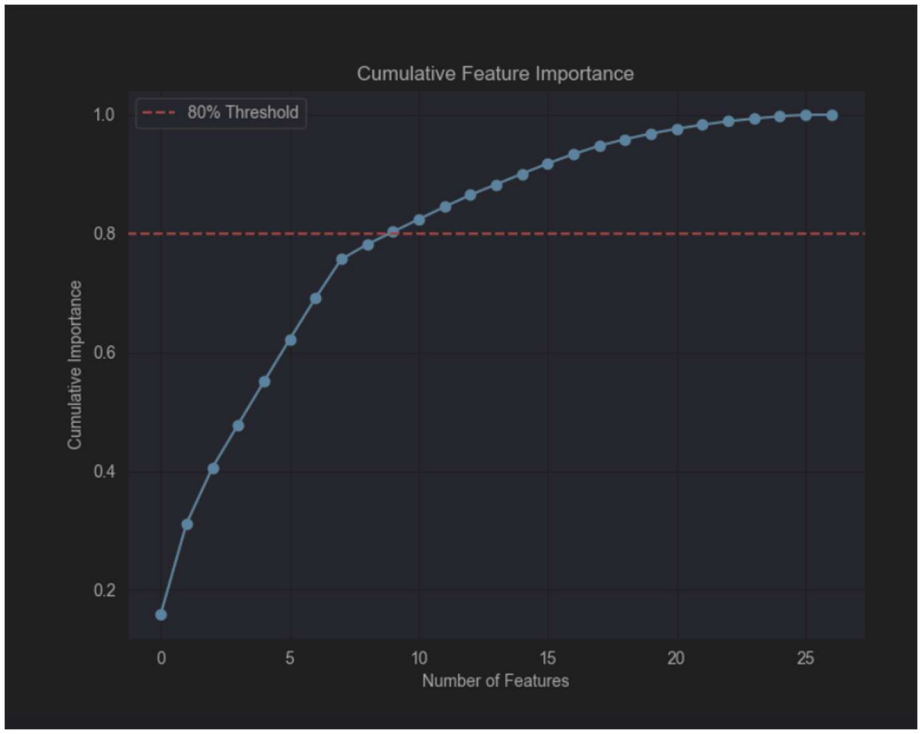

Figure 3 Number of optimal important features. Nine to Ten features provide 80% threshold.

4. Breathing and Metabolism Insights:

- Features like `TAC`, `WHtR_calculated`, and `OSA_Probability` (obstructive sleep apnea probability) reflect breathing insufficiency due to low activity or obesity.

- Laboratory markers (`Creatinine`, `LDH`, `Triglycerides`, `LogCRP`, `Uric acid`) indicate metabolic stress, potentially linked to hypoxia or inflammation, which are amplified by age.

Discussion

Key Findings:

- The model confirms that insufficient breathing (proxied by low `TAC`, high `WHtR_calculated`, and `OSA_Probability`) and metabolic stress (high `Creatinine`, `Triglycerides`, `LDH`, `LogCRP`, `Uric acid`) are significant predictors of mortality.

- SMOTE and balanced class weighting improved recall for the deceased class (from 0.14 to 0.39), identifying more true positives, though precision (0.38) indicates room for improvement.

- The low cross-validation F1-score (0.20 ± 0.06) suggests underfitting, possibly due to feature reduction or insufficient model complexity for capturing mortality patterns.

Limitations:

- The low F1-score for the deceased class (0.38) indicates challenges in predicting the minority class, likely due to class imbalance despite SMOTE.

- Feature reduction (e.g., dropping flag/comment columns, limiting to 13 features) may have excluded valuable signals (e.g., other activity metrics).

- Missing data in age-specific morphometric measurements and laboratory values (131–165 'NA') may have introduced bias, though mitigated by median imputation.

- The study population (age ≥49 years, mean 65.45) limits generalizability to younger cohorts.

Future Work:

- Optimize model hyperparameters (e.g., `min_samples_split`, `n_estimators`) to improve F1-score and reduce underfitting.

- Combine and Incorporate additional features such as occupation and MET (Metabolic Equivalent of Task), estimates prescription medication data (deferred in `2_DataAcquisitionandMerging.ipynb`), to capture more metabolic and behavioral signals and improve prediction.

- Validate findings with clinical expertise to confirm the biological relevance of breathing and metabolic proxies, such as using hypertension as the outcome, instead of mortality, thus expanding the number of participants to study.

- Explore survival analysis (e.g., Kaplan-Meier, Cox regression) using `permth_exm` to model time-to-event outcomes (`9_Mortality_Activity_Morphometrics_Labs_Analysis_Models.ipynb` suggests `lifelines` library use).

- Extend analysis to younger age groups or other NHANES cycles for broader applicability.

Conclusion

This project successfully integrated NHANES 2003-2006 data to predict mortality, duplicating Leroux et al analyses performed in R language (TLAC and TAC activity measures correlating strongly with mortality) (Leroux et al., 2019), and using morphometric, activity, and metabolic features. The Random Forest Classifier, enhanced by SMOTE and balanced weighting, identified key predictors like physical activity (`TAC`), age, and metabolic markers (`Creatinine`, `Triglycerides`, `LDH`, `LogCRP`, `Uric acid`, `WHtR_alculated`), supporting the hypothesis that breathing insufficiency and metabolic stress drive mortality risk. Despite improved recall for deceased cases, the model's precision and cross-validation F1-score indicate underfitting, necessitating further tuning and feature exploration. The cleaned and merged datasets (`NHANES_mort_bmx.csv`, `combined_demo_2003_2006.csv`) provide a robust foundation for future analyses, particularly in validating respiration-metabolism-mortality relationships.

References

Centers for Disease Control and Prevention: About the National Health and Nutrition Examination Survey (2017). URL: http://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Leroux, A., Di, J., Smirnova, E., McGuffey, E. J., Cao, Q., Bayatmokhtari, E., Tabacu, L., Zipunnikov, V., Urbanek, J. K., & Crainiceanu, C. (2019). Organizing and analyzing the activity data in NHANES. *Stat Biosci*, *11*(2), 262–287. https://doi.org/10.1007/s12561-018-09229-9

Smirnova, E., Leroux, A., Cao, Q., Tabacu, L., Zipunnikov, V., Crainiceanu, C., & Urbanek, J. K. (2020). The Predictive Performance of Objective Measures of Physical Activity Derived From Accelerometry Data for 5-Year All-Cause Mortality in Older Adults: National Health and Nutritional Examination Survey 2003-2006. *J Gerontol A Biol Sci Med Sci*, *75*(9), 1779–1785. https://doi.org/10.1093/gerona/glz193