**This document contains information on how to navigate the repo and the process behind creating this final project for BIOS669.**

The primary objective of this project is to reproduce the findings from “A Longitudinal Analysis of Allostatic Load among a Multi-Ethnic Sample of Midlife Women: Findings from the Study of Women’s Health Across the Nation” by Dr. Laura Chyu and Dr. Dawn Upchurch. The original paper is available in the “Additional Files” directory of this repo. Chyu and Upchurch utilized data from the Study of Women’s Health Across the Nation (SWAN) at baseline to create cumulative allostatic load scores across 10 different biomarkers. They also observed trends in allostatic load across the study period. I have derived these allostatic load scores using the Methods section and replicated Tables 1 (Biomarker and Allostatic Load Distribution at Baseline) and Figure 1 (Mean allostatic load by wave) from the original manuscript. All tables and figures, including an additional Figure 2, are available in the “Output” directory.

ICPSR, a research consortium, has all datasets for SWAN available for download as SAS transport files (<https://www.icpsr.umich.edu/web/ICPSR/series/253>). This file type was previously unfamiliar to me, and I originally believed the SAS datasets would be directly available to upload on SAS OnDemand. I have learned that a transport file is distinct from a SAS data set in that it is available in a transport format which was originally exported using the CPORT procedure. As such, it can be successfully imported using the CIMPORT procedure rather than the IMPORT procedure typically used for csv, delimited, and other file types. I imported all baseline through Visit 10 datasets (although only Visits 0, 1, 3, 4, 5, 6, and 7 were used for allostatic load) using PROC CIMPORT and pulled the data needed for my project from each dataset prior to merging and creating a final analysis data set. The codebooks, available from ICPSR, and the raw datasets are available in the “Codebooks” and “Data” directories of this repo.

Of important note, the final analytic data sets were created using the following criteria: women who had valid non-missing baseline data on all 10 biomarkers used to create the AL score, valid AL scores (complete data on all 10 biomarkers) for at least two subsequent visits, and valid (non-missing; not pregnant or breastfeeding) baseline menopausal transition stage information. Additionally, women recruited at the New Jersey site (n=432) were not analyzed in the original analysis because of low retention rate, but since the data which I have used is de-identified for public use, the study site variable is not available through ICPSR and thus all results from this replication will not be an exact match to the manuscript (<https://www.swanstudy.org/swan-research/data-access/process/>). Otherwise, I included only data that met the above criteria.

C-reactive protein, high-density lipoprotein cholesterol, total cholesterol, glucose, and triglycerides were not assessed at visits 2, 8, 9, or 10, so AL scores are unable to be calculated for participants at these visits. The variables corresponding to the 10 biomarkers are listed in the table below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Biomarker** | **BASELINE** | **V1** | **V3** | **V4** | **V5** | **V6** | **V7** |
| Systolic blood pressure | SYSBP10 | SYSBP11 | SYSBP13 | SYSBP14 | SYSBP15 | SYSBP16 | SYSBP17 |
| Diastolic blood pressure | DIABP10 | DIABP11 | DIABP13 | DIABP14 | DIABP15 | DIABP16 | DIABP17 |
| C-reactive protein | CRPRESU0 | CRPRESU1 | CRPRESU3 | CRPRESU4 | CRPRESU5 | CRPRESU6 | CRPRESU7 |
| High-density lipoprotein cholesterol | HDLRESU0 | HDLRESU1 | HDLRESU3 | HDLRESU4 | HDLRESU5 | HDLRESU6 | HDLRESU7 |
| Total cholesterol | CHOLRES0 | CHOLRES1 | CHOLRES3 | CHOLRES4 | CHOLRES5 | CHOLRES6 | CHOLRES7 |
| Body mass index | BMI0 | BMI1 | BMI3 | BMI4 | BMI5 | BMI6 | BMI7 |
| Waist-to-hip ratio | WAIST0, HIP0 | WAIST1, HIP1 | WAIST3, HIP3 | WAIST4, HIP4 | WAIST5, HIP5 | WAIST6, HIP6 | WAIST7, HIP7 |
| Fasting serum glucose | GLUCRES0 | GLUCRES1 | GLUCRES3 | GLUCRES4 | GLUCRES5 | GLUCRES6 | GLUCRES7 |
| Triglycerides | TRIGRES0 | TRIGRES1 | TRIGRES3 | TRIGRES4 | TRIGRES5 | TRIGRES6 | TRIGRES7 |
| Dehydroepiandrosterone | DHAS0 | DHAS1 | DHAS3 | DHAS4 | DHAS5 | DHAS6 | DHAS7 |

My original plan for this project also included replication of Table 2 (Demographic and Other Characteristics at Baseline), but due to the data being de-identified and a re-evaluation of how well it aligned with the replication of Table 1 and Figure 1, I decided that I was more interested in making a completely new figure. I decided to create Figure 2, an extension of Figure 1 with group by race, since allostatic load has been shown to result in worse health outcomes for marginalized populations (<https://doi.org/10.1038/s41598-022-20987-x>). The exact code used to import data, derive allostatic load scores, check the derived variables, create formats and the final analytic dataset, compute summary statistics, and generate reports and figures is all contained within a single SAS program. Although this made the file quite long, I found much easier to work in a single program and used the documentation skills practiced throughout the semester to detail my step-by-step process. This code is available in the “SAS Code” directory and its respective log is in “SAS Log”.

In short, the project included the following steps:

1. Import data using PROC CIMPORT.
2. Subset data to eligible cases (see above) using a series of data steps and data merges with conditional if statements.
3. Determine the quantile cutoff values at baseline using PROC UNIVARIATE and save these results in a usable dataset using ODS OUTPUT, data steps, and PROC TRANSPOSE.
4. Store cutoff values in macros using a data step and CALL SYMPUTX.
5. Calculate allostatic load scores at each visit using a macro data step and condition if statements that referenced the macro quantile cutoff variables.
6. Check derived variables using PROC FREQ and with a PROC PRINT of the first 20 observations, their respective biomarker values, the cutoff values, and their AL scores, and checking that for this selected subset, AL scores seem to be calculated appropriate.
7. Create the final analytic dataset using a data step and merge.
8. Prepare a dataset for Table 1, pulling from a PROC UNIVARIATE with summary statistics, which was exported to a SAS dataset and transposed.
9. Hard code new lines of data in preparation for PROC REPORT.
10. Use PROC REPORT to turn the table 1 dataset into a replication of Table 1 from the original manuscript, using a compute block for style techniques and define statements for additional formatting.
11. Use PROC MEANS to get summary statistics, including confidence intervals, for Figure 1.
12. Using data steps, merge, and the statistics saved from PROC MEANS, prepare data for visualization.
13. Create visualization in PROC SGPLOT, using both vbar and highlow to stack the vertical bar graphs and error bars.
14. Pulled in the race variable from baseline and added to the analytic dataset using PROC SQL to produce a visualization similar to Figure 1 with grouping by race.
15. Created Figure 2 (Figure 1 with mean by visit and race) using PROC SGPANEL to see distribution of AL scores over time by race.

The ADS, DCL, and RPT modules were leveraged heavily throughout this project as a foundation for the skills and techniques used to carry out this project. Overall, while I was unable to replicate the exact findings of the paper due to data access issues, the results that I obtained from both Table 1 and Figure 1 were very similar to what I expected to see based on the original manuscript. Figure 2 also displays an expected trend of racial disparity across AL, with Hispanic and Black Americans having consistently higher allostatic loads on average than White and Asian women.