Presentation draft

I am Bryan Ng, and I am going to introduce our presentation slide 2 of our projects.

Our analysis keeps the original dataset from slide 1. But we have refined our scientific questions to focus on blood-pressure outcomes—systolic and diastolic—and we removed our original Question 2 over concerns about reverse causation. The updated questions now ask, first, how demographic factors (for example, sex, age and BMI) influence systolic and diastolic over time after adjusting for education and time—and the questions 1 to 4 as shown here. This is just a brief introduction of the changes to slide 1. Next, we will introduce our exploratory data analysis.

To build intuition for our modeling approach, we began by examining two key visualizations. First, histograms of systolic and diastolic revealed their distributions: diastolic readings cluster tightly in a bell-shaped curve, while systolic readings are more variable and display a right skew with extreme outliers. These patterns highlighted the need for robust modeling methods to non-normal errors. Second, we generated spaghetti plots for a random sample of 100 subjects, overlaying each individual’s blood-pressure trajectory across visits. The strong within-person correlation—and the substantial variability in both baseline levels and time trends between people—confirmed that longitudinal mixed-effects models are necessary to capture both population-level effects and individual deviations.

Building on those insights, we summarized mean systolic and diastolic trajectories by key demographic and behavioral factors. When we overlaid curves stratified by sex, age, BMI, smoking status, and medication use, we observed significant differences: for example, treated participants diverged from untreated group, and smokers trended higher than non‐smokers. These subgroup plots visually confirmed the strong, time‐varying influence of demographics, smoking, and medication on blood pressure. Finally, a correlation heatmap of our predictors showed generally low pairwise correlations—except for total cholesterol vs. LDL cholesterol and systolic vs. diastolic—highlighting modest multicollinearity. So maybe we should do something to address the issue of multicollinearity in our further analysis part.

To address our first three scientific questions, we plan to perform two regression frameworks of longitudinal models.

For Question 1, demographic factors vary by individual, so we’re not only interested in the average effects of Sex, Age, and BMI on blood pressure but also in how each person’s baseline blood pressure and its trajectory over time differ. We therefore fit an LME model, which estimates both the population-level effects and individual-level effects, capturing subject-specific deviations from the overall trend. To address multi collinearity among predictors, we add a ridge penalty to the fixed-effects portion of the model.

For Question 2, our goal is to estimate the population-averaged effect of anti-hypertensive medication use on blood pressure rather than individual-level trajectories. GEE therefore provides a simpler, more reliable interpretation of the marginal drug effect. Again, we introduce a ridge penalty to address multi collinearity.

For Question 3, we aim to estimate the population-averaged effect of current smoking status and smoking intensity on systolic and diastolic blood pressure. Thus, we again employ a GEE framework, clustering on participant ID and specifying an exchangeable working correlation structure. As with Question 2, we incorporate a ridge penalty to curb potential multicollinearity among smoking status, smoking intensity, and additional covariates.

For Question 4, we prefer LME model because it directly captures within-subject correlation, naturally handles unbalanced repeated measurements, and lets us quantify both the average effect of these biomarkers and each individual’s deviation from that average. Moreover, EDA revealed a curved relationship between blood pressure and time, so we model TIME with a natural cubic spline to flexibly fit that nonlinearity while preserving linear tails. Finally, to address multicollinearity among total cholesterol, LDL cholesterol, systolic and diastolic, we add a ridge penalty to the model—stabilizing coefficient estimates without sacrificing the mixed-model framework.