2.

Part a:

Table 1 shows descriptive statistics for subjects over time where cholesterol is measured in mg/dL. Figure 1 shows a time plot of mean cholesterol levels over time by treatment group. Mean cholesterol in both groups is similar in the first four weeks. After 6 weeks the difference in means is 14mg/dL (rounded) and at 8 weeks the difference in means is 21mg/dL (rounded). The variance in both groups increases as time increases, from the smallest value of 230 to the largest value of 1645.

Table : Descriptive Statistics for Aim 1

|  | | **Week 0 Cholesterol** | | | | | | | **Week 2 Cholesterol** | | | | | | | | **Week 4 Cholesterol** | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **N** | | **Mean** | | **Median** | | **Var** | **N** | | **Mean** | | **Median** | | **Var** | | **N** | | **Mean** | **Median** | **Var** |
| **Treatment** | | 15 | | 219.87 | | 218.00 | | 230.12 | 15 | | 211.80 | | 215.00 | | 290.17 | | 15 | | 212.53 | 214.00 | 851.55 |
| **Control** | |
| **Experimental** | | 15 | | 220.53 | | 221.00 | | 251.55 | 15 | | 218.93 | | 217.00 | | 549.07 | | 15 | | 213.07 | 213.00 | 944.50 |
|  | | **Week 6 Cholesterol** | | | | | | | | | **Week 8 Cholesterol** | | | | | | | |
| **N** | | **Mean** | | **Median** | | **Var** | | | **N** | | **Mean** | | **Median** | | **Var** | |
| **Treatment** | | 15 | | 211.00 | | 217.00 | | 1046.71 | | | 15 | | 204.93 | | 222.00 | | 1645.07 | |
| **Control** | |
| **Experimental** | | 15 | | 197.40 | | 192.00 | | 946.40 | | | 15 | | 183.60 | | 173.00 | | 1576.26 | |



Figure : Time Plot for Aim 1

Table 2 shows the percentage of subjects with elevated cholesterol (greater than or equal to 200mg/dL) at each time point. The percentage in both groups at baseline is 87%. At the 8 week time point the percentage in the control standard of care group is 60% and the percentage in the experimental treatment group is 33%.

Table : Descriptive Statistics for Aim 2

| **group** | **N Obs** | **Variable** | **Label** | **N** | **Mean** | **Variance** |
| --- | --- | --- | --- | --- | --- | --- |
| Control | 15 | elevated0 elevated2 elevated4 elevated6 elevated8 | High Cholesterol Week 0 High Cholesterol Week 2 High Cholesterol Week 4 High Cholesterol Week 6 High Cholesterol Week 8 | 15 15 15 15 15 | 0.87 0.80 0.67 0.53 0.60 | 0.12 0.17 0.24 0.27 0.26 |
| Experimental | 15 | elevated0 elevated2 elevated4 elevated6 elevated8 | High Cholesterol Week 0 High Cholesterol Week 2 High Cholesterol Week 4 High Cholesterol Week 6 High Cholesterol Week 8 | 15 15 15 15 15 | 0.87 0.73 0.67 0.40 0.33 | 0.12 0.21 0.24 0.26 0.24 |

Part b.

Aim 1 of the grant proposal seeks to test the efficacy of the new drug therapy with respect to lowering the mean level of total cholesterol (relative to usual care) over an 8 week follow up period. The following regression model is used to test the null hypothesis that the difference in the change of mean cholesterol over time between groups is zero, against the alternative hypothesis that the difference in the change of mean cholesterol over time between groups is -3mg/dL per week. This translates to a hypothesis test of *β*3 being equal to 0 vs being equal to -3.

Model 1: *Yij* = *β*0 + *β*1*Xi* + *β*2*tj* + *β*3*Xitj* + *bi* + *t:ij*

The research team would like to have at least 85% power to detect this difference in mean cholesterol over time between groups, and they assume a one-sided 0.025 type I error rate.

According to the descriptive statistics above, the mean total cholesterol at baseline for both groups is 220mg/dL. The control group in the pilot study has a total decrease in mean cholesterol over 8 weeks of 15mg/dL, which translates to an average of ~2mg/dL per week. The experimental group in the pilot study has a total decrease in mean cholesterol over 8 weeks of 37mg/dL, which translates to an average of ~4.5mg/dL per week.

A mixed effects linear model is used to estimate within-subject variability in baseline cholesterol levels. According to the specification of model 1, the within-subject variability is 264(mg/dL)^2, as shown in table 3.

Table : Covariance Parameters for Model 1

| **Covariance Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cov Parm** | **Subject** | **Estimate** | **Standard Error** | **Z Value** | **Pr > Z** | **Alpha** | **Lower** | **Upper** |
| **g(1,1)** | **id** | 562.03 | 162.06 | 3.47 | 0.0003 | 0.05 | 342.83 | 1086.75 |
| **Residual** | **id** | 263.98 | 34.3427 | 7.69 | <.0001 | 0.05 | 207.75 | 346.70 |

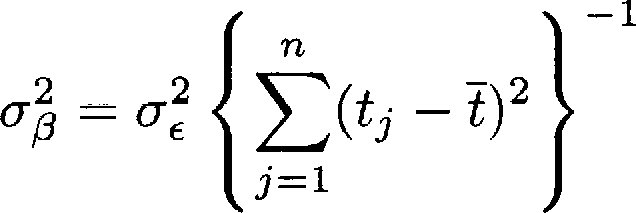
This variability estimate can be used in the formula below to calculate the total sample size needed to detect a difference between treatment groups of -3mg/dL cholesterol per week over the 8 week study. Z(1-alpha/2) is the standard normal CDF assuming a critical region of size alpha/2. In the proposed study the critical region is 0.025, but is one sided, making the value of this Z variable 1.96. Z(1-gamma) is the standard normal CDF where gamma represents the type II error, and thus 1-gamma is the desired power to reject the null hypothesis if the null hypothesis is false. The desired power in this study is 0.85 so the Z variable in this case is 1.04. Sigma squared sub beta represents the variance in the regression parameter estimate beta3.

Equation : Sample Size Calculation for AIM1



Sigma squared sub beta is calculated using the following formula below on the left where sigma squared sub epsilon is the within-subject variability, and the quantity inside the brackets can be calculated using the formula below on the right. In this study the visit time points are equally spaced so this approach is valid. Tau represents the study length, in this case 8 weeks, and n is the total number of visits, which is 5. The remaining quantities are estimated using the pilot data and the model explained above. The value of sigma squared sub epsilon, when using a random intercept, but not random slope, for the pilot data, is estimated to be 264mg/dL.

Equation : Aim 1 Estimate for Variance of Regression Parameter Beta3



Using the above formula, we obtain an estimate for sigma squared sub beta, the estimated variance of the regression parameter beta3, to be 6.6. Using equation 1, we find N to be 26.4. Rounded up to the closest even whole number, 28, we estimate we will need **14 subjects per group** to detect a difference in total cholesterol between treatment groups per week of -3mg/dL.

N = (1.96 + 1.04)^2 \* 4 \* 0.025 \* 264

(-3)^2

As a caveat, model 1 makes the very strong assumption of no correlation between baseline cholesterol and change in cholesterol levels over time. This may not be a scientifically valid assumption since there is a reasonable chance that people who start with a higher or lower cholesterol level will have a different outcome trajectory over time. Using a model that includes a random slope for time, the required sample size is 72.8, where we would need **27 subjects per group** to detect a difference in total cholesterol between treatment groups per week of -3mg/dL if this value is correlated to baseline cholesterol levels.

N = (1.96 + 1.04)^2 \* 4 \* ((0.025 \* 120) + 15.2)

(-3)^2

Part c.

Aim 2 of the grant proposal seeks to test the efficacy of the new drug therapy with respect to decreasing the proportion of patients with elevated (>= 200mg/dL) total cholesterol levels (relative to usual care) over an 8 week follow up period. The following regression model is used to test the null hypothesis that the difference in the rate of change in the log odds of elevated total cholesterol level over time between groups is zero, against the alternative hypothesis that this quantity is -0.15. This translates to a hypothesis test of *γ*3=0 vs *γ*3=-0.15. In this model, Yij is the outcome for subject i at time j, gamma0 is the log odds of elevated cholesterol at baseline across all subjects, gamma1 is the increment to log odds of elevated cholesterol for subjects in the experimental treatment group at baseline, gamma2 is the increment to log odds of elevated cholesterol for subjects in the control group per week in the study, gamma3 is the increment to log odds of elevated cholesterol for subjects in the experimental treatment group per week in the study. Also, x is the indicator variable for treatment group where x=0 is control and x=1 is experimental, and t is a drink with jam and bread but also time as a continuous variable. As a note, based on the fact that this is a randomized controlled trial, and the pilot data show the baseline risk of elevated cholesterol for both groups is equal, we expect the value of gamma1 to be zero.

Model 2: logit*{*Pr(*Yij ≥* 200)*}* = *γ*0 + *γ*1*Xi* + *γ*2*tj* + *γ*3*Xitj*

The research team would like to have 85% power to detect a difference of -0.15 in change in log odds over time between groups, and they assume a one-sided 0.025 type I error rate. The equation for the sample size (N = total for both groups) is:

Equation : Sample Size Calculation for Aim 2



Using equation 3, we need to estimate nu1 (the estimated variance in change of log odds of elevated cholesterol over time for the control group) and nu2 (the estimated variance in change of log odds of elevated cholesterol over time for the experimental group). As per our hypothesis test, we want to calculate the sample size such that we have 85% power to detect a difference in the change in log odds over time between groups of -0.15, so we set delta equal to -0.15.

To estimate the nu values we use the GEE model on the pilot data available to estimate the within-subject correlation (rho) between log odds of elevated cholesterol at different time points. Using an exchangeable correlation structure for the within-subject correlation between outcomes at different time points, we estimate this value to be 0.46. As shown in the descriptive statistics above, the estimated probability of elevated cholesterol at baseline for both groups is 0.87 (hence gamma1 is estimated to be 0), the estimated probability of elevated cholesterol at week 8 for the control group is 0.60, and the estimated probability of elevated cholesterol at week 8 for the experimental treatment group is 0.33. Plugging these probabilities into equations for the estimates of gamma0, gamma2 and gamma3, we mathematically specify a model with which to estimate the variance of the estimate of gamma3. We set gamma0 to 1.90, gamma2 to -0.187, and gamma3 to -0.15 (because this is the difference of the quantity between the null vs alternative hypothesis which we are trying to test). Using these regression parameters and rho equal to 0.46, the GEE model gives us an estimated standard error of gamma3 of 0.29, the square of which will give us the quantity in brackets with nu1 and nu2. Using equation 3 with our set and estimated parameters described above, we get N=33.6, so a total sample size of 34 where there would be **17 subjects per group**.

NOTE: Uses example code from 767 but I’m not sure I used/modified it as intended

title1 "Sample Size, genmod as an asymptotic variance calculator";

ods pdf file="ss06.pdf";

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**data** A;

one = **1**;

p0 = **0.3**; \* E[Y] at time t=0;

p1 = **0.15**; \* E[Y] at time t=2 in the active treatment group;

beta1 = log(p0/(**1**-p0));

beta2 = **0**; \* no time trend in the Placebo group;

beta3 = (log(p1/(**1**-p1)) - log(p0/(**1**-p0))) / **2**; \* H1;

put p0= p1= beta1= beta2= beta3= ;

do id = **1** to **2**;

trt = id - **1**;

do t = **0** to **2** by **0.5**;

eta\_0 = beta1 + beta2 \* t ; \* H0: beta3 = 0;

mu\_0 = **1** / (**1** + exp(-eta\_0)); \* H0;

eta\_1 = beta1 + beta2 \* t + beta3 \* trt \* t; \* H1;

mu\_1 = **1** / (**1** + exp(-eta\_1)); \* H1;

output;

end;

end;

label t = "Time (years)";

label trt = "Treatment (0=placebo, 1=active)";

label beta3 = "Difference in slopes (active - placebo)";

label mu\_0 = "The mean under H0";

label mu\_1 = "The mean under H1";

**run**;

Part d.

In order to address aim 1, researchers will need to randomize at least 14 subjects per treatment group, and to address aim 2, researchers will need to randomize at least 17 subjects per treatment group, and these 34 subjects would need to have data at week 8. Therefore, investigators should recruit 40 subjects total. The investigators of this study do not anticipate subjects will miss many visits, and do not feel the need to adjust sample size due to missing visits. However, it is possible that some subjects may miss some visits, or may drop out. The following is a statistical analysis plan which addresses missing data and drop-out.

Drop-out: For the purposes of this study, drop-out will be defined as any subject who is missing data for both weeks 6 and 8. Missing data at week 8 but not at week 6 does not constitute drop-out. Examining the trends in the pilot data, there is not much of a difference over time between control and experimental treatments until week 6. Therefore, it would be difficult to assess the differences in treatment over time if subjects do not have either week 6 or week 8 data. We will analyze whether the population of people who dropped out have a significantly different outcome trajectory from those who did not drop out by performing a longitudinal regression analysis with cholesterol levels as the outcome from all available data weeks 0-4. Complications may arise if sufficient 8 week data are missing, resulting in unbalanced numbers of observations between groups for addressing aim 2. Based on the trajectory of responses in each group it may not be scientifically reasonable to use a last-observation-carried-forward approach to impute week 8 data from week 6 data, so to address aim 2, week 8 data will be required.

Missing data: It is assumed that all subjects have baseline data. Both the multivariate linear regression and the GEE analysis are robust to small amounts of missing data and should provide variance estimates in line with the pilot data, and should not have too much bias as long as subjects have at least baseline data, and data from week 6 or week 8. Since other covariates are not included in the model, it is not reasonable to use a multiple imputation method to impute missing data.

NOTE: Most sample size calculations were done by hand on paper and/or using R as a calculator. I don’t have code for the calculations.