BIOST 540 Final

Katie McFarlane

Introduction

Candidiasis fungus, commonly known as a yeast infection, is an infection that can grow in many parts of the body, such as the mouth or vagina. Those with weakened immune systems due to HIV can be more susceptible to yeast infections. According to the Center for Disease Control (CDC), those with a CD4 count lower than 200 cells/mm^3 are at greater risk of contraction, but anti-retroviral therapy usage can decrease the chances of getting a yeast infection. Data was collected in a clinical trial setting from HIV-positive adult patients assigned female at birth to compare preventative treatments for mucosal candidiasis. Each patient was followed for up to a year, with follow-up visits taking place every three months to determine the presence or absence of candidiasis fungus. I am interested in analyzing the differences between these treatments on the patients and effects over time to determine the efficacy at preventing mucosal candidiasis.

Methods

Descriptive Analysis: I first present descriptive statistics for patients in each treatment group, shown as either the total number of participants per category per treatment group, or the mean and standard deviation where applicable (Table 1). I also analyzed missing data for the outcome variable both graphically and numerically: see Figure 1 for a histogram and display of the pattern of missingness across data points and see Table 2 for a table of proportions of individuals with the presence of candidiasis fungus across treatment group, as well the amount of missing data per measurement. Additionally, proportions of patients with characteristics at baseline were calculated for each treatment group at every three-month period to indicate the types of patients who may have dropped out or dropped back in at each follow-up period. *Confirmatory Analysis*: I utilized a Generalized Linear Mixed Model (GLMM) with robust standard errors and random slopes and intercepts to investigate the relationship between treatment type and the prevention of mucosal candidiasis, with variables for treatment type, time (using a variable indicating the month of follow-up), and an interaction term between the two. Time is treated as a continuous variable to simplify interpretations and allow for extrapolation. All models were used with the assumption that this trial was randomized, and as such confounders were not adjusted for. Missing data will be handled with multiple imputations using long format data and the method

"2l.bin" to impute missing outcomes, and the fit of these imputations will be assessed graphically (Figure 2). A GLMM approach was chosen so as to make inference on the effects of treatment on individuals using these investigational therapies, to allow for use the logit link function (as the outcome variable is binary) and to avoid making assumptions about the correlation structure. The random slopes and intercepts model was chosen so as not to assume the same underlying response between patients receiving treatment A and those receiving treatment B, and to also allow for different baseline responses for individuals. Wald tests at the 0.05 alpha level will be used to answer the questions of interest. Sensitivity analyses will be conducted through utilizing the same model as above on the observed data without imputations and on best-case scenario imputed outcomes for each treatment (listed in Table 3) to evaluate if missingness assumptions hold and if results are valid when using different methods.

Results

Descriptive Analysis: From Table 1, baseline characteristics differ slightly across treatment groups: notably, the mean CD4 count in the treatment A group is 13 cells/mm^3 higher than in the B group, and 9.4% more participants in the B group used anti-retroviral therapy at baseline than those on treatment A. Additionally, 4.7% more participants in group B had a history of vaginal candidiasis than in group A. Otherwise, baseline characteristics are well-balanced. Proportions of missing data are not balanced across treatment groups, however. Figure 1 illustrates the overall amount of missing data per visit and the pattern of missingness. Missing data is nonmonotone, and the amount of missing data increases over time, with the first follow-up visit at three months having 9% overall missing values, and the final follow-up visit at 12 months missing 28% of data. Table 2 illustrates the percentage of patients on each treatment arm who have a presence of candidiasis fungus at follow-up, along with the percentage of those missing data. Both treatment groups have about 26% occurrence of infection at baseline. Smaller proportions of individuals on treatment B have an infection at each follow-up point; however, there is a larger percentage of missing data on the treatment A arm than on treatment B, with up to 42.2% missing on treatment A, as compared to up to 14.2% on treatment B. By analyzing the percentage of patients with each characteristic at baseline who attended follow-ups, it appears that among patients on treatment A, increasingly less patients with lower baseline CD4 counts came to each follow-up visit, which may hint at the reasoning behind the missing data. *Confirmatory Analysis:* Using the GLMM model with multiple imputations described above, I find that the difference in log odds of

occurrence of candidiasis fungus between one individual on treatment A and one on treatment B is 0.08 (95% Wald-based CI: -0.39-0.55, p=0.74), or 1.08 times greater odds on treatment A, and the additional difference in log odds of occurrence of candidiasis fungus between an individual on treatment A and one on treatment B per month is 0.05 (95% Wald-based CI: -0.02-0.12, p=0.15), or 1.05 times greater odds per month on treatment A (see Table 3). Thus, I do not find evidence that there is a difference in the response between treatments A and B or a significant difference in these responses between treatments over time. Three different models were run as part of a sensitivity analysis: one that did not impute missing values, one that accounted for the best case for treatment A (where all missing values for patients receiving A were 0 and those receiving B were 1), and one that accounted for the best case for treatment B (missing values imputed conversely to the previous model). Table 3 includes all coefficients, standard errors, and p-values for each model. Estimates are similar between the multiple imputation and the non-imputed model, but there are slight differences, so data may not be Missing Completely at Random. Neither best-case model looks that similar to the multiple imputation model. While the non-imputed model showed a significant association between treatment and prevention of mucosal candidiasis over time (p=0.049), this is not echoed in the multiple imputation model. As such, neither best-case model can be used to affirm results about the impact of treatment or treatment over time. However, the imputed models were plotted against observed data (Figure 2), and this diagnostic tool gives no reason to believe this model is incorrectly specified.

Discussion

While conclusions were limited because of missing data, the multiple imputation data suggests that there is no significant association between treatment and prevention of mucosal candidiasis over time, while the observed data suggests that treatment B could be an effective treatment over time (p=0.049). Because we cannot know the reasoning behind missing data, further studies are warranted. The data itself is limited, too, as patients are fairly young, and race is solely grouped into ‘white’ and ‘non-white.’ As such, this information may not generalize to those who are older or not white. It is also possible that a different model was better suited for these analyses: the multiple imputations model is valid under Missing at Random data, which may not hold. Additional methods, such as inverse probability weighting, could be performed on existing data to try and better understand missingness. Thus, further randomized trials are needed, especially larger studies held over a longer period of time to investigate long-term effects of treatment.

Tables and Figures

Baseline Characteristics of Patients

	Trt A (n=161)	Trt B (n=162)
Mean age (SD)	36.0 (7.86)	37.6 (8.17)
Race		
White	25 (15.5%)	23 (14.2%)
Non-white	136 (84.5%)	139 (85.8%)
Mean CD4 count (SD)	228 (165)	215 (153)
Prior progression of disease	41 (25.5%)	40 (24.7%)
Anti-retroviral therapy use	121 (75.2%)	137 (84.6%)
History of vaginal candidiasis	82 (50.9%)	90 (55.6%)
History of oral candidiasis	75 (46.6%)	77 (47.5%)
Presence of candidiasis fungus	42 (26.1%)	42 (25.9%)

Table 1: Patients' baseline characteristics are displayed, divided by treatment. The number of patients with certain characteristics, followed by the percentage this number represents within that treatment group, is displayed for all variables except for patients' age and CD4 counts, where the mean and standard deviation are calculated.

Missing Data Across Follow-Up Visits and Patterns of Missingness

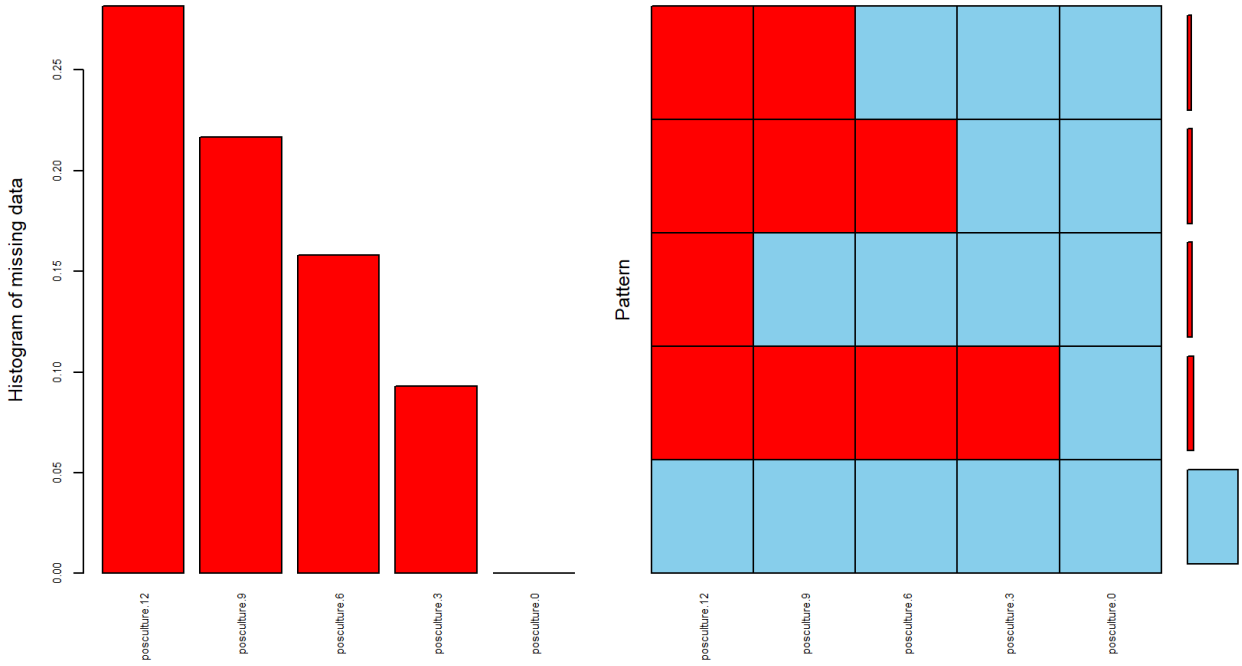

Figure 1: A histogram illustrating the total percentage of missing data at baseline and every

three months up to a year is displayed on the left, while the patterns of missing data is displayed to the right. Most patients have no missing data, as seen in the rightmost figure.

Percentages of Those with a Presence of Candidiasis Fungus

	Trt A (n=161)	Trt B (n=162)
Baseline	26.1%	25.9%
Month 3 (% missing)	24.6% (14.3%)	19.4% (4.3%)
Month 6 (% missing)	24.8% (22.4%)	15.6% (9.3%)
Month 9 (% missing)	21.8% (31.7%)	14.7% (11.7%)
Month 12 (% missing)	14.0% (42.2%)	6.5% (14.2%)

Table 2: For each visit, including at baseline, the percentage of participants in each treatment group with the presence of candidiasis fungus is displayed. In parentheses is the amount of missing data seen at each measurement time in that group. There is no missing data at baseline.

GLMM Analyses with Logit Link

	Intercept			Treatment		
	Estimate	SE	P-value	Estimate	SE	P-value
With imputations	-1.13	0.19	<0.001	-0.08	0.24	0.74
With missing data	-1.11	0.19	<0.001	-0.07	0.24	0.78
Best case, trt A	-1.02	0.19	<0.001	-0.13	0.23	0.57
Best case, trt B	-1.14	0.18	<0.001	0.09	0.23	0.70
	Months			Treatment*Months		
	Estimate	SE	P-value	Estimate	SE	P-value
With imputations	-0.05	0.03	0.09	-0.05	0.04	0.15
With missing data	-0.05	0.03	0.08	-0.07	0.04	0.049
Best case, trt A	-0.20	0.04	<0.001	0.10	0.04	0.009
Best case, trt B	0.10	0.03	<0.001	-0.38	0.05	<0.001

Table 3: Four methods were used to estimated log-odds for the presence of mucosal candidiasis with treatment, months, and an interaction term between the two as predictors. Methods include a model that used the ‘mice’ package in R to perform multiple imputations, a model that did not impute missing data, a model that used best-case imputations for treatment A, and a model that used best-case imputations for treatment B. This table indicates the coefficient estimate and associated standard error and p-value for each term.

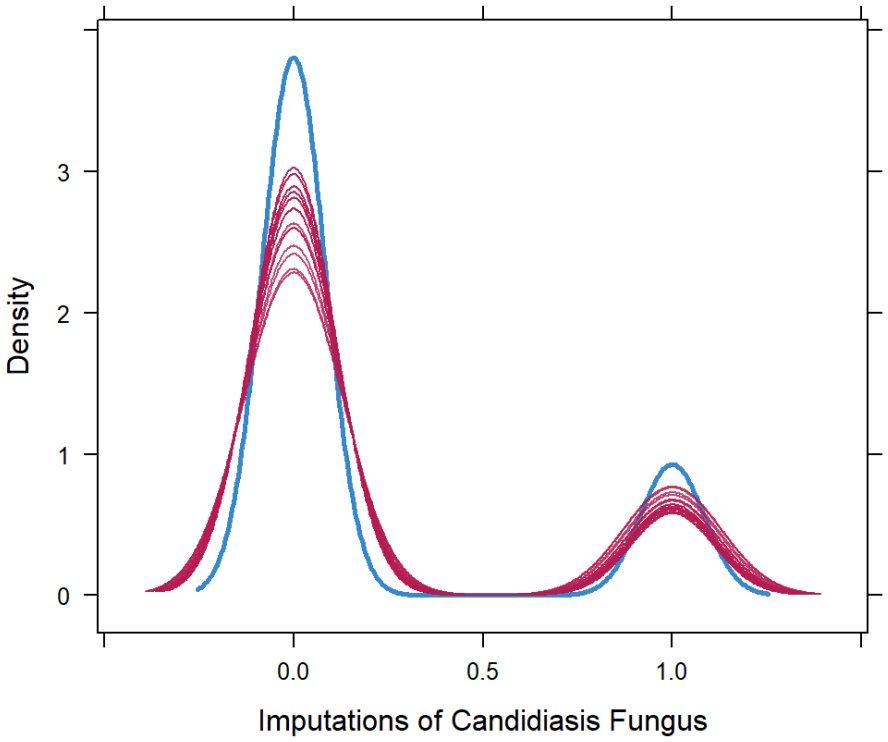

Figure 2: The density of observed data for the presence of candidiasis fungus is shown in blue, while pink lines indicate the densities observed for 20 imputations on the data.