**CHAPTER I**

**INTRODUCTION**

**Research Objectives**

Our project aims to apply logistic regression mixed models with ordinal and binary outcomes to non-tuberculous mycobacterial lung disease (NTM-LD) data. The modeling approaches take inspiration from a previous study done on a preliminary NTM-LD data set but incorporates more complex modeling choices and methods for handling different aspects of the data. The primary research questions we will explore are (1) Are there differences in the frequencies and severities for each of the specific CT features among the lung lobes in NTM-LD, and (2) Are there differences in (1) between mycobacterium avium complex lung disease (MAC-LD) and mycobacterium abscessus lung disease (M. abscessus-LD)? Answering these questions will pave the way to more targeted treatment of NTM-LD and serve as evidence for lifestyle and treatment choices that those with NTM-LD and their healthcare providers should consider in the treatment of this disease.

**Introduction to Non-Tuberculous Mycobacterial Lung Disease**

Non-tuberculous mycobacterial lung disease (which we will refer to as NTM-LD) is an infection of the lungs caused by organisms called, as the name suggests, non-tuberculous mycobacteria. As Miller (1994) and Erasmus et al. (1999) state, infection most often occurs through the inhalation of aerosolized water droplets containing the mycobacteria, although the bacteria can also be found in soil and various animal products like milk or fish, and therefore can also be acquired via ingestion or direct inoculation. Although NTM organisms are widespread across these many sources, most exposures do not result in infections. These are primarily experienced by individuals over the age of 50 or those experiencing underlying lung conditions or immunocompromised individuals. *Mycobacterium avium complex* (MAC) and *M. kansasii* are responsible for the majority of disease cases in the U.S. Two main forms of the disease (Miller, 1994; Erasmus et al. 1999).

**Statistical Background**

**Ordinal Logistic Regression**

Ordinal logistic regression, as the name suggests, allows for the analysis of ordinal outcomes, which are responses that fall into ordered categories (e.g., “no disease”, “mild disease”, “moderate disease”, and “severe disease”). A seminal example of an ordinal logistic regression model is the proportional odds model described by McCullagh (1980). Let be our ordinal variable of interest with ordered categories labeled , and let be the cumulative probability of falling in category or below, given covariates . The proportional odds assumption says that the cumulative odds follow:

where is a category-specific intercept or threshold and is a vector of regression coefficients common across categories. The model ensures through using the same for all cut points that comparing with depends only on and not on , the main idea of the proportional-odds framework. Put more simply, under the proportional odds assumption, the regression coefficients corresponding to our covariates is assumed to be equal across category thresholds. This means that the effect of the covariates on the odds of being at or below a particular category is constant, regardless of the threshold considered.

McCullagh’s model can be further extended to include random effects to account for within-cluster or within-subject correlation. Hedeker and Gibbons (1994) provide a multilevel formulation of an ordinal logistic regression model, where repeated observations are nested within higher-level units ; e.g. subjects or clusters). The repeated, individual observations are referred to as level-1, while the subjects are referred to as level-2. The representation of this ordinal logistic regression model is as follows:

where is the unobserved latent response strength for the observation in level-2 unit , are predictors whose effects vary by subject or cluster (i.e. the random effects), are predictors whose effects are constant across subjects (i.e. fixed effects), and is the residual error. We assume that follows a multivariate normal distribution with mean vector and covariance . This mapping of onto ordinal categories through multiple thresholds allows the model to extend the standard ordinal logistic regression framework while accounting for correlation in repeated measures per subject. This modeling approach reflects the expected heterogeneity across units and allows for more accurate inference to be made with data containing repeated measures or nested structures present in the data.

Agresti (2010) gives another interpretation of a random-intercept ordinal model with cluster-specific (i.e. subject specific) random effects, given by:

where is the ordinal response for observation in cluster , is the random effect for cluster , is the category-specific threshold, are the values of explanatory variables for observation in cluster , with corresponding beta coefficients . Here follows a normal distribution with variance , which describes how individual intercepts scatter around the population-average intercept. Including in the model addresses the fact that responses within the same cluster (e.g. repeated measures within the same subject) tend to be more similar than responses from different clusters. Agresti (2010) further notes that subjects with large positive tend to be more likely to fall into the upper categories of the ordinal scale, where the converse (subjects with negative tend to be in lower ordinal categories) also tends to be the case. The above approach outlined by Agresti aligns with the multilevel framework outlined by Hedeker and Gibbons (1994), where the random intercepts provide a subject-level shift of the ordinal cut points, which accounts for within-cluster correlation.

In a given proportional-odds model, a positive coefficient for a categorical predictor (e.g., a specific lung lobe compared to a reference lobe) indicates that observations in this category have higher odds of belonging to the more severe ordinal outcome category compared to the reference category. We also have that represents the odds ratio for being in outcome category or above. As an example, if we have that , then , meaning the odds of being in a higher ordinal outcome category multiply by 1.65 when an observation falls in that category of . The converse is also true; a negative implies that the category has lower probability of being in the higher-severity ordinal outcome levels. As the proportional-odds assumption uses one slope across all category cut points, the odds-ratio interpretation is the same across each cumulative split of the ordinal scale.

For our study’s ordinal outcomes, subjects’ disease severity is grouped into four ordered categories: “0” indicating no presence of the specified CT feature of interest in the given lobe, “1” indicating less than 25% involvement of the given lobe with the CT feature of interest, “2” indicating between 25 and 50% involvement of the given lobe with the CT feature of interest, and “3” indicating more than 50% involvement of the given lobe with the CT feature of interest. Fitting a proportional-odds model to our data allows us to capture how predictor variables (in our case, which lobe of the lung and which rater is assessing it) affect the likelihood of moving into more severe disease categories. For example, a significant positive coefficient on a specific lobe would indicate that this lobe is associated with a higher probability of falling into more severe disease categories relative to other lobes.

**Logistic Regression**

In contrast with ordinal logistic regression, which involves the analysis of ordinal outcomes, standard logistic regression is concerned with the analysis of dichotomous, or two-level, outcomes (e.g. “Yes or No” or “0 or 1”). Logistic regression models can be understood and expressed in the form laid out by Hosmer et al. (2013): Let be a binary outcome taking values 0 or 1, and let be the probability , where x is a vector of predictor variables. For logistic regression, we let:

which constrains to remain in the interval (0,1) for all values of . Through taking the logit transformation, given by:

we get a linear relationship with the parameters . This is conceptually like standard linear regression, but the outcome now follows a Bernoulli, rather than a normal, distribution.

Also by Hosmer et al. (2013), we know that follows a binomial process with mean , and as we stated before that , we also have that . Since for , we thus have that that which depends on , unlike with linear regression. We typically estimate parameters through maximum likelihood methods, providing coefficient estimates … which maximize the probability of the observed data under the logistic model.

Similar to our random-effects extensions to ordinal logistic models, binary logistic regression can also include random effects to account for unobserved heterogeneity or within-cluster correlation. As outlined by Larsen et al. (2000), let be our binary outcome and . A logistic regression model that incorporates random effects would thus be of the form:

with representing the fixed-effect parameters, the row of the design matrix for the fixed effects, the random effects that are normally distributed with mean 0 and variance matrix , and the row of the design matrix for the random effects. Through incorporating these random effects into our logistic regression model, we can account for correlation among observations that share the same higher-level grouping structure, like repeated measures within individuals or subjects nested within clusters. The random effects capture unobserved heterogeneity across these groups, allowing the model to adjust the log-odds of the outcome based on group-specific deviations from the population-average relationship defined by the fixed effects .