LOGISTIC AND ORDINAL LOGISTIC MODELING OF COMPUTED TOMOGRAPHY FEATURES ASSOCIATED WITH NON-TUBERCULOUS MYCOBACTERIA LUNG DISEASE

by

EDWARD JOHN VICTOR BOSKO

B.S., University of California, Irvine, 2019

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Biostatistics and Informatics

2025

This thesis for the Master of Science degree by

Edward John Victor Bosko

has been approved for the

Department of Biostatistics and Informatics

by

Matthew J. Strand, Chair

Nichole E. Carlson

Edward Chan

Date: August 15, 2025

Bosko, Edward John Victor (M.S., Biostatistics)

Logistic and Ordinal Logistic Modeling of Computed Tomography Features Associated with Non-Tuberculous Mycobacteria Lung Disease

Thesis directed by Professor Matthew J. Strand

**ABSTRACT**

Non-tuberculous mycobacteria lung disease (NTM-LD) is a chronic infection of the lungs caused by the inhalation of microbial organisms called non-tuberculous mycobacteria. NTM-LD is associated with radiologic features that can be observed through CT scans of the lungs, including atelectasis, bronchiectasis, consolidation, ground-glass opacities, tree-in-bud opacities, centrilobular nodules, and cavities. There is evidence in existing medical literature that NTM-LD is more severe and its associated features, especially bronchiectasis, are observed more frequently in certain regions of the lung, namely the right middle lobe and lingula. This analysis seeks to quantify the severity of these radiologic features among 166 subjects’ CT scans and compare this severity across six regions of the lung, namely the right upper lobe (RUL), right middle lobe (RML), right lower lobe (RLL), the left upper lobe assessed in two distinct areas: the left upper segment (LUS) and left lingular segment (LLS), and the left lower lobe (LLL). Nodule and cavity severity was scored on a two-point scale (“0” for absence of the feature and “1” for presence), and the remaining five features’ severity was scored on a four-point scale (“0” for absence of the feature, “1” for 0-25% of the lobe involved with the feature, “2” for 25-50% of the lobe involved with the feature, and “3” for >50% of the lobe involved with the feature). Features with binary scores were analyzed with logistic regression mixed models, while features with ordinal scores were analyzed with ordinal logistic regression mixed models, each incorporating random intercepts for subject and a nested random effect for rater to account for repeated measures over lobes within subjects. The results of the analysis indicate that atelectasis and bronchiectasis were most severe in the RML and LLS; consolidation in the RML and RUL; ground-glass opacities and nodules in the RLL, RUL, and LLL; thick wall cavities in the RUL and RLL; tree-in-bud opacities in the RLL and LLL; while thin wall cavities had no significant differences in lobar severity. These results confirm preferential involvement of lung regions with NTM-LD which can focus surveillance on the most affected regions and inform treatment recommendations.

**TABLE OF CONTENTS**

CHAPTER

**LIST OF TABLES**

TABLE

**LIST OF FIGURES**

FIGURE

**CHAPTER I**

**INTRODUCTION**

**Research Objectives**

Our project aims to apply logistic regression mixed models with ordinal and binary outcomes to non-tuberculous mycobacteria 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 question we will explore is “are there differences in the frequencies and severities for each of the specific CT features among the lung lobes in NTM-LD?” Answering this question will pave the way to more targeted surveillance 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 Mycobacteria Lung Disease**

Non-tuberculous mycobacteria (NTM) are a group of microbial organisms with about 200 species that are related to Mycobacterium tuberculosis and Mycobacterium leprae, the etiologic agents of tuberculosis and leprosy (Matsumoto et al., 2019). NTM lung disease (NTM-LD) is a chronic infection of the lungs and is the most common of the NTM infections. It is most often caused by Mycobacterium avium complex, Mycobacterium abscessus group, and Mycobacterium kansasii (Gopalaswamy et al., 2020). Infection most often occurs through the inhalation or aspiration of aerosolized soil, water, or biofilms that harbor NTM (Miller, 1994). Although NTM organisms are ubiquitous in the environment, most individuals do not develop NTM-LD through exposure. NTM-LD primarily develops in those with predisposing lung conditions or compromised immune function, such as the elderly (Miller, 1994; Erasmus et al., 1999). Individuals with severe immunocompromised states often develop disseminated NTM disease rather than isolated NTM-LD (Liu et al., 2021).

There are two main radiographic manifestations of NTM-LD: a fibrocavitary form characterized by opacification and cavitation of the upper lobes seen mostly in men with underlying emphysema and a nodular bronchiectasis form characterized by nodules and bronchiectasis that commonly affects the right middle lobe and lingula. Depending on the cohort studied, this form may be more frequently seen in women (Miller, 1994).

Erasmus et al. (1999) noted that the disease manifests in a number of radiologic patterns, including consolidation, cavitation, nodules, and bronchiectasis. In the existing literature, the severity and frequency of the disease tends to favor some parts of the lung more than others for some studies, while this was not observed in others.

For example, Reich and Johnson (1992) studied a group of 29 patients with Mycobacterium avium complex pulmonary disease and identified 6 elderly women from this group without predisposing pulmonary conditions with previously unexplained patterns of with a greater predisposition in the lingula and right middle lobe. They hypothesized that suppression of cough in these women may have led to the development of disease in these regions, using the term “Lady Windermere syndrome” to describe the pattern (Reich & Johnson, 1992). Moore et al. (1993) examined 40 culture-positive NTM patients and scored ten lung zones on a 3-point scale (mild, moderate, severe) for bronchiectasis, air-space disease, and nodules. Bronchiectasis was most pronounced in the right middle lobe and lingula, whereas nodules were evenly distributed. Hazelton et al. (2000) reported CT findings in 14 patients with Mycobacterium chelonae. Two radiologists used consensus scoring; bronchiectasis and nodules were present in 13/14 cases but were diffusely distributed across lobes for the most part. Two of these patients, however, did present with more severe bronchiectasis in the right middle lobe and lingula. Lee et al. (2013) compared 369 immunocompromised versus immunocompetent patients. Three radiologists rated each lobe on a 5-point severity scale for lesions with lobar scores summed per patient and scored bronchiectasis by determining its severity and extent and multiplying these values. Group differences were tested with paired t-tests to compare the groups and found that immunocompromised patients had significantly more nodules and cavities. For bronchiectasis, the top three most affected lobes for immunocompromised patients were the right middle lobe, left upper lobe, and lingula, while the top three most affected lobes for immunocompetent patients were the right upper lobe, right middle lobe, and lingula.

**Statistical Background**

**Ordinal Logistic Regression**

Ordinal logistic regression 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 assumed to be independently normally distributed with mean 0 and variance . 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.

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 .

**CHAPTER II**

**METHODS**

**Data**

Our dataset is comprised of data from lung CT images from 166 patients with NTM-LD. These patients were referred to this study conducted by National Jewish Health by their primary care physicians. Each subject’s lung CT images were assessed by one radiologist and one pulmonologist who were blind to each other’s ratings. Lung CT images were rated according to the presence of 8 radiologic features associated with NTM-LD in the CT images. These 8 features include tree-in-bud opacities, large nodules, ground-glass opacities, consolidation, bronchiectasis, atelectasis, and thin and thick wall cavities.

The following are definitions for each of these features provided by Hansell et al. (2008). Tree-in-bud opacities are small branching nodular opacities that look like a budding tree in lung CT images and are a result of mucous impaction and inflammation. Nodules are round opacities measuring up to 3 centimeters in diameter. For the purposes of our study, we consider large nodules to be nodules greater than 1 centimeter in diameter. Ground-glass opacities appear as hazy opacities in CT images of the lung and are caused by fluid in the airspaces of the lungs or thickening of lung tissue. Consolidation, which appears opaquer than ground-glass opacities on CT images, refers to the condition of the lung being rendered solid because of fluid or other disease byproduct replacing the alveolar air of the lungs. Bronchiectasis refers to permanent dilation of the bronchi, or airways, resulting from infection, obstruction, or congenital abnormalities. This can be identified in CT scans from the widening and lack of tapering of bronchi. Atelectasis is the collapse or reduction in inflation of all or part of the lung and is identified in scans through a reduction in volume and increase in opacity. Cavities are gas-filled spaces within areas of consolidation, masses, or nodules. Example CT images for each of these features can be found in Appendix A Figures (Hansell et al., 2008).

For each subject, 6 regions of the lungs were assessed for these features: the right upper lobe (RUL), right middle lobe (RML), right lower lobe (RLL), left upper segment (LUS), left lingular segment (LLS), and left lower lobe (LLL), with the left upper lobe divided into the LUS and LLS per what is standard in the literature for the study of NTM-LD (Hazelton et al., 2000; Lee et al., 2013). We refer to these 6 regions as “lobes” for the remainder of the paper. The raters scored each of the subject’s 6 lobes separately according to the involvement of the feature present in the given lobe. 5 of the features (tree-in-bud opacities, ground-glass opacities, consolidation, bronchiectasis, and atelectasis) were assigned an ordinal score of 0 through 3 for each lobe, with 0 indicating no involvement of the feature in the given lobe, 1 indicating involvement of less than 25% of the given lobe with the feature, 2 indicating involvement of between 25 and 50% of the given lobe with the feature, and 3 indicating more than half of the lobe’s involvement with the feature. These percentages were visual estimates by the raters, not exact measurements taken from the CT images. The other 3 features (large nodules, thin wall cavities, and thick wall cavities) were assigned a binary score of 0 or 1, with 0 indicating the absence of the feature in the given lobe, and 1 indicating its presence. Although ground-glass opacities was assigned a binary score, it exhibited sparseness in the “2” and “3” categories. To facilitate stable model estimation, this feature was dichotomized into a binary variable, with 0 representing no involvement in the given lobe, and 1 indicating involvement in the lobe (i.e. the “1”, “2”, and “3” categories were collapsed into the “1” binary score). With 6 lobes scored for each subject and 2 different raters performing the scoring, there are 12 repeated measures per subject for each feature.

The data was provided by the lead pulmonologist of the study in the form of an Excel document, with separate sheets for the two raters’ scores. Data was arranged in wide format, with each subject’s scores provided in a single row and lobes as columns, with every group of 6 columns representing the scores for a new feature. Minimal data cleaning was performed, including steps to assign each subject with a new ID to deidentify the data and convert the data to long format (making a new row within each subject for each of the 6 lobes) for ease of modeling purposes.

**Modeling**

Our approach to modeling the data for this study was inspired by statistical analyses performed in a pilot study of this data with 71 subjects. In this pilot study, features with ordinal outcomes were modeled using ordinal logistic regression mixed models with a cumulative logit link and random intercepts for subjects. Similarly, features with binary outcomes were modeled using logistic regression mixed models with a logit link and random intercepts for subjects. In the pilot study, a consensus score was achieved between the two raters, meaning only one set of scores was analyzed. In our study, however, we include rater as a variable for our modeling and retain both sets of raters’ scores for our analysis.

**Features with Binary Outcomes**

For the features with binary outcomes (ground-glass opacities, large nodules, thin wall cavities, and thick wall cavities), we modeled using two approaches. The first approach, which we refer to as our full model, used a logistic regression mixed modeling approach, with binary score as the outcome, lobe and rater as fixed effects, a random intercept for subject, a nested random effect for rater within subject, and a logit link function. In a similar way to Larsen et al. (2000), we can express this model using statistical modeling notation as follows. Let be subject, be lobe, and be rater. Our statistical model can thus be expressed as:

,

where is the random intercept for subject with , and is the nested random effect for rater within subject with .

Our second modeling approach, which we refer to as our reduced model, again used a logistic regression mixed modeling approach, with binary score as the outcome, lobe and rater as fixed effects, a random intercept for subject, with a logit link function, but without a nested random effect for rater within subject. Again by Larsen et al. (2000), we can express this model using statistical modeling notation as follows. Let be subject, be lobe, be rater, and be the cut-points between the ordinal categories. Our statistical model can thus be expressed as:

,

where is the cut-point-dependent intercept and is the random intercept for subject with . This reduced approach, like in the ordinal features case, was taken primarily for the binary features where the full model did not converge due to the complex random effects structure. We also fit this reduced model for the features where the full model did converge as a means of sensitivity analysis to check if the simpler random effects structure changes the model estimate for the effect of lobe, which is our main estimate of interest in addressing the research question.

This reduced model also has some inherent trade-offs to keep in mind. The lack of a random rater term means that now rater is only treated as a fixed effect, so the model cannot parse subject versus rater variation, which does not allow for the calculation of ICC as we have outline it in the Interrater Reliability section (see below). Also, with only a single subject-level intercept, the repeated measures structure may also be oversimplified, assuming every observation within a subject is equally correlated.

**Features with Ordinal Outcomes**

For the features with ordinal outcomes (tree-in-bud opacities, consolidation, bronchiectasis, and atelectasis), we modeled using three approaches. The first approach, which we refer to as our full model, used an ordinal logistic regression mixed modeling approach, with ordinal score as the outcome, lobe and rater as fixed effects, a random intercept for subject, a nested random effect for rater within subject, and a cumulative logit link function. This nested random effect for rater produces unique random rater effect estimates for each subject, as opposed to an unnested random effect for rater which would only produce 2 estimates. In a similar way to Hedeker and Gibbons (1994), we can express this model using statistical modeling notation as follows. Let be subject, be lobe, be rater, and be the cut-points between the ordinal categories. Our statistical model can thus be expressed as:

,

where is the cut-point-dependent intercept, is the random intercept for subject with , and is the nested random effect for rater within subject with .

Our second modeling approach, which we refer to as our reduced model, again used an ordinal logistic regression mixed modeling approach, with ordinal score as the outcome, lobe and rater as fixed effects, a random intercept for subject, with a cumulative logit link function, but without a nested random effect for rater within subject. In a similar way to Agresti (2010), we can express this model using statistical modeling notation as follows. Let be subject, be lobe, be rater, and be the cut-points between the ordinal categories. Our statistical model can thus be expressed as:

,

where is the cut-point-dependent intercept and is the random intercept for subject with . This reduced approach was taken primarily for the features where the full model did not converge due to the complex random effects structure. We also fit this reduced model for the features where the full model did converge as a means of sensitivity analysis to check if the simpler random effects structure changes the model estimate for the effect of lobe, which is our main estimate of interest in addressing the research question. Please note that the same caveats we mentioned for our binary reduced model apply here as well.

Our third modeling approach sought to relax the proportional odds assumption outlined by McCullagh (1980) that the other two modeling approaches abide by, thereby allowing the effect of lobe to differ across the category thresholds. We again modeled this using an ordinal logistic regression mixed modeling approach, with ordinal score as the outcome, lobe and rater as fixed effects that we allowed to differ across the category thresholds, a random intercept for subject, and a cumulative logit link function. We can express this model using statistical notation similar to that outlined by Lin et al. (2021). Let be subject, be lobe, be rater, and be the cut-points between the ordinal categories. Our statistical model can thus be expressed as:

,

where is the cut-point-dependent intercept, is the cut-point-dependent coefficient vector for lobe, is the cut-point-dependent coefficient vector for rater, and is the random intercept for subject with .

**Estimation Method**

All models were fit by maximum-likelihood estimation. In the case of generalized linear mixed models, the likelihood function cannot be expressed in closed form due to the presence of random effects. These random effects must be integrated out to achieve the marginal likelihood, since the random effects are non-observable values, and this integral with respect to the random effect is what makes the likelihood estimation an intractable problem. Adaptive Gaussian quadrature is a method of numerical integration that approximates this marginal likelihood by evaluating the integrand at a set of abscissas and summing weighted values of the integrand at these points (Pinheiro & Chao, 2006). This approximation can be expressed as follows:

,

where is a probability density function, is a function integrated against it, is the number of quadrature points, are the weights, and are the abscissas. In standard Gaussian quadrature, the integrand, and thus the placement of abscissas, is assumed to be centered at zero and symmetric. Adaptive Gaussian quadrature improves upon this by centering and scaling the placement of abscissas depending on the density function of the random effects and by incrementally increasing the number of quadrature points until the marginal log-likelihood stabilizes (Capanu et al., 2013; Pinheiro & Chao, 2006).

**Comparison of Model Fit**

To assess how well each of the modeling approaches fits our data, we calculated the Akaike Information Criterion (AIC), given by:

,

where is the maximum likelihood of the model given the data evaluated at the estimated parameters, and is the total number of estimated parameters. AIC offers a balanced view of the model between goodness-of-fit and model complexity, as the term penalizes models with more parameters to avoid overfitting (Burnham & Anderson, 2004).

**Software Used**

The non-proportional odds model for our ordinal features were fit and all analyses for this model was performed in R Version 4.4.2 using the clmm2 function. All other full and reduced models were fit in SAS Version 9.4 using the GLIMMIX procedure, along with all analysis performed for these models.

**Comparisons Between Lobes**

**Features with Ordinal Outcomes**

To address the research question regarding the differences in the frequencies and severities for each of the specific CT features among the lung lobes in NTM-LD, we constructed pairwise contrasts of the fixed-effect estimates for the lobe variable in our models using “estimate” statements for our proportional odds models fit in SAS. As our ordinal logistic regression models compare lower levels to higher levels by default (in other words, lower disease severity to higher disease severity), we first flipped the signs on these estimates before calculating the contrasts to get the comparisons in the other direction. We then exponentiated the results of these contrasts to get the odds ratios of being in a higher disease severity category between the lobes. Confidence intervals for these odds ratios were calculated by exponentiating the Wald 95% confidence limits produced by GLIMMIX.

For the non-proportional odds model, which produces three fixed-effect estimates for the 5 non-reference lobes (one for each severity cut-point), pairwise contrasts were constructed between each of the lobes within each cut-point, again by subtracting the corresponding log-odds estimates. Standard errors were determined through the model’s variance-covariance matrix, Wald 95% confidence intervals were formed, and resulting estimates were exponentiated to yield values on the odds scale.

**Features with Binary Outcomes**

Similar to the features with ordinal outcomes, we performed lobe-specific analyses for those with binary outcomes. Pairwise contrasts were constructed for each of the estimates for the lobe variable in our models. Unlike with our ordinal logistic regression models, we did not flip the signs on these estimates before constructing the contrasts. Confidence intervals for these odds ratios were again calculated by exponentiating the Wald 95% confidence limits produced by GLIMMIX.

**Multiple Testing Correction**

As there are 6 lobes we are performing pairwise contrasts for, there are 15 unique pairwise contrasts performed in total for the models with a single fixed-effect estimate for the lobe variable and 45 unique pairwise contrasts performed in total for the non-proportional odds model with 3 fixed-effect estimates for the lobe variable. In order to correct for these multiple tests, we used the False Discovery Rate (FDR) multiple testing correction approach outlined by Benjamini and Hochberg and provided below.

Suppose we test hypotheses based on corresponding p-values , let be the ordered p-values, and let be the null hypothesis corresponding to . For independent test statistics, the following procedure controls the FDR at : let be the largest for which , then reject all (Benjamini & Hochberg, 1995).

As we are evaluating these tests at a nominal significance threshold of , our FDR-corrected p-values are thus , and hypotheses with were determined to be significant.

**Interrater Reliability**

Two raters performed scoring of each subject’s lung CT images. Since rater is a variable included in our analysis, we were interested in the interrater reliability, or the degree of agreement between both raters’ scores for subject. To determine interrater reliability, we calculated the interclass correlation coefficient (ICC) related to rater agreement for our models that were able to fit both random intercept for subject and the nested random effect for rater within subject. We calculate the ICC for our ordinal features drawing on the method outlined by Liljequist et al. (2019): let be the random intercept estimate for subject (representing between-subject variability) and be the random effect estimate for rater nested within subject. For interrater reliability, we want to know how much of the total variation in scores is attributable to true differences between subjects as opposed to variation due to raters or random error. Our specific interest in inter-rater reliability leads us to define the ICC as the proportion of total variance attributable to between-subject differences. Our formula for ICC is thus (Liljequist et al., 2019).

To determine the 95% confidence interval for our estimate of the ICC, we used a Delta Method approach proposed by Casella and Berger (2002). Let be random variables with means and let and . For a differentiable function , an approximate estimate of the variance is given by a first-order Taylor series expansion of about as

,

(Casella & Berger, 2002). In our case, we have parameters, where is our estimate for , is our estimate for , and is our estimate for ICC. We thus have that and , both by the quotient rule. Expanding the Casella and Berger formula for our case yields

,

where the , , and terms are the estimated variance and covariance of the variance component estimates, obtained from the covariance matrix provided by the GLIMMIX procedure. The standard error (SE) of the ICC estimate was obtained by taking the square root of . An approximate 95% confidence interval (CI) for the ICC was constructed as .

We also applied Cohen’s kappa, which is a measure of agreement for categorical data interpreted as the proportion of ratings with agreement between raters after adjusting for chance agreement (Cohen, 1960). Cohen’s kappa is defined as follows:

where is the observed agreement proportion and is the chance agreement proportion. Chance agreement here is defined as follows:

where is the column 1 marginal, is the column 2 marginal, is the row 1 marginal, is the row 2 marginal, and is the number of observations (see Figure X in Appendix A for a representation of these values) (McHugh, 2012). 95% confidence intervals were calculated with the following formula:

where is determined through the following formula:

(McHugh, 2012). Cohen (1960) provides suggestions for interpreting rater agreement through the kappa values, with values less than or equal to 0 indicating no agreement, 0.01 to 0.2 as no to slight agreement, 0.21 to 0.4 as fair agreement, 0.41 to 0.6 as moderate agreement, 0.61 to 0.8 as substantial agreement, and 0.81 to 1 as near-perfect agreement.

For ordinal measures, we also applied a weighted Cohen’s kappa, which incorporates a degree of disagreement between raters, allowing for partial credit. Unlike unweighted kapa, which treats all disagreements equally, weighted kappa applies weights to each disagreement based on the distance between ratings, thus penalizing large disagreements more heavily than small ones. The formula for weighted kappa is as follows:

where is the weight assigned to the comparison between categories and , is the observed proportion in that cell, and is the expected proportion under chance agreement. 95% confidence intervals were again calculated through the following formula:

where is defined as follows:

(Cohen, 1968). Applying Cohen’s weighted kappa is important for the features with ordinal scores, since a 1-point disagreement is less concerning than a 3-point mismatch. The weights applied during the calculation of the weighted kappa will account for the severity of these mismatches.

Although both the ICC and Cohen’s kappa outlined above provide a measure of interrater reliability, they have different interpretations and are produced from different calculation frameworks. The ICC is model based, extracted from the variance components of our mixed models outlined in the Modeling section, and represents the proportion of total variability attributable to differences between subjects once rater effects have been removed. Cohen’s kappa is a descriptive measure computed from the cross-classification table of raw scores as opposed to the model itself, and is an evaluation of absolute agreement, controlling for chance. By reporting both measures of interrater reliability, we provide a fuller picture of agreement in scores.

**CHAPTER III**

**RESULTS**

**Table 1** displays the demographics of our cohort stratified by the pathogen responsible for their lung disease. A majority (109) of the cohort’s lung disease was a result of NTM belonging to mycobacterium avium complex, 31 had NTM-LD as a result of mycobacterium abscessus complex, 16 had lung disease as a result of both MAC and m. abscessus, and 10 had lung disease as a result of other pathogens. Age group and gender proportions were similar in each group, being largely female and over 65 years of age.

**Comparisons Between Lobes**

We now get into the results for our lobewise comparisons to address the research question regarding the differences in severities of the disease among the lung lobes.

Our atelectasis feature was only able to fit the reduced ordinal logistic regression modeling approach. Based on our lobewise comparisons for this feature, we see that the two lobes that are most severely affected are the right middle lobe and lingula, which are significantly more affected by NTM-LD than the other four lobes analysed. The largest odds ratio was 24.4 (95% CI: 14.9, 39.9) between the RML and RLL, indicating that RML has 24.4 times the odds of being in a higher disease severity category compared to RLL. The results for all comparisons for atelectasis can be seen in Table X.

Figure X is a forest plot of the odds ratios for each of the 15 pairwise comparisons between lobes. The dot in the center is the estimate for the odds ratio, representing the odds of being in a higher severity category compared to the comparison lobe, with the bars on either side of the dot representing the 95% confidence bounds. The x-axis represents the value of the odds ratio, and the y-axis is labeled according to the pairwise comparison in question. The estimates and error bars translucent if the estimate is not significant at the FDR corrected significance threshold.

Figure X in Appendix A is a heatmap of the odds ratios, where each cell is colored a shade of red depending on the magnitude of the associated odds ratio between the reference lobe on the x axis and the compared lobe on the y axis

For the bronchiectasis feature, we were able to fit both the full and reduced models. The estimates for the reduced are shown in blue with triangles representing the estimate of the odds ratio, and the error bars again representing the 95% CIs (Figure X). We see that the estimates for the full model are larger than the estimates for the reduced model, although not significantly so, since their 95% CIs overlap for all comparisons. For this feature, we again see that the RML and lingula are the most severely affected lobes, which is in accordance with what we know from the literature about these lobes’ involvement with bronchiectasis. The largest odds ratio was 22.6 (95% CI: 15.5, 32.8) between RML and LUS, indicating RML has 22.6 times the odds of being in a higher disease severity category compared to LUS.

We were also able to fit the non-proportional odds modeling approach for bronchiectasis, which allowed us to attain 3 different odds ratio estimates for each of the pairwise comparisons between lobes, one for each cut-point between the severity thresholds. The estimates in red in the top panel are the estimates for the odds ratios corresponding to the cut-point between the 0 and 1 severity categories, those in green in the middle are the odds ratios corresponding to the cutpoint between the 1 and 2 severity categories, and those in blue in the bottom panel are the odds ratios corresponding to the cutpoint between the 2 and 3 severity categories (see Figure X).

Based on these odds ratios, we see that the RML and lingula are still the most severely affected lobes. One point to note about our non-proportional odds estimates is that the odds ratio confidence intervals for the RML vs LUS comparison do not overlap when comparing the top to the middle panel and the top to the bottom panel. This suggests that these comparisons significantly differ across the severity thresholds, information which would have been lost if we viewed the proportional odds models alone.

In spite of this, the AIC, our measure which assesses model fit, was lower for the full proportional odds model when compared to the non-proportional odds model, indicating that this significant difference in pairwise comparisons across severity thresholds alone does not improve model fit (see Table X for the AIC for all models).

For the consolidation feature, we were again able to fit both the full and reduced proportional odds models. The odds ratio estimates for the full and reduced models are very similar, even more so than what we saw in bronchiectasis, suggesting that the simpler random effects structure in the reduced model does not significantly affect our fixed effect estimates for lobe, and thus our results regarding the research question (see Figure X). These results indicate that the RML and RUL are the lobes with the greatest severity of disease. The largest odds ratio is 6.5 (95% CI: 4.2, 10.0) for the comparison between the RML and LUS, suggesting that RML has 6.5 times the odds of being in a higher disease severity category compared to LUS.

The ground-glass opacities feature was a feature with ordinal outcome. However, it had very sparse scores in the upper disease severity categories, and thus we had to dichotomize the outcome in order to get the models to fit. This means that the ordinal score of 0 retained a score of 0, and the ordinal scores of 1-3 became a score of 1, and the data was fit using the binary logistic regression approaches. The full and reduced models yield similar estimates. The full model shows that RLL, RUL, and LLL are the three most severely affected lobes, each more severely affected than the lingula and RML. However, only the RLL remains more significantly affected by disease in the reduced models, suggesting we lose important information with the simpler random effects structure (see Figure X). The largest odds ratio had a value of 3.0 (95% CI: 1.6, 5.4) for the comparison between RLL and the lingula, suggesting the RLL has 3.0 times the odds of being in a 1, 2, or 3 disease severity category compared to the lingula

Large nodules is the first feature with a true binary outcome that we have presented so far. This means that its odds ratios represent the odds of the feature being present, since a score of 1 indicates the presence of the feature in the given lobe. We were only able to fit the reduced model for this feature, and the results indicate that the RLL, RUL, and LLL have significantly greater disease severity than the LUS, RML, and lingula (see Figure X). The largest OR was 4.7 (95% CI: 2.5, 9.0) for the RLL vs LUS comparison, indicating the odds of large nodules in the RLL is 4.7 times that of the odds of large nodules in the LUS.

Thin wall cavities are another binary feature which could only be fit with the reduced model. Although there are some odds ratio estimates whose error bars do not cross the dashed 1 line (which indicates equal severity), after applying the FDR multiple testing correction, the results suggest that there are no significant differences in severity by lobe (see Figure X).

Thick wall cavities are our last binary feature, which again could only be fit with the reduced modeling approach. The results indicate that that the RUL and RLL were the most significantly affected lobes, with RUL having greater severity than all other lobes, and RLL having greater severity than all but RUL (see Figure X). The largest odds ratio was 58.3 (95% CI: 17.8, 190.7) for the RUL vs RML comparison, suggesting that RUL has 58.3 times the odds of having thick wall cavities compared to the RML.

Tree-in-bud is our final ordinal feature observed, and we were able to fit both the full and reduced models to the data. The estimates between the two are very close, suggesting the simpler random effects structure does not impact the fixed effect estimates for lobe (see Figure X). The results suggest that the RLL and LLL have significantly greater severity of tree-in-bud feature compared to the LUS, lingula, and RUL. The largest odds ratio is 6.5 (95% CI: 4.6, 9.3) for the comparison between RLL and LUS, suggesting RLL has 6.5 times the odds of being in a higher severity category compared to the LUS.

We were also able to fit the NPO modeling approach for the tree-in-bud feature. The results of the pairwise comparisons confirm that RLL and LLL remain the most significantly affected lobes for this feature. Similar to what we saw with the bronchiectasis NPO results, the odds ratio confidence intervals for the RLL vs LUS, LLL vs LUS, and RML vs LUS comparisons do not overlap for their respective estimates in the top and middle panels. This suggests that these comparisons significantly differ across the severity thresholds, information which would have been lost if we viewed the proportional odds models alone.

Unlike the bronchiectasis NPO model, the tree-in-bud NPO model has a lower AIC than the proportional odds tree-in-bud model. This indicates that the proportional odds assumption does not hold for this feature, and that we lose valuable information by not considering the NPO model results (see Figure X).

**Interrater Reliability**

Table X shows the results for our interrater reliability based on the ICC measure. The closer to 1, the better the agreement between the raters. We see that all of our agreement measures are very high, especially when we look at the raw data for a feature like bronchiectasis and see that one rater tends to rate more severely than the other (see Figures X-X in Appendix A for all features’ raw scores). In the model we fit, rater appears both as a fixed effect, which accounts for any differences in average scoring between raters, and as a random effect nested within subject, which captures the residual variation that remains after those differences have been removed. Thus, the variance attributed to the random rater term is smaller than the total rater variance that would have been present had rater been modelled only as a random effect. This may explain why the ICC is so high, since part of the rater-to-rater variability has already been accounted for by the fixed rater term.

The difference we see in the kappa results in Table X, which are a model-free measure of interrater reliability, highlight this, where the agreement for bronchiectasis is much lower. In spite of the low agreement for bronchiectasis, most estimates for agreement and their 95% confidence intervals fall within the moderate to substantial agreement categories. The weighted kappa, which penalizes less heavily when non-equal ratings are close together, is higher in comparison to all of the corresponding unweighted kappas, suggesting that although perfect agreement was not always achieved, most scores were not far apart on the 4-point ordinal scale.

A close-up of a number

AI-generated content may be incorrect.Table 1: Cohort Demographics

Table 2: Atelectasis Reduced Model Pairwise Comparison Results

A screenshot of a computer

AI-generated content may be incorrect.

Table 3: Bronchiectasis Full Model Pairwise Comparison Results

A screenshot of a computer screen

AI-generated content may be incorrect.

Table 4: Consolidation Full Model Pairwise Comparison Results

A screenshot of a table

AI-generated content may be incorrect.

Table 5: Ground-Glass Opacities Full Model Pairwise Comparison Results

A screenshot of a table

AI-generated content may be incorrect.

Table 6: Large Nodules Reduced Model Pairwise Comparison Results

A screenshot of a computer

AI-generated content may be incorrect.

Table 7: Thick Wall Cavities Reduced Model Pairwise Comparison Results

A screenshot of a data

AI-generated content may be incorrect.

Table 8: Thin Wall Cavities Reduced Model Pairwise Comparison Results

A screenshot of a computer screen

AI-generated content may be incorrect.

Table 9: Tree-In-Bud Non-Proportional Odds Pairwise Comparisons Results for All Cutpoints

A screenshot of a computer

AI-generated content may be incorrect.

A table of numbers with black text

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Atelectasis Proportional Odds Model Forest Plot

A graph with red dots

AI-generated content may be incorrect.

Figure 2: Bronchiectasis Proportional Odds Models Forest Plot

A graph with red and blue dots

AI-generated content may be incorrect.

Figure 3: Bronchiectasis Non-Proportional Odds Model Forest Plot

A graph of a diagram

AI-generated content may be incorrect.

Figure 4: Consolidation Proportional Odds Models Forest Plot

A graph with lines and dots

AI-generated content may be incorrect.

Figure 5: Ground-Glass Opacities Proportional Odds Models Forest Plot

**A graph of a graph

AI-generated content may be incorrect.**

Figure 6: Large Nodules Model Forest Plot

**A graph with red lines and dots

AI-generated content may be incorrect.**

Figure 7: Thick Wall Cavity Model Forest Plot

**A graph with red dots

AI-generated content may be incorrect.**

Figure 8: Thin Wall Cavity Model Forest Plot

A graph with lines and dots

AI-generated content may be incorrect.

Figure 9: Tree-In-Bud Proportional Odds Models Forest Plot

A graph with lines and dots

AI-generated content may be incorrect.

Figure 10: Tree-In-Bud Non-Proportional Odds Model Forest Plot

A screenshot of a graph

AI-generated content may be incorrect.

**CHAPTER IV**

**DISCUSSION**

**Results in Context**

As a high-level overview of our results pertaining to the research question, we found that atelectasis and bronchiectasis were most severe in the RML and lingula, consolidation was most severe in the RML and RUL, ground-glass opacities and large nodules were most severe in the RLL, RUL, and LLL, thick wall cavities were most severe in the RUL and RLL, tree-in-bud opacities were most severe in the RLL and LLL, and thin wall cavities had no significant differences in lobar severity.

The greater severity of atelectasis and bronchiectasis observed in the RML and lingula is unsurprising and reflects findings commonly observed in the literature. Choi et al. (2019) suggest that anatomic features in these lobes, namely more narrowed or more acutely angled bronchi, may be responsible, as narrowed or acutely angled bronchi may be less conducive to clearing out secretions from these lung regions during forced exhalation. Lee et al. (2017) lends credence to this claim, as they found that RML and lingula bronchi angles were more acute in NTM patients than in healthy subjects, and that the angles of the bronchi in segments of the lung involved with NTM were also more acute than those in segments not involved with the disease (Choi et al., 2019; Lee et al., 2017).

Based on our results for the thick wall cavity feature, the RUL seems to have greater severity of thick wall cavities compared to all other lobes. Our PI, Dr. Edward Chan, hypothesizes this might be for a few reasons, and offers some suggestions in the treatment of patients with upper lobe cavities as a result.

Upper lobes have decreased blood perfusion in the upright lung, as the blood must work against gravity to get to the upper lobes, and follows a more tortuous path to the RUL (Powers & Dhamoon, 2023). This results in decreased flow of immune cells and IV antibiotics to the upper lobes. Upper lobes also experience greater physical stress on the lung tissue upon inspiration in an upright lung due to the weight of the lower lobes (West & Matthews, 1972). This may warrant lifestyle or treatment practices, such as having patients refrain from intense exercise, and lying down more when resting or receiving antibiotics.

**Assessment of Rater Agreement**

Our two measures of interrater reliability confirm that lobar scores were generally in agreement between the pulmonologist and radiologist performing the scoring. Our mixed-model based ICC measure (ICC = 0.82-0.96) quantifies the proportion of total variance attributable to subject-to-subject differences after adjusting for the effect of rater. Since rater appears in our model twice, both as a fixed and a random effect, the random rater variance is smaller than it would have been had we only modeled with a random effect for rater, and thus our ICC is higher than it would have been in this case.

Our Cohen’s kappa values (Unweighted = 0.22-0.80, Weighted = 0.55-0.75), which are instead calculated directly from the raw scores, make no model-based adjustments for the effect of rater, and thus provide a more direct, descriptive view of agreement. Since our weighted kappa scores are all larger than their corresponding unweighted scores, we can infer that most disagreements in scores were not far off from each other on the 4-point ordinal scale.

Taking these two measures together, we see that while the raters were not identical in their scoring of the subjects’ lung lobes (based on the kappa scores), most differences in scores for subjects were due to true differences in subject-to-subject lobar severity, as opposed to rater or random error (based on the ICC scores).

**Limitations**

There are also limitations of our study to account for. In several of our models, we could not fit the full random effects structure with both a random intercept for subject and a nested random effect for rater. This may not account for all the correlation due to repeated observations within the same subject. We also noticed, namely in the ground-glass opacities feature, that when comparing the full and reduced model, the reduced model had far fewer significant pairwise comparisons, and only the RLL comparisons remained significant. We thus might be losing valuable information regarding lobar severity by using only the simpler random effects structure. Also, we fit separate models for each feature. This may not account for correlated scores between features if some features are associated or correlated with others. We also observed low rater agreement for bronchiectasis in our unweighted kappa measure of interrater reliability. This may require further examination for the results regarding the severity of this feature, since one rater clearly rates the lobes more severely than the other. Finally, there was some degree of referral bias present in our cohort, as patients were referred to this National Jewish Health study by their primary care physicians. Especially of note is the much larger female representation in our cohort compared to males. Due to this, our results may not be generalizable to a broader NTM-LD population.

In spite of these limitations, our analyses of the data confirm lobar dominance for all CT features but the thin wall cavity feature exhibited in NTM-LD and highlight the need for lobe-targeted severity metrics in the research, prevention, and treatment of this disease.

**REFERENCES**

**APPENDIX A – SUPPLEMENTARY TABLES AND FIGURES**

**APPENDIX B – SAS CODE**

**APPENDIX C – R CODE**