**Introduction**

* Non-tuberculous mycobacteria lung disease is a chronic lung infection most often caused by the inhalation of aerosolized water droplets, soil, or biofilms harboring microbial organisms known as non-tuberculous mycobacteria
* 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
* There are more than 200 species of mycobacteria, two of the most common types being mycobacterium avium complex and mycobacterium abscessus complex. These are responsible for a majority of disease observed in our cohort.
* This disease is associated with several radiographic features that can be observed through CT scans of the lungs
  + Atelectasis: 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. Depicted here is atelectasis of the right middle lobe.
  + Bronchiectasis: 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. Depicted here is varicose bronchiectasis, characterized by irregular dilations of the bronchi (as opposed to cylindrical [more even] and cystic [more sac/pouch like])
  + Cavities: Cavities are gas-filled spaces within areas of consolidation, masses, or nodules
  + Consolidation: Consolidation refers to the condition of the lung being rendered solid because of fluid or other disease byproduct replacing the alveolar air of the lungs.
  + Ground-glass opacities: 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.
  + Nodules: Nodules are round opacities measuring up to 3 centimeters in diameter.
  + Tree-in-bud: 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
* 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. 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.
* Which brings us to our research question
* 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)

**Methods**

* Our data is made up of scores determined by two raters. Unlike many of the studies in the existing literature, where a single score was achieved in a consensus between the raters, we retain both raters’ scores for our analyses. Features with ordinal outcomes are tree-in-bud, ground glass opacities, consolidation, bronchiectasis, and atelectasis, whose scores depend on the percent of the lobe affected by the disease, with 0 meaning no involvement, 1 meaning <25% of involvement, 2 meaning 25-50% involvement, and 3 meaning >50% involvement. These were visual estimates by the raters, not exact percentages. Features with binary scores include large nodules and thin and thick wall cavities, with 0 indicating absence of this feature in the lobe and 1 indicating its presence
* Our study has 166 subjects, with 12 repeated measures per subject (6 scores for each of the lobes and 2 raters scoring each subject). The variables present in our preliminary analysis include the subject’s ID, lobe, score of the lobe, and rater. The data was provided to us in excel document in wide format. Minimal data cleaning was performed, mostly to get the data from wide to long format. I also put each feature in its own sheet for ease of extracting during modeling.
* 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.
* Our second modeling approach, which we refer to as our reduced 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, with a logit link function, but without a nested random effect for rater within subject. This reduced approach 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.
* For the features with ordinal outcomes (tree-in-bud opacities, consolidation, bronchiectasis, and atelectasis), we modeled using three approaches. The first approach, which we again 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.
* 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
* 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.
* Our third modeling approach sought to relax the proportional odds assumption 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.
* To address the research question regarding 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 for our proportional odds models. 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.
* 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.
* Since our data was comprised of scores from two raters, we were interested in determining interrater reliability, or the degree to which the raters agree on scores. For the models in which we were able to fit the full random effects structure (meaning both the random intercept for subject and nested random effect for rater), we determined the interrater reliability via the interclass correlation coefficient (ICC). For ICC, 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. Let sigma squared subject be the random intercept estimate for subject and sigma squared rater be the random effect estimate for rater nested within subject. We thus define ICC as sigma squared subject over sigma squared subject plus sigma squared rater
* Since we weren’t able to fit the full random effects structure for all models, we also calculated Cohen’s kappa as an alternative way to assess interrater agreement. Let P-O be the observed proportion of agreement between raters, and P-E, the expected proportion of agreement by chance. Cohen’s kappa is calculated using the following formula: kappa equals (P-O minus P-E) divided by (1 minus P-E). This adjusts the observed agreement to account for agreement that might occur just by chance. 95% confidence intervals for kappa were constructed as kappa plus or minus 1.96 times the standard error of kappa. Because some of our CT features are ordinal, we also applied weighted Cohen’s kappa, which gives partial credit for close disagreements. For example, a mismatch between scores of 2 and 3 is penalized less heavily than a mismatch between scores of 0 and 3. This makes weighted kappa a better fit when the degree of disagreement matters.

**Results**

* This is a table displaying the demographics of our cohort stratified by the pathogen responsible for their lung disease. A majority 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
* 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
* On the right side of the screen 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.
* On the left 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
* 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 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
* 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. 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% confidence intervals 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 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.
  + 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, an 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
* 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. These results indicate that the RML and RUL are the lobes with the greatest severity of disease. The largest odds ratio is 6.5 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. The largest odds ratio had a value of 3. for the comparison between RLL and the lingula, suggesting the RLL has 3 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. The largest OR was 4.71 for the RLL vs LUS comparison, indicating the odds of large nodules in the RLL is 4.71 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.
* **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. The largest odds ratio was 58.3 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. The results suggest that the right lower lobe and left lower lobe have significantly greater severity of tree-in-bud feature compared to the LUS, lingula, and RUL. The largest odds ratio is 6.5 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.
* These are 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. Because ICC is derived from the variance components of the mixed model, its magnitude depends on how much rater–to–rater variability is left in those components. In the model we fitted 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 only the residual, subject-specific variation that remains after those systematic 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 solely 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, 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 applies partial credit when the two raters differ by one category, 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.

**Discussion**

* 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. Ed 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
  + 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
* This may warrant lifestyle or treatment practices, such as having patients refrain from intense exercise, and lying down more when resting or receiving antibiotics
* 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 of the correlation due to repeate observations within the same subject
  + 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. Due to this, our results may not be generalizable to a broader NTM-LD population.

**Appendix**

* Estimation of the model parameters in all our above generalized linear mixed models was performed through the adaptive Gaussian quadrature method. 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 (quadrature points) and summing weighted values of the integrand at these points.
  + L(theta) marginal likelihood, L(theta|b\_i) is conditional likelihood given random effects, phi(b\_i) is probability density of random effects, theta is placeholder for unknowns
* Suppose we test hypotheses H\_1,H\_2,…,H\_m based on corresponding p-values P\_1,P\_2,…,P\_m, let P\_((1) )≤P\_((2) )≤⋯≤P\_((m)) be the ordered p-values, and let H\_((i)) be the null hypothesis corresponding to P\_((i)). For independent test statistics, the following procedure controls the FDR at q^\*: let k be the largest i for which P\_((i))≤ i/m q^\*, then reject all H\_((i)),i=1,2,…,k (Benjamini & Hochberg, 1995).
* As we are evaluating these tests at a nominal significance threshold of q^\*=α=0.05, our FDR-corrected p-values are thus p\_((i))^FDR=m/i P\_((i)), and hypotheses with p\_((i))^FDR<0.05 were determined to be significant.
* 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 T\_1,…,T\_k be random variables with means θ\_1,…,θ\_k and let T=(T\_1,…,T\_k ) and θ=(θ\_1,…,θ\_k). For a differentiable function g(T), an approximate estimate of the variance is given by a first-order Taylor series expansion of g about θ as
* Var\_θ g(T)≈∑\_(i=1)^k▒〖[g\_i^' (θ)]^2 Var\_θ T\_i 〗+2∑\_(i>j)▒〖g\_i^' (θ) g\_j^' (θ)Cov\_θ (T\_i,T\_j)〗,
* (Casella & Berger, 2002). In our case, we have k=2 parameters, where T\_1 is our estimate for σ\_subject^2, T\_2 is our estimate for σ\_rater^2, and g(T)=T\_1/(T\_1+ T\_2+ π^2/3) is our estimate for ICC. We thus have that g\_1^' (θ)=∂g/(∂T\_1 )= (T\_2+ π^2/3)/(T\_1+T\_2+ π^2/3)^2 =A and g\_2^' (θ)=∂g/(∂T\_2 )= (-T\_1)/(T\_1+T\_2+ π^2/3)^2 =B, both by the quotient rule. Expanding the Casella and Berger formula for our case yields
* Var\_θ g(T)≈(A)^2 Var\_θ T\_1+(B)^2 Var\_θ T\_2+2 (A)(B)Cov(T\_1,T\_2),
* where the Var\_θ T\_1, Var\_θ T\_2, and Cov\_θ (T\_1,T\_2) 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 Var\_θ g(T). An approximate 95% confidence interval (CI) for the ICC was constructed as ICC\_estimate±1.96×SE