Modeling Lung Disease and Its Progression In A COPD Cohort

Modeling Lung Disease and Its Progression In A Chronic Obstructive Pulmonary Disease Cohort

By

LIQIN XU

A Project
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2006) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Modeling Lung Disease and Its Progression in A Chronic Obstructive Pulmonary Disease Cohort

AUTHOR: Liqin Xu

SUPERVISOR: Dr. Lehana Thabane

NUMBER OF PAGES: vi, 87

Abstract

Chronic obstructive pulmonary disease (COPD) is an irreversible, slowly progressive lung disease, usually associated with smoking and with respiratory infections. The objective of the study was to investigate the association of lung function and its progression with smoking, infection and inflammation. I discussed the optimal modeling strategy by comparing results from different methods of analysis and methods of handling missing data.

It was well documented that smoking significantly accelerated lung function decline at a rate of 89.6 mL/year and was robust to adjustment for age, gender, height and breathlessness. C. *pneumoniae* was associated with decreased lung function, but was not statistically significant. Log MMP-9, log ratio of MMP-9 to TIMP-1 and log CRP were strongly associated with slope change in FEV_1 and robust to covariate adjustment in weighted least squares models. Thus, they were good biomarkers for COPD progression.

There were no significant differences among generalized estimating equations, mixed-effects and robust random-effects models for measuring lung function at multiple-time. Smoking was positively related to lung function, but was not robust to covariate adjustment. C. *pneumoniae* was not significantly correlated with lung function. Log MMP-9 and log ratio of MMP-9 to TIMP-1 were associated with decreased lung function, but were not statistically significant. However, log CRP was significantly associated with lung function and robust to covariate adjustment. Thus, log CRP was the best biomarker for modeling lung function measured at multiple- time.

The GEE modeling results of different data sets imputed by group mean estimation, last observation carried forward, hot-deck and multiple imputation methods were consistent, but multiple imputation was recommended to handle the missing data.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Lehana Thabane for his constant support, expert guidance and encouragement towards the completion of this project.

I would also like to thank Dr. Marek Smieja for providing me with the opportunity to work on this research topic, explaining the related background and serving on my committee.

Special thanks to Dr. Roman Viveros - Aguilera and Dr. Noori Akhtar - Danesh for serving on my supervisory committee and giving valuable advices.

Many thanks to Mrs. June Graham for checking my English.

I would like to thank Dr. Peter Macdonald, Dr. Narayanaswamy Balakrishnan, Dr. Aaron Childs, Dr. Ernie Mead and Dr. Abdel EI-Shaarawi for their teaching and guidance.

Thanks to all my friends for their friendships and support.

Finally, I would like to thank my family, especially my parents, my husband and my son whose love, care and support encouraged me to finish the project.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background

Chronic obstructive pulmonary disease (COPD), also called chronic airflow limitation (CAL), including chronic bronchitis and emphysema, is a chronic, irreversible, slowly progressive lung disease. It is characterized by a chronic cough, sputum, airflow limitation and breathlessness. COPD is increasing in prevalence and it is the fifth leading cause of death in Canada and the fourth in the United State [1; 2]. It has been estimated that 750,000 Canadians, about 3 percent of the population, have COPD [3]. Direct and Indirect costs associated with treating COPD are estimated at $3.2 billion in 2002 [4]. Research predicts that by 2020, COPD will be the third leading cause of death worldwide [5].

According to COPD international guidelines [6], COPD patients can be determined by lung function FEV_1 less than 80% of predicted and ratio of FEV_1 to FVC of less than 70%. FEV_1 is the forced expiratory volume at first second. FVC is the forced vital capacity. Research indicates that the normal decline in FEV_1 of COPD patients is from 25 ml to 100 ml per year. But a decline in FEV_1 is estimated at 34 to 62 ml per year for the patients 40-65 years of age [7]. More than 80% COPD cases are caused by cigarette smoking, but not all smokers develop COPD [8]. It is commonly considered that COPD occurs more frequently in genders. However, some research showed that the relative risk of developing COPD was not significantly higher in men than in women [9].

Chlamydia *pneumoniae* (C. *pneumoniae*) infection, as measured by DNA detection in blood or sputum, has been associated with accelerated progression of COPD [10-12]. Some research showed that C. *pneumoniae* DNA existed in 59% of patients with severe COPD [13]. Inflammatory biomarkers, such as matrix metalloproteinases 9 (MMP-9) and its inhibitor, tissue inhibitor of metalloproteinases (TIMP)-1, have been related to COPD by measurement in

sputum [14; 15]. C-reactive protein (CRP), an important inflammatory biomarker, was increased in COPD patients compared with disease-free controls in a meta-analysis [16].

COPD is not a reversible disease but it can be prevented. Smoking cessation is the most important way to prevent COPD for improving symptoms and preventing respiratory exacerbations. On the other hand, antibiotic treatments, vaccination shots for influenza or pneumonia, appropriate medication, oxygen use and surgery are commonly used methods to decrease COPD progression [7].

All abbreviations for medical and statistical terms used in this project were shown in Table 1.1.

1.2 Objectives

The overall goal of this project was to determine the optimal modeling strategy for measuring lung function and its decline by comparing results from different methods of analysis and methods of handling missing data. The project included primary and secondary objectives and objectives of subgroup and sensitivity analysis, as detailed below.

1.2.1 Primary Objectives

The primary objectives were to examine COPD progression measured as yearly change in FEV_1, also called slope change in FEV_1. I compared the results of weighted least squares (WLS), multiple linear regression (WLR) and robust linear regression models (ROBUST) in determining the fitted models and the suitable predictors for slope change in FEV_1. The primary objectives included:

- To determine the relationship between slope change in FEV_1 and smoking.
- To examine the association of slope change in FEV_1 with C. *pneumoniae* infection.

1.2.2 Secondary Objectives

The secondary objectives were to examine baseline lung function and follow-up, called lung function measured at multiple-time, with different statistical modeling approaches. Specifically, I compared the results of generalized estimating equations (GEE), mixed-effects (MIXED) and robust random-effects models (RRE) in determining the fitted models and the appropriate predictors for lung function at repeated-time. The secondary objectives included the following:

- To determine the relationship between baseline lung function and smoking.
- To determine the relationship between baseline lung function and C. *pneumoniae.*
- To explore the association of lung function measured at multiple-time with smoking.
- To explore the association of lung function measured at multiple-time with C. *pneumoniae.*

1.2.3 Objectives of Subgroup Analysis

Epidemiologic research indicates that COPD is strongly correlated with higher inflammatory marker concentrations [17]. There are two biomarkers of particular interest in this study: CRP and MMP-9. Thus, the following objectives in subgroup analysis were proposed:

- To validate MMP-9 as a risk marker for lung function and its decline.
- To confirm the association of lung function and its decline with CRP levels.
- To explore the relationship of lung function and its progression with adenovirus or herpesviridae viral infection.

1.2.4 Objective of Sensitivity Analysis

The sensitivity analysis was based on different imputation methods to handle missing data. In this cohort study, most of patients missed some follow-up visits and made the data set incomplete. I employed three single imputation methods, specifically, last observation carried forward (LOCF), group mean estimation and

hot-deck, and one multiple imputation method to handle the missing data. I compared the results of GEE models based on different imputed data sets in determining the appropriate imputation methods.

1.3 Study design

This was a prospective cohort study with over 2 years of follow-up. There were 200 patients recruited from February 2002 to October 2004, 184 from Firestone Institute for Respiratory Health (FIRH) and 16 from Family Practice Clinics in Hamilton, Ontario. There were a total of 779 visits. We randomly chose 81 subjects from 200 patients as the sub-sample used in subgroup analysis. For each recruited subject, patients' demographics, smoking status, vaccination shots and respiratory symptoms were recorded. Also, lung function was measured by spirometry and samples of blood were collected to assess infection and inflammation for the subgroup sample.

1.3.1 Outcomes

The primary outcome of this observational cohort study was the slope change in FEV_1. This slope was obtained by the estimated coefficients in simple linear regression, where FEV_1 was considered as a function of time variable YEARS, which means yearly interval from any visit to first visit and first visit t_0 set equal to zero.

The secondary outcome was lung function measured by spirometry according to the American Thoracic Society (ATS) specifications. There were two important lung function measures used in this study: FEV_1 and FEV_1PP, which is a percent FEV_1 predicted by age, gender and height. Both of them were continuous variables with units in liters.

1.3.2 Variables

We chose patients' demographics (age, gender and height), smoking status (current smoking, pack years of smoking and quit years of smoking), respiratory symptoms (current cough, productive cough, sputum, sputum amount and breathlessness) and vaccination shots (flu shot, number of flu shots in last 5 years and pneumonia shot) as the predictors (summarized in Table 3.1.1). Among these predictors, age, gender, height, current smoking and breathlessness are not only strongly correlated with lung function, but also are known clinical predictors. Thus, those predictors were always used to adjust all the models.

C. *pneumoniae* infection was considered as explanatory variables in primary and secondary analysis. Inflammatory variables, such as CRP, MMP-9, ratio of MMP-9 to TIMP-1 and viral infection, such as adenovirus or herpesviridae, were explanatory variables in subgroup analysis.

1.4 Inclusion Criteria and Exclusion Criteria

1.4.1 Inclusion Criteria

The subjects who meet all of the followings were included in the study:

- Age 40 to 79 years.
- Current or ex-smoker.
- Smoked $\geq$ 20 packs per year.
- COPD patients with $FEV_1 < 80\%$ predicted and $FEV_1/FVC < 70\%$.
- Able to sign inform consent and likely to attend clinic every 3 or more months.

1.4.2 Exclusion Criteria

The subjects who had any of the followings were excluded in the study:

- $EV_1 < 30\%$ predicted.
- Current exacerbation requiring hospitalization or home oxygen.

- Enrollment in another interventional clinical study.
- Unable to comprehend and sign informed consent.
- Pregnant women or women of child-bearing potential.
- Patients with HIV disease.
- Patients with concurrent major respiratory diagnosis (e.g. pulmonary fibrosis, known or suspected cancer).

1.5 SAMPLE SIZE

The primary objective in the study was to detect the lung function decline measured as the slope change in FEV_1. We assumed that these slopes were independent. However, we justified our sample size in secondary objectives for the correlated data within each patient which requires a large sample size. We used a multivariable generalized estimating equations (GEE) model to determine the relationship of lung function with smoking and infection. To account for possible correlation between measurements within a patient, we need to inflate the sample size [18]. Sample size was calculated by using following formula:

$$n = \frac{(Z_{\alpha/2}+Z_{\beta})^2 \times 2\sigma^2 \times VIF}{\delta^2} \qquad (1.1)$$

where $\frac{\delta}{\sigma}$: relative effect size; α: type I error ; β: type II error; 1-β: test power and VIF was calculated by following formula:

$$VIF = 1 + (m-1) \times \rho \qquad (1.2)$$

where m: average cluster size. ρ: intra-class correlation coefficient (ICC).

We assumed that a medium relative effect size of 1/3, test power of 80% and significant level of 5%, and chose that variation inflation factor (VIF) of 1.40, intra-class correlation coefficient of 0.05 and average cluster size of 9. We can obtain the total sample size of 197 from Table 2.1. Moreover, this multivariable analysis would have about 6 predictors. Simulation research demonstrates that each predictor

requires 10 to 15 events to produce stable estimates for continuous outcome [19; 20]. Therefore, the given sample size of 200 would be sufficient to address the secondary hypothesis. Table 1.3 lists all sample sizes and number of observations in different data sets used in this study.

1.6 SCOPE OF THE STUDY

The project is arranged as follows. In Chapter 2, I first review the research models used in longitudinal data and introduce the general statistical methods used in this study. Then, I address primary, secondary, subgroup and sensitivity analysis with different statistical modeling approaches and various imputation methods. In Chapter 3, I report all the results of unadjusted and adjusted models in primary, secondary, subgroup and sensitivity analyses. I discuss the comparisons of different modeling results and different imputation methods for handling the missing data in Chapter 4. I draw conclusions of appropriate models and predictors for lung function and its progression in Chapter 5.

\

Chapter 2 Methods

2.1 Introduction

2.1.1 Introduction to Modeling Approaches

For independent observations, generalized linear models (GLMs) [21] and quasi-likelihood [22] have been used for discrete and continuous outcomes. For longitudinal or cluster correlated data, Liang and Zeger (1986, also Zeger and Liang 1986), extended GLMs approach to GEE which allows modeling to take into account the correlation of measurement within a patient taken over time [23]. The GEE approach is based on a "working correlation matrix" and only requires specification of marginal mean and covariance functions. We should assume an autoregressive correlation structure (AR1) which is appropriate for longitudinal data [24]. Mixed-effects models (MIXED), also called random-effects models, are appropriate for the study of an individual's growth. Mixed linear models for continuous longitudinal data are in common use [25]. But for non-Gaussian outcomes, mixed generalized linear models have become of research interest recently [26]. Regression estimators from these classical methods are very sensitive to potential outliers or extreme values [27]. The traditional robust regression method of MM-estimation was commonly used for independent data [28]. Thus, it is not suitable for longitudinal data. In STATA 9.1, there are several robust regression models used to analyze longitudinal data. The robust random-effects (RRE) approach was used to fit the data by using generalized least squares estimator [29]. Therefore, we chose GEE, MIXED and RRE models to analyze the data for secondary and subgroup outcomes. For primary outcome of slope change in FEV_1, we used weighted least squares (WLS), multiple linear regression (MLR) and robust linear regression (ROBUST) approaches to model COPD progression.

2.1.2 General Statistical Methods

I first checked obvious data entry errors, then examined for outliers and distribution assumptions by using graphical techniques, such as box plots, histograms and normal probability plots. I also did normality test for distribution assumption by using Shapiro-Wilk Test and Kolmogorov-Smirnov test in SPSS 11.5 software.

Then, I summarized all the outcomes and variables based on original incomplete data by using descriptive measures, expressed as mean (standard deviation [SD]) or median (maximum [max] - minimum [min]) for continuous variables, such as age, height, number of flu shots, sputum amount, breathlessness, inflammatory variables and lung function, and count (percent) for categorical variables, such as gender, current smoking, flu shot, pneumonia shot, sputum, current cough, productive cough and C. *pneumoniae* infection detected at baseline or follow up.

I followed three steps to fit the models. First, I did univariate (or unadjusted) analysis for each variable. Taking account for particular clinical predictors and clinical importance, age, gender, height, current smoking and breathlessness should be used to adjust all the models even though some of them are not statistically significant. Second, we did multiple linear regression models to check the multicollinearity among these variables. Multicollinearity will inflate the variance, standard error and p-value. Variance inflation factors (VIF) are often used as an indicator of the severity of multicollinearity. The value of VIF larger than 10 is of concern. Finally, we did multivariable analysis to build the adjusted models using those significant and independent variables.

All statistical tests were examined using two-sided tests at the 0.05 level of significance. I compared the results of above different statistical modeling approaches. The results were expressed as coefficient, corresponding 95% confidence interval (CI), and associated p-value which was reported to four decimal. It was written as < 0.0001 if it was less than 0.0001. The comparison of different modeling results was assessed by goodness-of-fit test (such as likelihood ratio test and residual analysis) [30]. Different software packages were used in the study,

specified in Table 2.1.

2.2 Primary Analysis

The purpose of primary analysis was to examine the association of slope change in FEV_1 with smoking and C. *pneumoniae* infection. I generated a data set from 200 patients by choosing subjects with at least 3 visits and follow up beyond 6 months. Thus, 139 subjects are interpretable for slope change in FEV_1. Considering some outliers existed in the data set, I employed MLR, WLS and ROBUST models to analyze data and results were compared.

2.2.1 Weighted Least Squares Regression

The weighted least squares regression is a technique for correcting the problem of heterogeneity by weights (w_i) that adjust the errors of prediction. Consider a normal linear model

$$Y_i = X_{ij}\beta_j + \varepsilon_i \qquad (2.1)$$

where i = 1,2,...,n; j = 1,2,...,p. Y_i is an $n \times 1$ vector of slope change in FEV_1; X_{ij} is an $n \times p$ design matrix of predictors including age, gender, height, current smoking, breathlessness, infection or inflammation; ε_i is an $n \times 1$ vector of error term which is assumed as independent and identical normal distribution $N(0, \sigma^2)$, and β_j is a $p \times 1$ unknown coefficient which will be estimated. The coefficient β_j was estimated by minimizing

$$\sum_{i=1}^{n} w_i (Y_i - X\beta)^2 \qquad (2.2)$$

where w_i is the reciprocal of square of standard errors in each simple linear regression, where FEV_1 was treated as a function of time variable YEARS for each patient. If all the w_i's equal one, WLS is the same as MLR.

2.2.2 Robust Linear Regression

Robust linear regression is an important tool for analyzing data with outliers. The main purpose of robust regression is to provide stable results in the presence of outliers. These outliers can exist in x space, y-direction or both.

The ordinary least squares estimates are significantly influenced by a single outlier. When outliers exist in the data, we can not use it for estimating β. In SAS 9.1, Proc ROBUSTREG provides MM method to estimate the coefficients β as following steps [28]:

- Compute an initial value $\beta^{'}$ estimated by the least trimmed squares.
- Find estimate of variance $\sigma^{'}$ such that

$$\frac{1}{n-p}\sum_{i=1}^{n}\chi\left(\frac{y_i - x_i^T\beta^{'}}{\sigma^{'}}\right) = \theta \qquad (2.3)$$

where $\theta = \int \chi(s)d\Phi(s)$ and there are two kinds of functions for χ: Tukey and Yohai [28].

- Minimize $\beta_{MM}^{'}$ of $\beta_{MM} = \sum_{i=1}^{n}\varphi\left(\frac{y_i - x_i^T\beta^{'}}{\sigma^{'}}\right)$ (2.4)

and there are two kinds of φ functions: Tukey and Yohai, corresponding to the functions for χ [28].

2.3 Secondary Analysis

Based on original incomplete data set with 200 patients and 779 visits, we examined the association of lung function measured at multiple-time with smoking and C. *pneumoniae* infection by using GEE, MIXED and RRE models. In addition, we explored the relationship of baseline lung function with smoking and C. *pneumoniae* infection by using MLR models.

2.3.1 Mixed-Effects Models

In mixed-effects model, we treated patient ID as a random effect and patients' demographics, smoking status, respiratory symptoms, vaccination shots, infection and inflammation as fixed effects. The general mixed-effects model [31] is:

$$Y = X\alpha + Z\beta + \varepsilon \qquad (2.5)$$

where Y is a vector containing the multi-responses, X is a design matrix for the fixed effects, α is the vector containing all the fixed effects parameters. Z is a design matrix for the random-effects; β is the vector containing all the random effects variables. But in this study, β has only one variable - patient ID and ε is the vector of random errors.

We assumed that both β and ε were independent normal distributions with zero-mean vector and variance-covariance matrices D and Σ, respectively. Therefore, the response vector Y has a multivariate normal distribution with mean vector $\mu = X\alpha$ and covariance matrix $V = ZDZ' + \Sigma$.

Different methods have been developed to estimate the parameters in covariance matrix V. Likelihood methods are currently in common use. In SAS procedure MIXED, the default estimate method is the residual (restricted) maximum likelihood (REML) estimation which is my choice in this study.

2.3.2 Robust Random-Effects Models

Robust regression models are more powerful to check the accuracy and stability of estimates if a regression model contains many outliers in the data.

In STATA 9.1, there is a robust regression for longitudinal data which employed an extension of random-effects models with GLS (generalized least squares, also called weighted least squares) and a robust approach to estimate the regression intercepts, coefficients, standard errors, 95% confidence intervals and p-values [29]. The GLS has following formula:

$$(Y-w\hat{Y})=(X-w\hat{X})\alpha+(Z-w\hat{Z})\beta+(\varepsilon-w\hat{\varepsilon}) \qquad (2.6)$$

where w is the weight, an inverse function of variance of ε within and between panels. All assumptions of α, β and ε, the meanings of X, Z and Y are the same as mixed-effects models addressed in Section 2.3.1.

The main difference between mixed-effects models and robust random-effects models for longitudinal data analysis is using different methods for estimating parameters of fixed and random effects. The MIXED model used REML method to estimate the coefficients. The RRE model minimized the generalized least squares to obtain the estimators.

2.3.3 Generalized Estimating Equations Models

Liang and Zeger (1986) [23] introduced generalized estimating equations to a regression setting with correlated observations within subjects. Let Y_{ij} represents the j^{th} outcome on the i^{th} subject; X_{ij} is the vector of independent variables for the j^{th} outcome on the i^{th} subject, and β is the vector of regression parameters.

Since the response variable Y_{ij} (FEV_1 or FEV_1PP) was continuous and close to normal distribution, the link function would be identity and marginal regression model should be

$$E(y_{ij}) = \mu_{ij} = X_{ij}\beta \qquad (2.7)$$

We estimated β by solving the generalized estimating equations:

$$\sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) = 0 \qquad (2.8)$$

where $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$, $D_i = \frac{\partial \mu_i}{\partial \beta}$, $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})^T$ and V_i is the working correlation matrix of Y_i, and was decomposed by working correlation matrix.

$$V_i(\alpha) = A_i^{1/2} R_i(\alpha) A_i^{1/2} \qquad (2.9)$$

where A_i is $n_i \times n_i$ diagonal matrices with variance of Y_{ij} as the j^{th} diagonal element and $R_i(\alpha)$ = corr (Y_i) is $n_i \times n_i$ working correlation matrix, and α is the vector of unknown parameters, which is a constant for all the subjects.

The estimate of the variance-covariance matrix of β is:

$$V_G = n(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i)^{-1} \left[\sum_{i=1}^{n} D_i^T V_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} D_i \right] (\sum_{i=1}^{n} D_i^T V_i^{-1} D_i)^{-1} \quad (2.10)$$

In SAS procedure GENMOD, we chose AR (1) as the structure of working correlation matrix to estimate the regression coefficient, corresponding standard error, p-value and 95% confidence interval for each predictor.

2.4 Subgroup Analysis

We randomly chose 81 subjects from total 200 patients for testing the relationship between lung function measured at multiple-time and inflammation with GEE, MIXED and RRE modeling approaches. There were 75 subjects from 81 patients with at least 3 visits and over 6 months follow-up, which were used to test the validity of MMP-9 or CRP as a risk biomarker for COPD progression by using MLR, WLS and ROBUST models. We also examined the relation between baseline lung function and inflammation by applying MLR models.

2.5 Sensitivity Analysis

The data set used in this project was from a prospective cohort study with over 2 years of follow up. Since some patients moved to other cities, got lung cancer or other serious diseases, they quit the follow-up visits. Twenty-eight (14%) patients had one visit, thirty-two (16%) patients had two, thirty-three (16.5%) patients had three and four visits, respectively. The maximum number of visits was 9. I wanted to impute the missing visits and let each subject have 9 visits. Thus, the missing rate for outcome FEV_1 was 56.67% (1021/1800). Some missing values also existed in the predictors,

especially, in the predictor of breathlessness with missing rate 55.8% (116/200).

The most appropriate approach to handle missing data will depend on the missing data mechanisms. There are three types of missing data: missing completely at random (MCAR), missing at random (MAR) and non-ignorable missing. MAR is a commonly used assumption since the missing pattern can be predicted by other variables in the data. We assumed MAR existed in this data and applied single and multiple imputation methods to handle the missing data.

2.5.1 Imputation Methods

Single imputation substitutes a known value for each missing value. Mean estimation, last observation carried forward (LOCF) and hot-deck are commonly used for handling missing values in longitudinal data [32]. Mean estimation replaces missing data with the mean of non-missing values. The standard errors and standard deviations are underestimated. In order to increase the variation in the imputed values, we used group mean estimation to handle the missing values. The data was grouped by each visit of all the patients, the missing values of each group were replaced by the means of corresponding groups. LOCF imputes the missing values by the corresponding known values of last observation if one observation has missing data in some variables. It has the same shortcoming as mean estimation. Hot-deck imputation stratifies and sorts data by key covariates, substitutes missing data from another record in the same strata. Underestimation is still a problem in this method. On the other hand, multiple imputation (MI) combines results from repeated single imputation and requires MAR missing type. Markov Chain Monte Carlo (MCMC) method randomly imputes the missing data. Therefore, it can reduce the variability of results and increase the accuracy of an estimated parameter. In SAS 9.0, PROC MI creates multiple imputed data sets, and PROC MIANALYZE combines results after analysis. The following flowchart is the MI procedure in SAS [33].

Create n imputed data sets and add a variable _IMPUTATION_ to each set Using Proc MI

⇓

Do your standard analysis (GEE models) repeated n times Using BY _IMPUTATION_ statement

⇓

Combine the n sets of results to quantify the estimations due to imputation Using Proc MIANALYZE

Rubin [34] showed that the efficiency of an estimate based on n imputations was

$$\text{Efficiency} = (1+\gamma/n)^{-1} \times 100\% \qquad (2.11)$$

where γ is the rate of missing data and n is the imputation number.

In sensitivity analysis, we used these four imputation methods to impute missing data in the outcomes of lung function and the predictor of breathlessness and compared the results of GEE models based on these imputed data sets. We used age, gender and height as key covariates to impute lung function in hot-deck method.

In primary, secondary and subgroup analyses, we used hot-deck method to impute missing values in the predictor of breathlessness. We chose age, grouped by 40-50 years, 50-60 years, 60-70 years and 70+ years, current smoking and quit years of smoking, grouped by 0-10 years, 10-20 years, 20-30 years and 30-40 years, as key covariates, since they were highly correlated with breathlessness.

Chapter 3 Results

3.1 Summary

Descriptive summaries of outcomes and variables in the analyses were addressed in Table 3.1.1. The statistics of variables in secondary analysis were all based on a cohort of 200 patients, some variables have missing values. The statistics of variables in subgroup analysis were all based on 81 patients, randomly chosen from 200 patients. However, the statistics of lung function measured at multiple-time was based on 200 patients with 779 visits, and the statistics of slope change in FEV_1 was based on 139 patients.

The 200 study subjects had mean (SD) age of 64.3 (9.4) years, 55% were men, and 39.5% were current smokers. Mean FEV_1 was 1.5 (0.6), mean FEV_1 as a percent predicted by age, gender and height was 51.1(15.6) %. Only 139 subjects were interpretable for the slope change in FEV_1. Among these subjects, mean of the slope change in FEV_1 was -19.9 (266) ml/year.

I checked FEV_1 trend over time by plots of FEV_1 with visits. In general, FEV_1 had a decreasing trend over time, even though it increased at some points in time for some patients. See Figure 3.1 for more detail.

I also did normality test of residuals for weighted least squares model by using Shapiro - Wilk or Kolmogorov - Smirnov tests. From the following results, both of p - values were larger than 0.05, we cannot reject the null hypothesis of normal distribution.

Tests of Normality in WLS Model

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RESCHI	.071	127	.200	.985	127	.173

We also did histograms and Q-Q plots for the inflammatory variables CRP, MMP-9

and ratio of MMP-9 to TIMP-1 which were likely to be skewed right in Figure 3.2 and Figure 3.3. Thus, they were transformed by employing a logarithmic transformation.

3.2 Results of Primary Analysis

We assumed that smoking and C. *pneumoniae* infection would lead to lung function decline. Thus, we examined the association of slope change in FEV_1 with smoking and C. *pneumoniae* infection with three different modeling approaches. Specifically, MLR, WLS and ROBUST models and compared the modeling results.

3.2.1 Detection of Outliers

Outlier is an observation that is unusually large or small relative to the other values in a data set. There are two common used methods to detect outliers. One is z-score method using mean and standard deviation to define the distance of outliers from the mean of sample. If z-score is larger than 2 or smaller than -2, the observation can be detected as an outlier. Another method is the box-plot method. Outlier can be defined as a point which falls more than 1.5 times the interquartile (IQR = the third quartile - the first quartile). In this study, I detected outliers by using box plots of residuals for each fitted model. For linear regression models, I found that the subjects having residual values outside the range -0.4 to 0.4 in the models would be outliers. There were 12 outliers in total in the fitted models. I removed all the outliers at one time, and rebuilt the models. The modeling results with and without outliers were very different. For example, in the WLS models, current smoking (-0.0869, 95% CI (-0.1601, -0.0138); p = 0.0203) was significantly associated with lung function decline and predicted a decline at 86.9 ml/year in data set without outliers. However, current smoking (0.0426, 95% CI (-0.0612, 0.1463); p = 0.4186) was not significantly associated with lung function decline in the data set with outliers. Moreover, the values of VIF in gender were 10.3095 and 2.3506 in the data sets with and without outliers, respectively. Thus, WLS models were sensitive to outliers. In addition, after removing outliers, the values of VIF for all the predictors in three models were less

than 10. Thus, there was no significant multicollinearity among these predictors. See Table 3.2.1 and Figure 3.4 for more detail. Therefore, I removed the outliers for these three linear regression models.

3.2.2 Slope Change in FEV_1 with Smoking

Current smoking accelerated FEV_1 lung function decline at significant level 0.05 and was estimated a decline in lung function FEV_1 at 86.9 ml/year in the WLS model. Also, current smoking was significantly associated with decreased lung function in univariate analysis (-0.0753, 95% CI (-0.1383, -0.0123); p = 0.0198). Thus, current smoking was robust to adjustment by age, gender, height and breathlessness. Moreover, gender (0.1325, 95% CI (0.0567, 0.2082); p = 0.0007) and height (-0.0092, 95% CI (-0.0139,
-0.0044); p = 0.0002) were significantly associated with slope change in FEV_1 in the WLS model, and age was a significant factor in decreasing lung function FEV_1 (-0.003, 95% CI (-0.0063, 0.0002), p = 0.0699). See Table 3.2.2 and Table 3.2.3 for more detail.

3.2.3 Slope Change in FEV_1 with C. *pneumoniae*

In the WLS model, C. *pneumonaie* infection (-0.0584, 95% CI (-0.1283, 0.0114); p = 0.1002) potential significantly associated with decreased lung function, but was not statistically significant in other two models. In addition, current smoking (-0.0898, 95% CI (0.1625, -0.0171); p = 0.016) was strongly debilitating lung function at 89.8 ml/year. See Table 3.2.4 for more detail.

3.3 Results of Secondary Analysis

We assumed that smoking and C. *pneumoniae* infection would decrease lung function which was demonstrated in the primary analysis. Now, we want to examine the assumptions by other methods. We considered lung function measured at multiple-

time as the outcome to examine the relationship of lung function with smoking and C. *pneumoniae* infection adjusted by age, gender, height and breathlessness. First, we examined the relationship between baseline lung function, measured at first visit, and smoking, C. *pneumoniae* infection by employing MLR models. Then, we explored the association of lung function measured at multiple-time with smoking and C. *pneumoniae* infection by applying GEE, MIXED and RRE models to determine the association in univariate and multivariable analyses. In addition, we explored the correlation between lung function measured at first visit and other visits of same subject in GEE models.

3.3.1 Baseline Lung Function with Smoking

In univariate analysis, age, gender, height and pneumonia shot were highly related to baseline lung function FEV_1. Current smoking was significantly associated with lung function FEV_1 (0.2746, 95% CI (0.1174, 0.4317); p = 0.0007) and FEV_1PP (4.9169, 95% CI (0.4805, 9.3532); p = 0.03). However, after adjustments for age, gender, height and breathlessness, current smoking was not significantly correlated with baseline lung function FEV_1 (0.0675, 95% CI (-0.0796, 0.2146); p = 0.366), but it was still significantly associated with baseline lung function FEV_1PP (5.1728, 95% CI (0.7101, 9.6354); p = 0.0233). See Table 3.3.1 to Table 3.3.4 for more detail.

3.3.2. Baseline Lung Function With C. *pneumoniae*

There was no significant association of C. *pneumoniae* infection with lung function FEV_1 (-0.1855, 95% CI (-0.5272, 0.1562); p = 0.2856) and FEV_1PP (-7.082, 95% CI (-18.8952, 4.7312); p = 0.2385). This relationship was independent to covariate adjustment (-0.1001, 95% CI (-0.05303, 0.3300); p = 0.6467 for FEV_1 and (-7.9181, 95% CI (-19.8085, 3.9723); p = 0.1906 for FEV_1PP in univariate analysis). See more detail in Table 3.3.1 to Table 3.3.2 and Table 3.3.5 to Table 3.3.6.

3.3.3 Lung Function Measured at Multiple-Time with Smoking

We detected outliers in GEE model for fitting lung function measured at multiple-time with smoking. From the box plot of residuals, there were only two data points detected as the outliers. Comparing the modeling results with and without outliers, we found that there was no significant difference between them. Thus, GEE model was robust to outliers, I didn't remove any outliers for the longitudinal methods. See Table 3.3.7 for more detail.

Current smoking was significantly associated with lung function FEV_1 (0.2734, 95% CI (0.1156, 0.4313); p = 0.0007) and FEV_1PP (4.5148, 95% CI (0.177, 8.8525); p = 0.0414) in univariate analysis. But in adjusted GEE models, current smoking was not statistically significant to lung function FEV_1 (0.024, 95% CI (-0.1006, 0.1486); p = 0.7055) and was potentially significant to FEV_1PP (3.5433, 95% CI (-0.6963, 7.7830); p = 0.1014). However, age (-0.017, 95% CI (-0.0243, -0.0098); $p < 0.0001$), height (0.0246, 95% CI (0.0149, 0.0344); $p < 0.0001$) and breathlessness (-0.1271, 95% CI (-0.2203, 0.0339); p = 0.0075) were highly correlated with lung function FEV_1 in the GEE model. The correlation of observations within a subject, denoted by ρ, was quite high (from 0.88 to 0.95) in univariate analysis. ρ is the correlation between second visit and first visit within a subject, ρ^2 is the correlation between third visit to first visit, and so on. The farther from first visit, the lower the correlation. Detail results showed in Table 3.3.8 to Table 3.3.11.

3.3.4 Lung Function Measured at Multiple-Time with C. *pneumoniae*

C. *pneumoniae* infection was associated with decreased lung function, but was not statistically significant (-0.0449, 95% CI (-0.3036, 0.2138); p = 0.7339 for FEV_1 and (-3.2143, 95% CI (-12.4137, 5.9851); p = 0.4935 for FEV_1PP in the GEE model). But age, gender, height and breathlessness were significantly related to lung function in univariate and multivariable analyses. See Table 3.3.12 and Table 3.3.13 for more detail.

3.4 Results of Subgroup Analysis

We examined whether higher concentrations of log MMP-9 and log CRP were associated with greater decline in lung function. We explored the relationship from three parts: (1) slope change in FEV_1 with inflammation; (2) baseline lung function with inflammation; (3) lung function measured at multiple-time with inflammation.

3.4.1 Slope Change in FEV_1 with Inflammation

- **Slope Change in FEV_1 with MMP-9**

In univariate models, log MMP-9 was significantly associated with decreased FEV_1 in three models (-0.1779, 95% CI (-0.2346, -0.1212); $p < 0.0001$ in the WLS). Also, log MMP-9 was independent to covariate adjustment. Higher concentration of log MMP-9 was significantly associated with decreased lung function (-0.1814, 95% CI (-0.2473, -0.1156); $p < 0.0001$ in the WLS). Similar results could be found in log ratio of MMP-9 to TIMP-1 (-0.1322, 95% CI (-0.1786, -0.0858); $p < 0.0001$ in the WLS). See Table 3.4.1 to Table 3.4.3 for more detail.

- **Slope Change in FEV_1 with CRP**

Log CRP (0.0798, 95% CI (0.026, 0.1337); $p = 0.0042$) significantly improve lung function in univariate analysis and was robust to covariate adjustment (0.0740, 95% CI (0.0151, 0.1328); $p = 0.0145$) in the WLS model. But log CRP was not significantly associated with lung function in MLR and ROBUST models.

- **Slope Change in FEV_1 with Viral Infection**

Among 81 subjects, adenovirus was detected in 7 (8.6%) and herpesviridae in 46 (56.8%). The composite of adenovirus and herpesviridae was associated with FEV_1 decline, but was not statistically significant (-0.0292, 95% CI (-0.0831, 0.0247); $p = 0.2828$ in the MLR). This relationship was independent to covariate adjustment

(0.0271, 95% CI (-0.026, 0.0801), p = 0.3126 in unadjusted MLR). See Table 3.4.5 for more detail.

3.4.2 Baseline Lung Function with Inflammation

- **Baseline Lung Function with MMP-9**

Higher concentration of log MMP-9 was not statistical significantly associated with baseline lung function FEV_1 (-0.0775, 95% CI (-0.3692, 0.2142); p = 0.5983) and FEV_1PP (-3.6566, 95% CI (-13.3225, 6.0093); p = 0.4536). Similar results could be found in log ratio of MMP-9 to TIMP-1. See Table 3.4.6 to Table 3.4.9 for more detail.

- **Baseline Lung Function with CRP**

Higher levels of log CRP was potential significantly associated with decreased baseline lung function FEV_1 (-0.1659, 95% CI (-0.3685, 0.0368); p = 0.1071) and FEV_1PP (-6.2396, 95% CI (-12.5096, 0.0305); p = 0.0511). See Table 3.4.10 and Table 3.4.11 for more detail.

- **Baseline Lung Function with Viral Infection**

The composite of adenovirus and herpesvirdae was associated with decreased baseline FEV_1 (-0.0177, 95% CI (-0.3063, 0.2709); p = 0.9031), but was not statistically significant. However, the composite of adenovirus and herpesvirdae significantly decreased baseline FEV_1PP (-4.6955, 95% CI (-9.6138, 0.2228); p = 0.0610). See Table 3.4.12 and Table 3.4.13 for more detail.

3.4.3 Lung Function Measured at Multiple-Time with Inflammation

Among 81 subjects, we examined the relationship between lung function measured at multiple-time and inflammation by employing GEE, MIXED and RRE models.

- **Lung Function Measured at Multiple-Time with MMP-9**

Log MMP-9 was associated with lung function FEV_1 measured at multiple-time, but was not statistically significant in three models (-0.1802, 95% CI (-0.4189, 0.0586); p = 0.1392 in the GEE model). However, log MMP-9 was significantly correlated with lung function FEV_1PP (-7.1560, 95% CI (-14.8068, 0.4948); p = 0.0668) in the GEE model, but was not significant in other two models. Similar results could be found in log ratio of MMP-9 to TIMP-1 (-0.1221, 95% CI (-0.2978, 0.0535); p = 0.1730 for FEV_1 and -4.7287, 95% CI (-10.3992, 0.9418); p = 0.1022 for FEV_1PP in the GEE model). See Table 3.4.17 to Table 3.4.21 for more detail.

- **Lung Function Measured at Multiple-Time with CRP**

Log CRP was highly associated with decreased lung function FEV_1 (-0.2673, 95% CI (-0.5002, -0.0344); p = 0.0245) and FEV_1PP (-7.5017, 95% CI (-12.9637, -2.0397); p = 0.0071) in univariate analysis. After the covariate adjustment, log CRP was still significantly associated with lung function FEV_1 (-0.2242, 95% CI (-0.3922, -0.0562); p = 0.0089 in the GEE model) and FEV_1PP (-7.4112, 95% CI (-12.8041, -2.0182); p = 0.0071 in the GEE model) in three models. The higher CRP level, the lower the lung function. Therefore, log CRP was the best biomarker for lung function measured at multiple-time. The working correlation of ρ is 0.9108 in the GEE model. That means the correlation between first visit and second visit was very high. See Table 3.4.20 and Table 3.4.21 for more detail.

- **Lung Function Measured at Multiple-Time with Viral Infection**

The composite of adenovirus and herpesviridae was significantly associated with decreased lung function FEV_1 (-0.2395, 95% CI (-0.4362, -0.0365); p = 0.0205) and FEV_1PP (-5.102, 95% CI (-9.7881, -0.416); p = 0.0328) in unadjusted GEE models.

After covariate adjustment, it was still significantly correlated with decreased lung function FEV_1 (-0.1742, 95% CI (-0.3026, -0.0459); p = 0.0078) and FEV_1PP (-5.4163, 95% CI (-9.8763, -0.9563); p = 0.0173) in three models. See Table 3.4.7 to Table 3.4.8

and Table 3.4.22 to Table 3.4.23 for more detail.

3.5 Results of Sensitivity Analysis

We employed group mean estimation, LOCF, hot-deck and multiple imputation (MI) methods to handle the missing data and generated four complete data sets in which each patient had 9 visits. For the multiple imputation method, we let the imputation number equal 10. According to the formula 2.10, we obtained an efficiency of 94.63% ($[1+(1021/1800)/10]^{-1} \times 100\%$). We examined the relationship between lung function measured at multiple-time and smoking, C. *pneumoniae* infection in univariate and multivariable GEE models.

3.5.1 Lung Function Measured at Multiple-Time with Smoking

In univariate analysis, current smoking was highly correlated with lung function FEV_1 measured at multiple-time and was robust to different imputation methods (0.1563, 95% CI (0.0688, 0.2439); p = 0.0006 in the MI data). Current smoking was also significantly associated with lung function FEV_1PP measured at multiple-time except in the MI data (2.7761, 95% CI (0.2098, 5.3425); p = 0.034 in the hot-deck data). But after covariate adjustment, current smoking was not significantly associated with lung function FEV_1 measured at repeated-time (0.051, 95% CI (-0.0395, 0.1414), p = 0.2640 in the MI data) in four data sets, but it was significantly related to lung function FEV_1PP except in the MI data (2.3176, 95% CI (-0.0577, 8.3460); p = 0.0731 in the hot-deck data). In addition, age, gender and height were significantly associated with lung function FEV_1 measured at-multiple-time and robust to all the imputed data sets in adjusted and unadjusted models. The working correlation of ρ in LOCF, mean estimation, hot-deck and multiple imputation data sets was 0.9544, 0.2670, 0.1948 and 0.1485, respectively. The lowest correlation of observations existed in the MI data, the highest correlation of observations existed in LOCF data. See Table 3.5.1 to Table 3.5.4 for more detail.

For other predictors in univariate analysis, breathlessness was significantly associated with decreased lung function FEV_1 (-0.1414, 95% CI (-0.2423, -0.0405); p = 0.0103 in the MI data) and FEV_1PP in different imputation methods. Pneumonia shot was significantly related to decreased lung function FEV_1 and FEV_1PP. Flu shot and number of flu shots in past 5 years were strongly correlated with decreased lung function FEV_1 in the hot-deck and the MI data sets. However, current cough, productive cough, sputum, and sputum amount were not significantly associated with lung function FEV_1, but sputum and productive cough were significantly associated with FEV_1PP in some imputed data sets, but were not statistically significant in other data sets. See Table 3.5.1 and Table 3.5.2 for more detail.

3.5.2. Lung Function Measured at Multiple-Time with C. *pneumoniae*

In univariate analysis, C. *pneumoniae* infection was not significantly associated with decreased lung function FEV_1 (-0.1116, 95% CI (-0.2714, 0.0482), p = 0.2492 in the MI data) and FEV_1PP (-1.9461, 95% CI (-6.2069, 2.3146), p = 0.3688) in all the data sets. However, in multivariable analysis, C. *pneumoniae* infection was significantly associated with decreased lung function FEV_1 (-0.1041, 95% CI (-0.2196, 0.0114); p = 0.0772) in the group mean estimation data set, but was not statistically significant in other data sets. In addition, age, gender and height were significantly associated with lung function FEV_1. However, current smoking was significant associated with lung function FEV_1PP except the MI data. But breathlessness was highly correlated with lung function FEV_1 and FEV_1PP. See Table 3.5.5 and Table 3.5.6 for more detail.

Chapter 4 Discussion

This project was focused on determining the optimal modeling for lung disease and its progression. We employed multiple linear regression, weighted least squared and robust linear regression to examine the relationship between slope change in FEV_1 and smoking, C. *pneumoniae* infection and inflammation. We also applied GEE, Mixed-effects and robust random-effects models to determine the association between lung function measured at multiple-time and smoking, C. *pneumoniae* infection and inflammation. For handling missing data, we used group mean substitution, LOCF, hot-deck and multiple imputation methods to generate different complete data sets and built GEE models using these imputed data. In Chapter 3, we addressed all the modeling results. But which models were better to fit the data and which imputation methods were suitable to handle the missing data? We could compare the estimate, 95% confidence interval and p-value of each variable in different models in order to find the differences and similarities. We could examine the residuals for checking the suitability of statistical models after the models had been fitted. Graphical methods are often used in checking residuals. Thus, we drew Q-Q plots of residuals for each multivariable model. We could also check goodness-of-fit test for each adjusted model. In this Chapter, we would discuss the comparisons of different modeling and imputation methods by using above approaches.

4.1 Modeling Comparison

4.1.1 Modeling Comparison in Primary Analysis

We compared the modeling results of weighted least squares, multiple linear regression and robust linear regression in determining the association of slope change in FEV_1 with smoking and C. *pneumoniae* infection. In these models, we assumed that the random error terms were identical and independently distributed normal N (0, σ^2).

There was a linear regression between Y (slope change in FEV_1, rather than FEV_1 itself) and X (age, gender, height, current smoking, breathlessness, infection or inflammation).

- **Similarities and Differences**

For slope change in FEV_1 with smoking in Table 3.2.3, the estimated coefficients of age and breathlessness were very close to each other in all three models. For other predictors, coefficients in the MLR and ROBUST models were very close, but large differences existed in the WLS model. For example, the estimated coefficients, 95% confidence intervals and p-values of current smoking in the MLR, ROBUST and WLS models were -0.0457, 95% CI (-0.1045, 0.013); p = 0.126, -0.0422, 95% CI (-0.1029, 0.0185); p = 0.0203, and -0.0869, 95% CI (-0.1601, -0.0138); p = 0.1733, respectively. 95% confidence interval in the WLS was smaller than others and its p-value was significant. Similar results could be found in gender and height.

However, for slope change in FEV_1 with CRP in Table 3.4.4, log CRP was significantly associated with FEV_1 decline in the WLS model, but was not statistically significant in the other two models.

Therefore, WLS model was better than others based on significance for clinical predictors.

- **Residuals Analysis**

Q-Q plots of WLS, MLR and ROBUST were very similar, some points were a slightly away from the Q-Q line, but most of residual points lay around the Q-Q line. Thus, we believed that all three models were suitable to fit the data from residuals analysis. See Figure 3.5 for more detail.

- **Goodness-of-fit Test**

We checked R-square for each model. The R-square value is an indicator of how well the model fits the data. The larger the value of R-square, the better the model fit. For the relationship of slope change in FEV_1 with MMP-9, the values of R-square were

0.445, 0.1096 and 0.0877 in the WLS, MLR and ROBUST models, respectively. Thus, weighted least squares model was the best model to fit the data according to the value of R-square. See Table 4.1.1 for more detail.

Similar results could be found in the models of slope change in FEV_1 with smoking, C. *pneumoniae* infection and CRP inflammation. See Figure 3.8, Table 3.4.2 to Table 3.4.4 for more detail.

Some research showed that the estimator was the best linear unbiased estimator (BLUE) if we used the reciprocal of variance as the weights in WLS model. In this study, we used the reciprocal of square of standard errors as the weights to fit WLS models. Therefore, we considered WLS model as the best to fit the data in determining the association of slope change in FEV_1 and smoking, C. *pneumoniae* infection and inflammation. However, WLS model was sensitive to outliers, we need to check and remove outliers when we use WLS models to fit the data.

4.1.2 Modeling Comparison in Secondary Analysis

We examined the association of lung function measured at multiple-time with smoking, C. *pneumoniae* infection and inflammation by fitting GEE, MIXED and RRE models.

- **Similarities and Differences**

For lung function FEV_1 measured at multiple-time with C. *pneumoniae* infection in Table 3.3.11, the modeling results including estimated coefficients, 95% confidence intervals and p-values were very close. There was a little difference among three models. For example, the coefficients of C. *pneumoniae* infection in GEE, MIXED and RRE models were -0.0449, 95% CI (-0.3036, 0.2138), -0.0338, 95% CI (-0.2847, 0.2170) and -0.0346, 95% CI (-0.2860, 0.2168), respectively. However, there was no significant difference for height in three models. The estimators of height for GEE, MIXED and RRE models were 0.0247, 95% CI (0.0149, 0.0345); 0.0250, 95% CI (0.0160, 0.0340) and 0.0251, 95% CI (0.0161, 0.0341) with same p-value < 0.0001,

respectively. Moreover, the estimates, 95% confidence intervals and p-values of age, gender and height were almost same in MIXED and RRE models. Also, a little difference existed in breathlessness and current smoking between these two models. GEE model had different results, but the difference was not significant.

However, for lung function measured at multiple-time with log MMP-9 in Table 3.4.16, log MMP-9 was significantly associated with lung function FEV_1 in the GEE model, but was not significant in other two models. Similar results in log ratio of MMP-9 to TIMP-1. Therefore, GEE model is better than other two models based on significance of clinical predictors.

- **Residuals Analysis**

We checked Q-Q plots of residuals in each GEE and Mixed model, but neglected RRE model whose results were very similar to the MIXED model. Most of residual points lay in the middle of regression line. Some points were a little away from the regression line at the bottom and the top. Thus, we could say that all three models were suitable to fit this data which was very robust to different modeling approaches. Similar results could be found in other models in secondary and subgroup analysis. See Table 3.3.11, Table 3.4.10 to Table 3.4.12 and Figure 3.6 for more detail.

- **Goodness-of-fit Test**

We checked the values of log likelihood and AIC (Akaike Information Criterion) for each GEE and MIXED model. The smaller the AIC, the better the model fit. The values of AIC in the MIXED models were much smaller than those for the GEE models. Thus, MIXED model was better than GEE model to fit the data from goodness-of-fit test. See Table 4.1.2 for more detail.

However, in the GEE model, we assumed the observations were correlated within same subject and were independent between the subjects. It only uses marginal distribution and working correlation matrix to estimate the coefficients and variance-covariance of parameters. It is more realistic to analyze the longitudinal data. In addition,

GEE model is very easy to implement in SAS 9.1 and STATA 9.1, Therefore, I recommended that GEE model was the best choice to fit the data.

4.2 Comparisons of Imputation Methods

We fitted GEE models based on different data sets imputed by group mean estimation, LOCF, hot-deck and multiple imputation (MI) methods. We assessed the similarities and differences among various imputation methods by comparing the estimated coefficients, 95% confidence intervals and p-values. We also compared Q-Q plots of residuals and checked goodness-of-fit for each fitted model.

- **Similarities and Differences**

For lung function FEV_1 measured at multiple-time with smoking in Table 3.5.3, the modeling results for each predictor were consistent with different imputation data. Age was significantly associated with decreased lung function (-0.0175, 95% CI (-0.0244,-0.0106); $p < 0.0001$), (-0.0074, 95% CI (-0.0110,-0.0038); $p < 0.0001$), (-0.0167, 95% CI (-0.0209, -0.0124); $p < 0.0001$) and (-0.0113, 95% CI (-0.0157,-0.0070); $p < 0.0001$) in the LOCF, group mean estimation, hot-deck and MI data sets, respectively. Consistent results could be found in the predictors of gender, height, current smoking and breathlessness.

- **Residuals Analysis**

From Q-Q plots of GEE models using four different imputed data, multiple imputation method was the best. Hot-deck method was better than others. Multiple imputation using Markov Chain Monte Carlo (MCMC) method randomly imputes the missing data. Thus, it can reduce the variability of results and lead to unbiased estimators. Moreover, hot-deck method had similar modeling results with multiple imputation method. LOCF method was better than group mean estimation by Q-Q plots of residuals. But the values of ρ in LOCF data were beyond 0.9, because many imputed

data points were the same, which increased the correlation within subjects. See Figure 3.7 for more detail.

- **Goodness-of-fit Test**

We checked values of Deviance/DF in GEE model based on various imputed data and original data. Mean estimation data had the smallest deviance, MI data had the largest deviance in the GEE model of FEV_1 measured at multiple-time with smoking and C. *pneumoniae* infection. The value of original data was between the smallest and the largest. Even though mean estimation data had the lowest deviance, it had the worst residual plots and a disadvantage of underestimation. Therefore, multiple imputation method is the best choice to impute a large number of missing data. It can improve the efficiency by increasing the imputation number [34]. Moreover, it is easy to implement in SAS 9.1 with PROC MI and PROC MIANALYZE. Hot-deck is a good choice when few missing data exist in the data set. It is also easy to implement in STATA 9.1 with HOTDECK function. See Table 4.2.1 for more detail.

4.3 Future Study

There were several limitations, which should be addressed in future studies:

- GEE model requires sample space (any two successive visits) to be equal for binary outcome [36], otherwise it generates inaccurate results. However, the data set in this study came from 200 patients with follow-up over two years. Some patients visited once over 3 months, some patients visited once over or less than 1 month. In order to make sample space equal for the continuous outcome of lung function, we defined a variable VISIT, a time interval of any two successive visits, as 2 months ± 1 month. We removed some data points if any two successive visits were less than 2 months. Thus, sample space was close to equal. This might be the reason that some modeling results were very interesting. For example, smoking would be better to lung function. In the future, we can try SAS macro GLIMMIX and Proc NLMIXED, which allow the observations to be equally or unequally spaced [35].

• For checking the model validity, we can use an effective method which is to obtain one more follow-up data point in one year. Comparing the fitted values with these observed values, we can determine which model fits the data better.

•. From all the modeling results, we only found that C. *pneumonaie* infection (0.0584, 95% CI (-0.1283, 0.0114); p = 0.1002) was potentially associated with FEV_1 decline in the WLS model. We couldn't find significant relationship between lung

function and C. *pneumoniae* infection in any other models. There were only 12 of 200 subjects detected C. *pneumoniae* infection in their sample of blood in this study. Therefore, the future study will enlarge sample size, optimize the modeling strategy and improve the accuracy for estimating the magnitude of the association.

Chapter 5 Conclusion

The overall goal in this study was to determine the optimal modeling strategy by comparing results from different methods of analysis and methods of handling missing data. We examined the lung disease and its progression from three aspects:

• Let slope change in FEV_1 as a function of explanatory variables including patients' demographics, smoking status, respiratory symptom, vaccination shots, infection and inflammation by employing weighted least squares, multiple linear regression and robust linear regression models in determining the association of lung function decline with smoking, C. pneumoniae infection and MMP-9, CRP inflammation.

• Let baseline lung function as a function of the explanatory variables by applying multiple linear regression models in determining the relationship of baseline lung function with smoking, infection and inflammation.

• Let lung function measured at multiple-time as a function of the explanatory variables by employing GEE, mixed and robust random-effects models and comparing the modeling results in determining the relationship of lung function measured at multiple-time with smoking, infection and inflammation

For slope change in FEV_1, I drew the conclusions as follows:

• Current smoking strongly accelerated lung function decline at a rate of 86.9 ml per year. Current smoking was robust to adjustment for age, gender, height and breathlessness.

• C. *pneumoniae* infection decreased the lung function, but was not statistically significant.

• Log MMP-9 and log ratio of MMP-9 to TIMP-1 were significantly correlated with the slope change in FEV_1. The higher concentrations of log MMP-9 and log ratio of MMP-9 to TIMP-1, the lower lung function. Log MMP-9 was robust to covariate adjustment. Thus, it was a good biomarker for COPD progression.

• Log CRP was significantly correlated with the slope change in FEV_1, but higher levels of log CRP would have better lung function. Log CRP was robust to covariate adjustment. Therefore log CRP was another good biomarker for COPD progression.

• The composite of adenovirus or herpesviriade viral infection was associated with decreased lung function FEV_1, but was not statistically significant.

• Weighted least squares model is recommended to fit the data based on R^2 and significance for clinical predictors, such as smoking, but it was sensitive to outliers.

For baseline lung function, I drew the following conclusions:

• Current smoking was not significantly correlated with baseline lung function FEV_1, but was strongly related to baseline lung function FEV_1PP.

• C. *pneumoniae* infection would decrease baseline lung function, but was not statistically significant.

• Log MMP-9 and log ratio of MMP-9 to TIMP-1 would decrease baseline lung function, but was not statistically significant.

• Log CRP decreased baseline lung function FEV_1PP at significant level 0.05, but was not significantly decrease baseline lung function FEV_1.

•. The composite of adenovirus and herpesviridae infection significantly decrease baseline lung function FEV_1PP, but was not significantly associated with baseline FEV_1.

For lung function measured at multiple-time, I drew the following conclusions:

• Current smoking was not significantly correlated with lung function FEV_1 and FEV_1PP measured at multiple-time in adjusted models.

• C. *pneumoniae* infection was not significantly associated with lung function measured at multiple-time in both unadjusted and adjusted models.

• Log MMP-9 and log ratio of MMP-9 to TIMP-1 were associated with decreased lung function measured at multiple-time, but were not statistically significant.

• Log CRP was significantly associated with decreased lung function FEV_1 and FEV_1PP measured at multiple-time and was robust to covariate adjustment. Thus, log CRP was the best biomarker for lung function measured at multiple-time.

• The composite of adenovirus and herpesviriade viral infection was significantly associated with decreased lung function FEV_1 and FEV_1PP.

• There was no significant difference among GEE, MIXED and RRE models.

For sensitivity analysis, we drew the conclusion as follows:

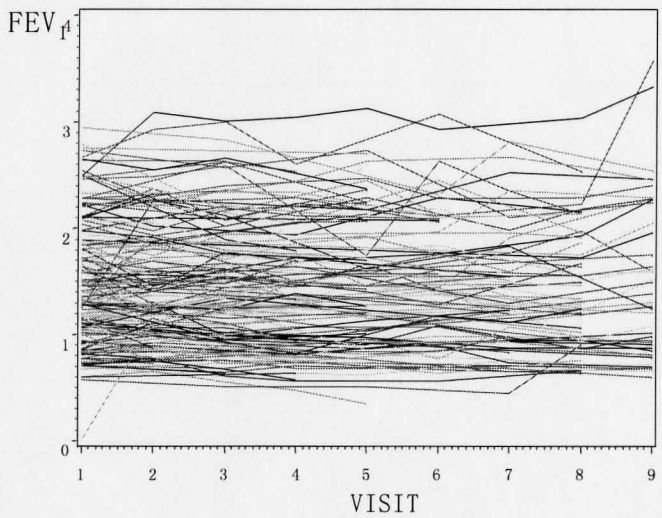
- GEE modeling results for different imputed data were consistent.
- Multiple imputation method is the best method to impute missing data.
- Comparing with original data, GEE modeling results based on LOCF data were almost same as those of original data. But MI data improved p-values, even though there were small differences in estimators.

Summarized conclusions were showed in Table 5.1 to Table 5.3.

Chapter 6 Appendix

6.1 Figures

Figure 3.1 Plots of FEV_1 Measured at Multiple-Time



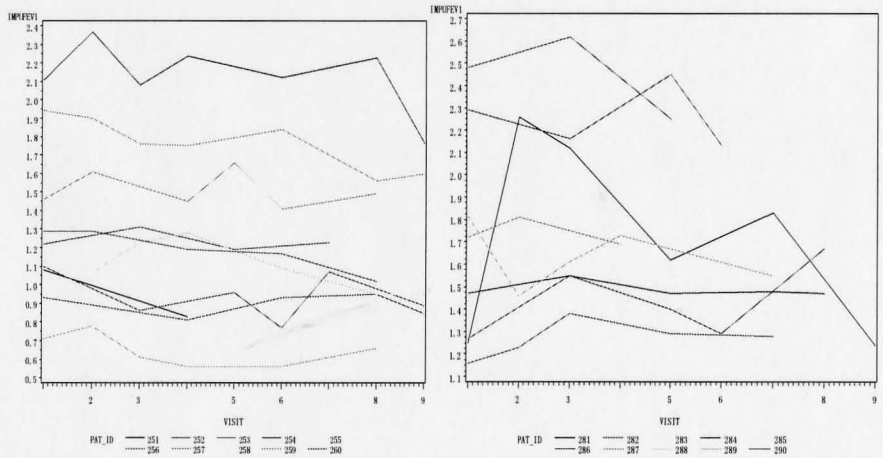
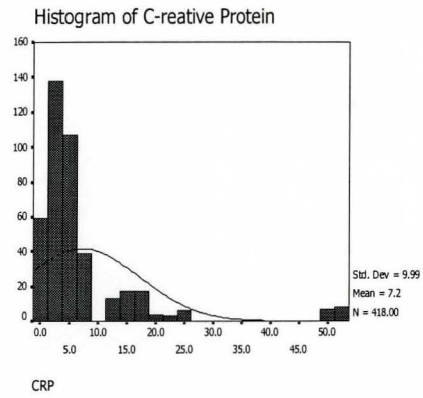
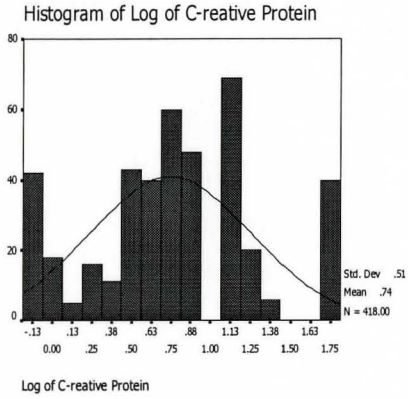
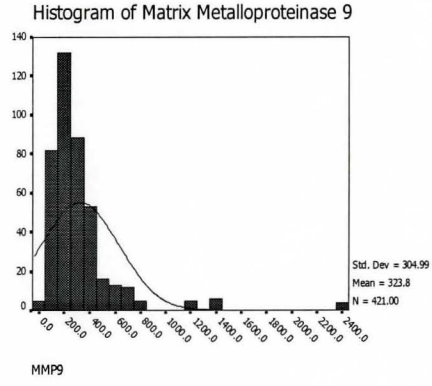
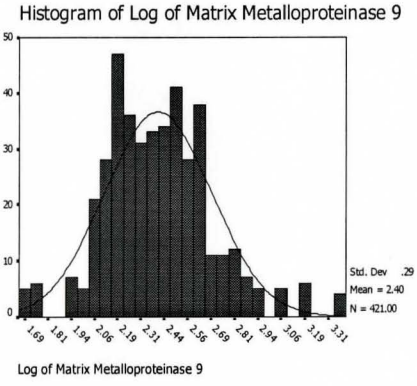


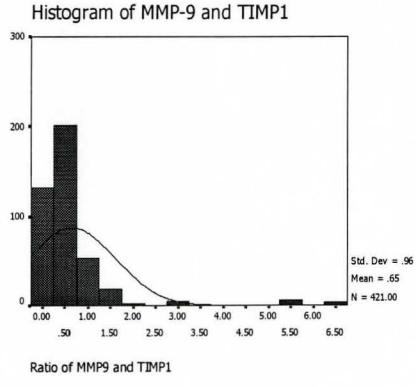
Figure 3.2 Histograms of Inflammatory Variables











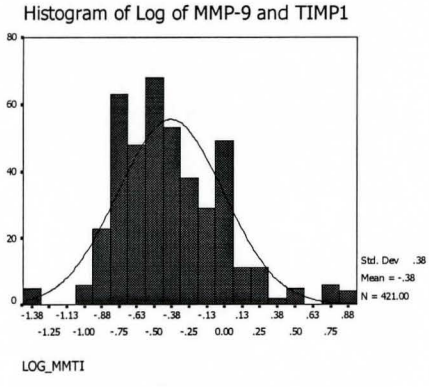
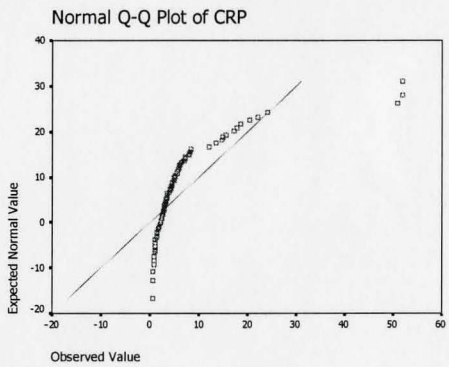
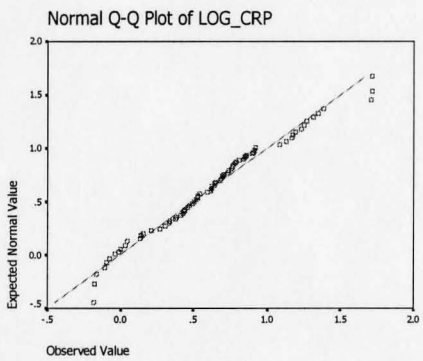
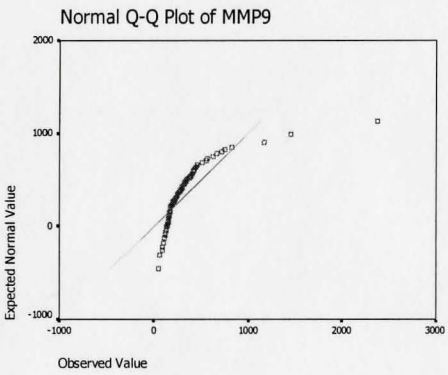
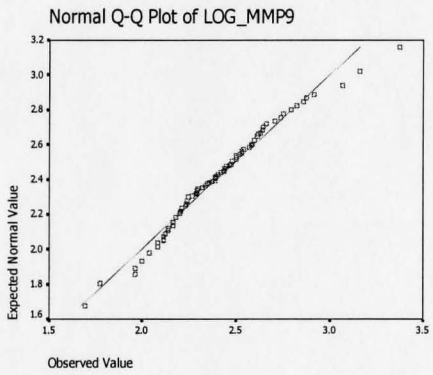


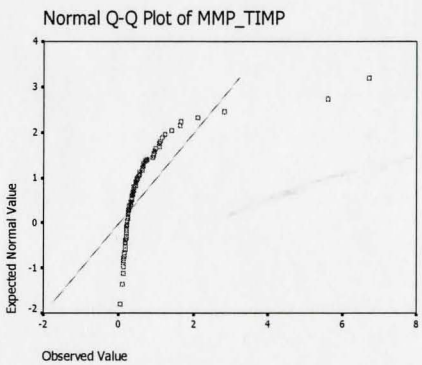
Figure 3.3 Q-Q Plots of Inflammatory Variables











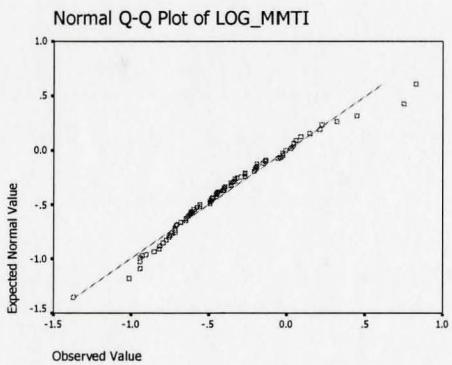
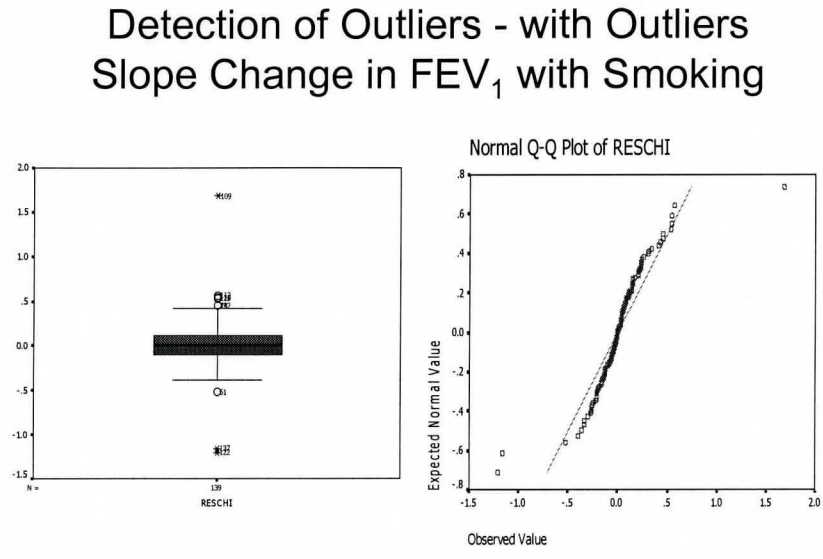


Figure 3.4 Detection of Outliers in Slope Change in FEV_1



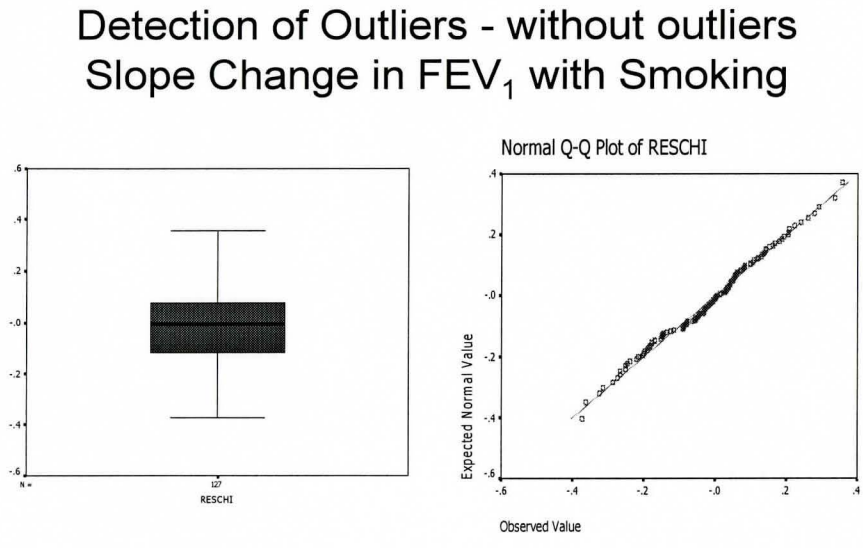
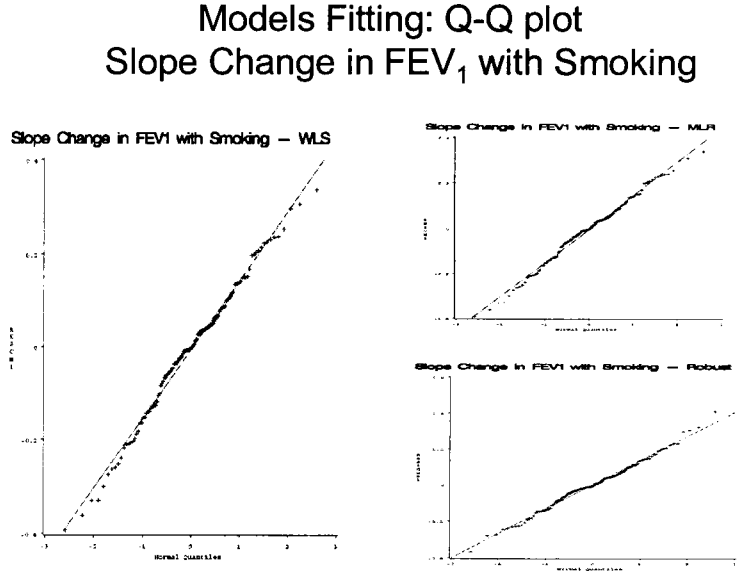
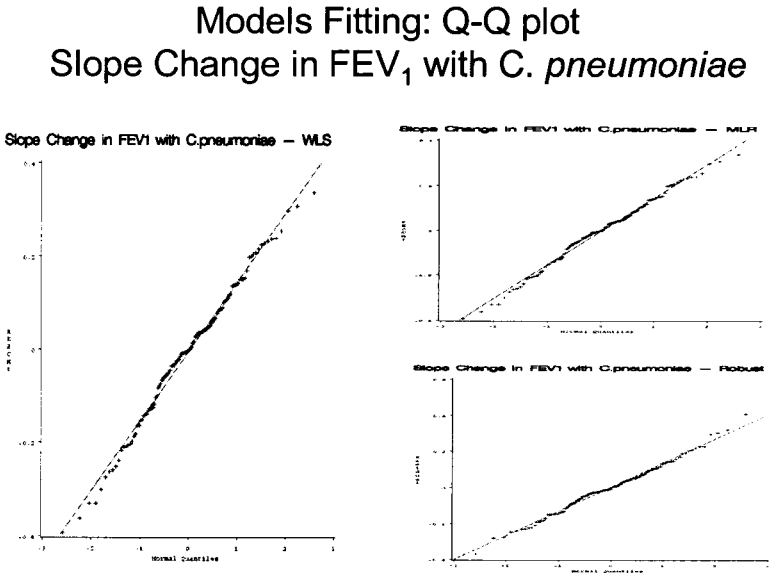
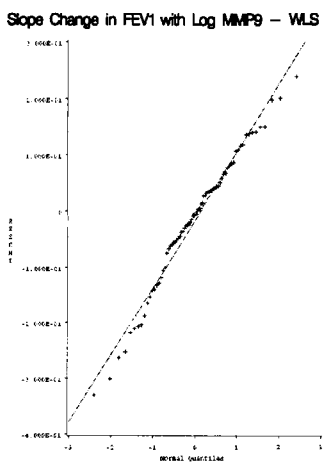


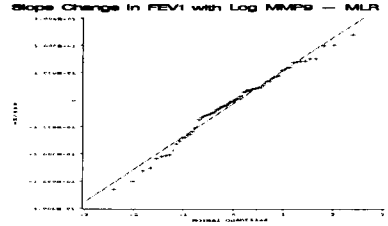
Figure 3.5 Models Fitting in Slope Change in FEV_1

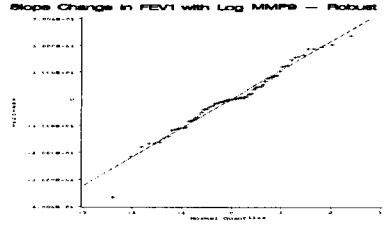




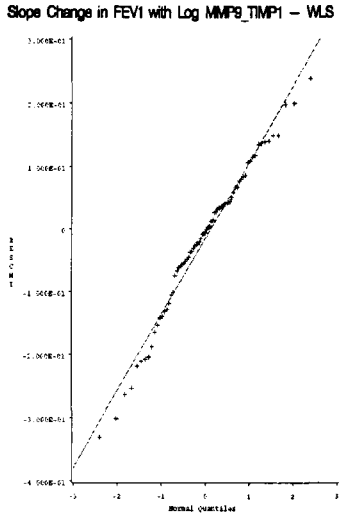
Models Fitting: Q-Q plot
Slope Change in FEV_1 with Log MMP-9

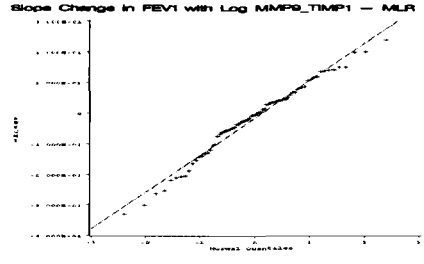


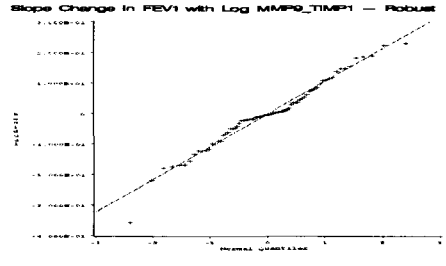




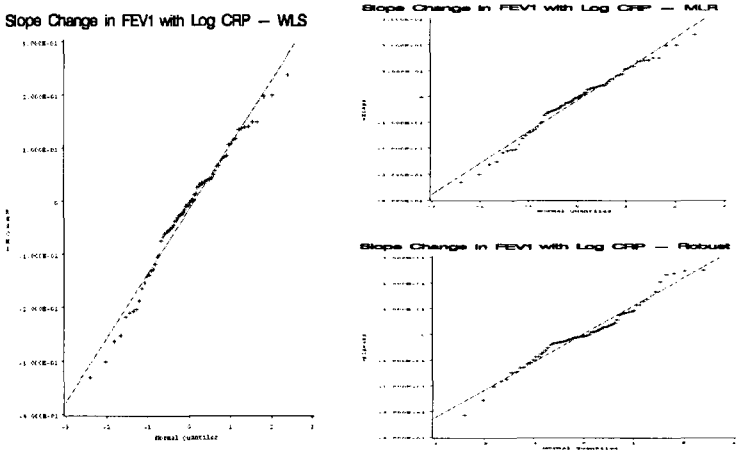
Models Fitting: Q-Q plot
Slope Change in FEV_1 with Log MMP9_TIMP1







Models Fitting: Q-Q plot
Slope Change in FEV_1 with Log CRP



Models Fitting: Q-Q plot
Slope Change in FEV_1 with Viral Infection

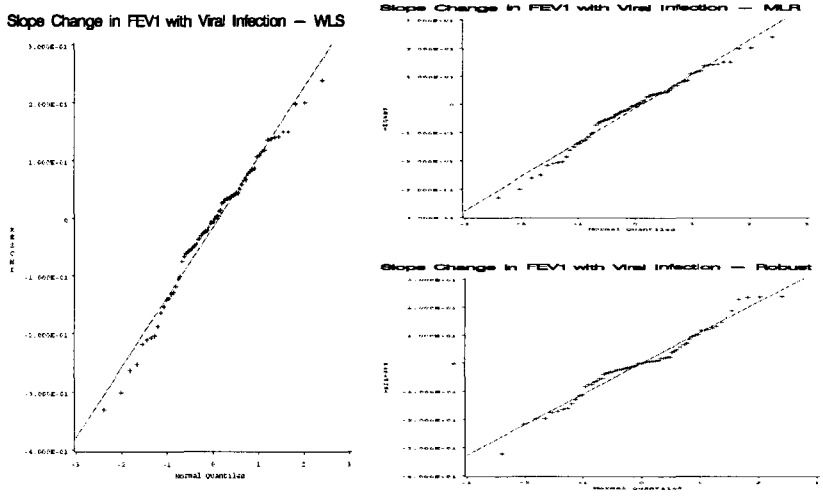
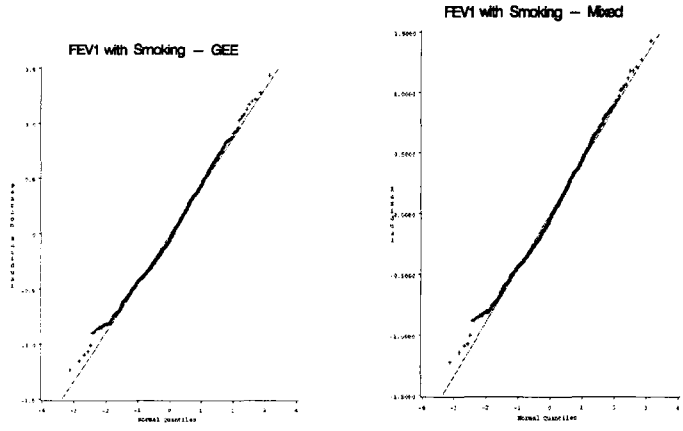
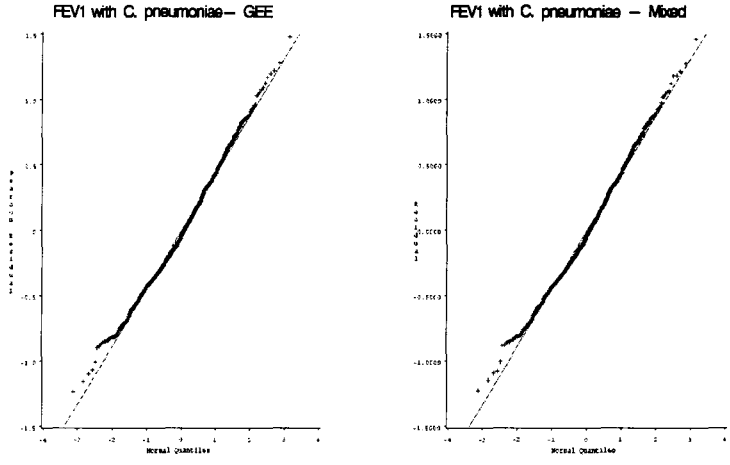


Figure 3.6 Models Fitting in Lung Function Measured Over Time

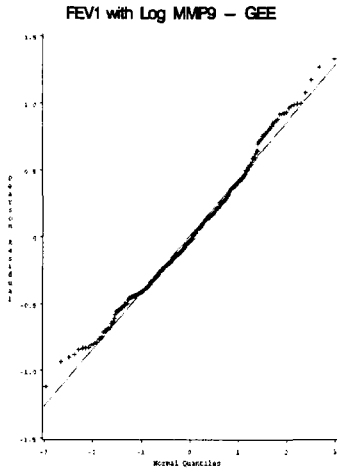
Models Fitting : Q-Q Plot
FEV_1 with Smoking

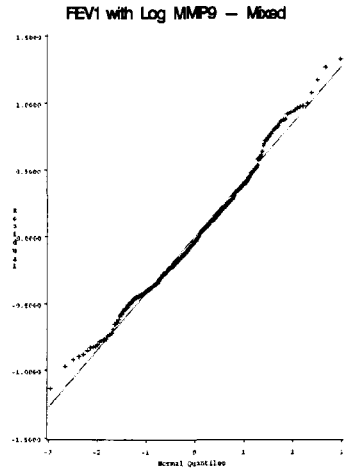


Models Fitting: Q-Q Plot
FEV_1 with C. *pneumoniae*

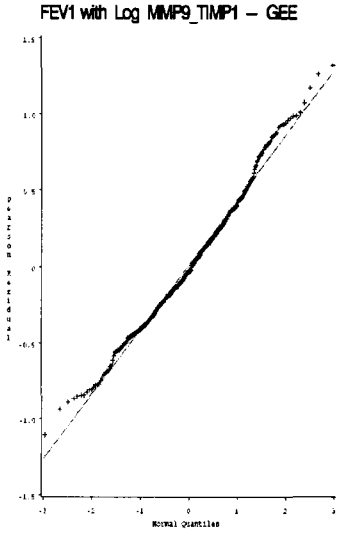


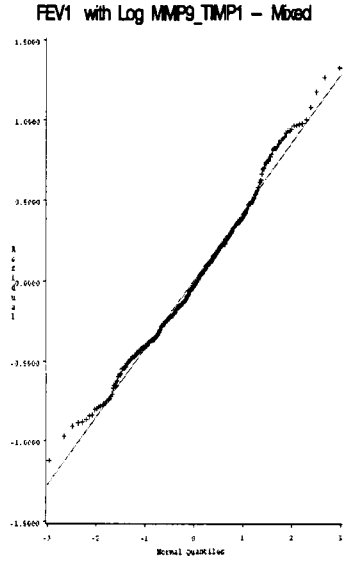
Modeling Comparison: Q-Q Plot
FEV_1 with Log MMP-9



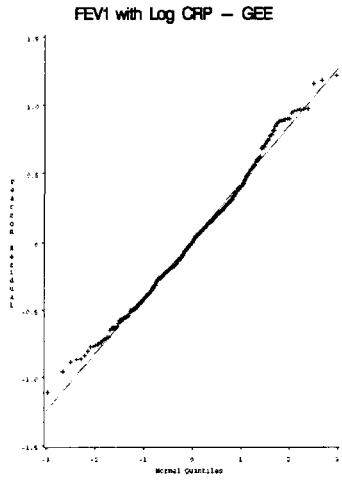


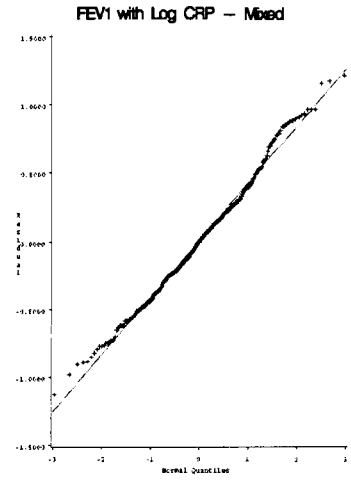
Models Fitting: Q-Q Plot
FEV_1 with Log MMP-9_TIMP-1



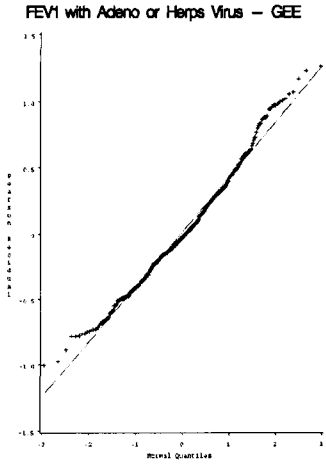


Models Fitting: Q-Q Plot
FEV_1 with Log CRP





Models Fitting : Q-Q Plot
FEV_1 with Viral Infection



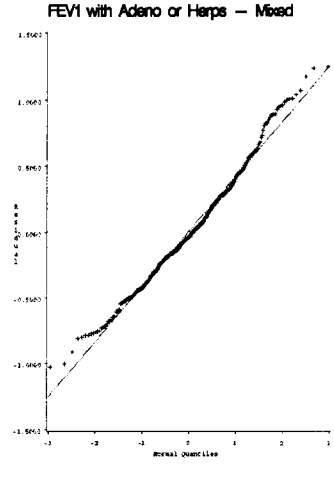
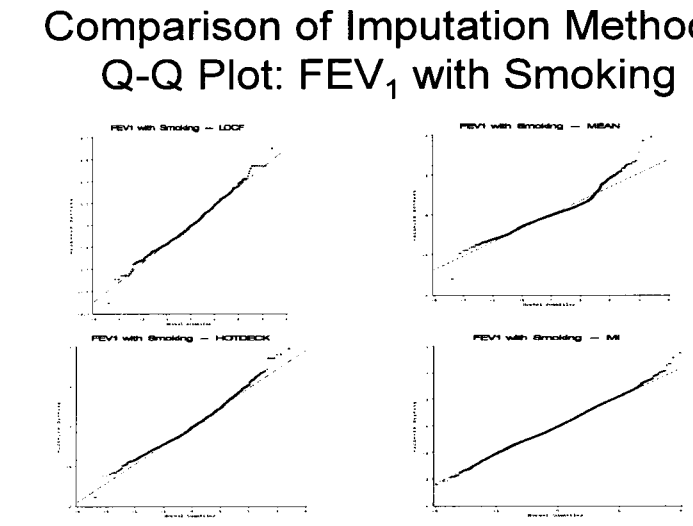
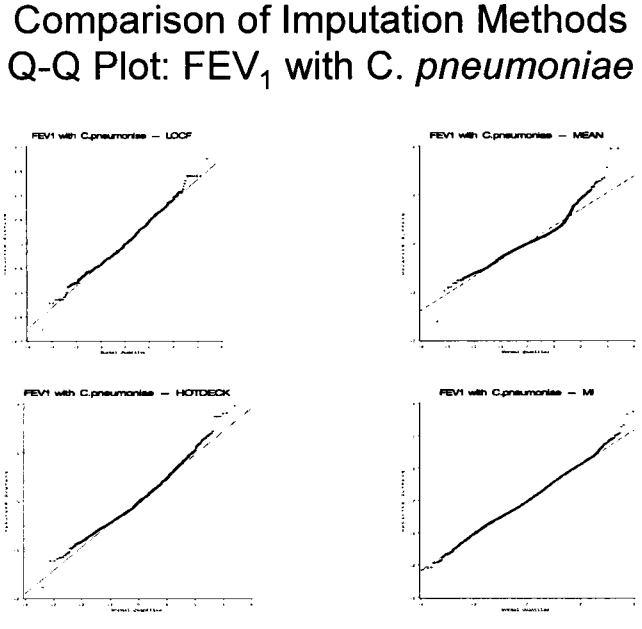


Figure 3.7 Comparisons of Imputation Methods





6.2 Tables

Table 1.1 Abbreviations Used in the Project

Abbreviation	Term
COPD	Chronic Obstructive Pulmonary Disease
CAL	Chronic Airflow Limitation
MMP-9	Matrix Metalloproteinase 9
TIMP-1	Tissue Inhibitor of Metalloproteinase 1
CRP	C - reactive Protein
C. *pneumoniae*	Chlamydia *pneumoniae*
FEV1	Force Expiratory Volume at first second
FEV1PP	Percent FEV1 Predicted by age, gender and height
FVC	Forced Vital Capacity
VIF	Variation Inflation Factor
WLS	Weighted Least Squares
MLR	Multiple Linear Regression
ROBUST	Robust linear regression
MIXED	Mixed-effects model
GEE	Generalized Estimating Equations
RRE	Robust Random-Effects model
GLM	Generalized Linear Model
MI	Multiple Imputation
AIC	Akaike Information Criterion
ICC	Intra-class correlation coefficient
ρ	Working correlation

Table 1.2 Sample Sizes Calculation

Index	ICC (ρ)	VIF	Sample Size
1	0.01	1.08	152
2	0.02	1.16	163
3	0.03	1.24	175
4	0.04	1.32	186
5	**0.05**	**1.40**	**197**
6	0.06	1.48	209
7	0.07	1.56	220
8	0.08	1.64	231
9	0.09	1.72	242
10	0.10	1.80	254
11	0.11	1.88	265
12	0.12	1.96	277
13	0.13	2.04	288
14	0.14	2.12	299
15	0.15	2.20	310

Table 1.3 Sample Sizes in Different Data Sets

Data Sets	Data Analysis	Sample Size
Slope Change in FEV1	Primary	139 subjects
		127 subjects without outliers
Slope Change in FEV1	Subgroup	77 subjects
		75 subjects without outliers
Baseline Lung Function	Secondary	200
Lung Function Measured at Multiple-time	Primary	200
	Subgroup	81
	Sensitivity	1800

Table 2.1 Software for Data Analysis

Software	Functions	Data Analysis
SPSS 11.5	Descriptive Box plot Q-Q plot Shapiro-Wilk test	Frequency, Descriptive of Data Detection of Outliers Residuals Analysis Normality Tests
SAS 9.1	PROC REG	Weighed Least Squares Multiple Linear Regression
	PROC GENMOD	GEE Model
	PROC MIXED	Random Effects Model
	PROC STANDARD	Mean Imputation
	PROC MI	Multiple Imputation
	PROC MIANALYZE	Multiple Imputation Modeling Analysis
	PROC UNIVARIATE	Residual Analysis, Detection of Outliers
	PROC ROBUSTREG	Robust Linear Regression
S-PLUS 6.2	PROGRAMMING	Group Data By Visit for Mean Estimation
		LOCF Imputation
STATA 9.1	HOTDECK	Hot- Deck Imputation
	ROBUST REGRESSION FOR LONGITUDINAL DATA	Robust Random-effects Model

Table 3.1.1 Descriptive of Variables and Outcomes

Patients' Demographics	
Age (yr)	n = 200, Mean 64.3 (SD 9.4)
Gender (male/female)	n = 200, 110 (55.0)
Height (cm)	n = 200, Mean 166.6 (SD 9.4)
Smoking Status	
Current Smoking (current/former smoker)	n = 200, 79 (39.5)
Pack Years of Smoking	n = 200, Mean 46.6 (SD 19.4)
Quit Years of Smoking (yr)	n = 196, Mean 6.0 (SD 8.5)
Vaccination Shots	
Flu Shot in Last Year (yea/no)	n = 171, 148 (86.5)
Flu Shots Number in Last 5 Years	n = 171, Mean 3.7 (SD 1.8)
Pneumonia Shot (yes/no)	n = 163, 94 (57.7)
Respiratory Symptom	
Current Cough (yes/no)	n = 200, 163 (81.5)
Productive Cough (yes/no)	n = 199, 151 (75.9)
Sputum (yes/no)	n = 200, 160 (80.0)
Sputum Amount	n = 99, Mean 1.4 (SD 1.4)
Breathlessness	n = 84, Mean 3.0 (SD 1.0), Min=1, Max=5
Secondary Variables (Infection)	
C. *pneumoniae* at Baseline (yes/no)	n = 200, 7 (3.5)
C. *pneumoniae* at Base / follows (yes/no)	n = 200, 12 (6.0)
Subgroup Variables (Inflammation)	
C-reactive Protein (CRP)	n = 81, Mean 7.2951 (SD 10.1759)
Matrix Metalloproteinase 9 (MMP 9)	n = 82, Mean 326.9 (SD 323.02)
Tissue Inhibitor of Metalloproteinase 1	n = 82, Mean 652.5 (SD 189.7)
Ratio of MMP-9 toTimp-1	n = 82, Mean 0.656 (SD 0.998)
Adenovirus or Herpesviriade (yes/no)	n = 82, 55 (67.1)
Outcomes	
Slope Change in FEV_1	N =140, Mean -0.0199 (SD .266)
FEV_1	N = 779, Mean 1.5 (SD 0.6)
FEV_1PP	N = 779, Mean 51.1 (SD 15.6) %

Table 3.2.1 Comparisons of WLS Models with and without Outliers

Slope Change in FEV_1 with Smoking

Variable	with Outliers				without Outliers			
	Estimate	95% CI		P-Value	Estimate	95% CI		P-Value
		LCL	UCL			LCL	UCL	
Age	-0.003	-0.0079	0.0019	0.2322	-0.0022	-0.0055	0.0011	0.1858
Gender	0.2197	0.1594	0.2799	**<0.0001**	0.1325	0.0567	0.2082	**0.0007**
Height	-0.0097	-0.0158	-0.0037	**0.0019**	-0.0092	-0.0139	-0.0044	**0.0002**
Current Smoking	0.0426	-0.0612	0.1463	0.4186	-0.0869	-0.1601	-0.0138	**0.0203**
Breathlessness	0.0006	-0.0304	0.0315	0.9709	0.0125	-0.0149	0.0399	0.3689

Table 3.2.2 Univariate Models for Slope Change in FEV_1

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	MLR	0.0013	-0.0020	0.0045	0.4489
	WLS	-0.0012	-0.0043	0.0020	0.4628
	ROBUST	0.0016	-0.0013	0.0046	0.2804
Gender	MLR	0.0417	-0.0181	0.1014	0.1686
	WLS	-0.0541	-0.1056	-0.0026	**0.0396**
	ROBUST	0.0120	-0.0429	0.0669	0.6693
Height	MLR	-0.0002	-0.0034	0.0030	0.8884
	WLS	-0.0038	-0.0067	-0.0008	**0.0136**
	ROBUST	-0.0013	-0.0041	0.0014	0.3451
Current Smoking	MLR	0.0100	-0.0528	0.0727	0.7524
	WLS	-0.0753	-0.1383	-0.0123	**0.0198**
	ROBUST	-0.0174	-0.0731	0.0383	0.5405
C. *pneumoniae*	MLR	-0.0048	-0.0909	0.0813	0.9119
	WLS	-0.0497	-0.1140	0.0145	0.1273
	ROBUST	-0.0214	-0.0996	0.0567	0.5909

Table 3.2.3 Slope Change in FEV_1 with Smoking

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
	MLR	-0.0028	-0.0059	0.0003	**0.0762**
	WLS	-0.0022	-0.0055	0.0011	0.1858
Age	ROBUST	-0.0030	-0.0063	0.0002	**0.0699**
	MLR	0.0373	-0.0315	0.1061	0.2855
	WLS	0.1325	0.0567	0.2082	**0.0007**
Gender	ROBUST	0.0210	-0.0495	0.0915	0.5591
	MLR	-0.0019	-0.0057	0.0020	0.3365
	WLS	-0.0092	-0.0139	-0.0044	**0.0002**
Height	ROBUST	-0.0015	-0.0054	0.0024	0.4555
	MLR	-0.0457	-0.1045	0.0130	0.1260
	WLS	-0.0869	-0.1601	-0.0138	**0.0203**
Current Smoking	ROBUST	-0.0422	-0.1029	0.0185	0.1733
	MLR	0.0141	-0.0105	0.0387	0.2586
	WLS	0.0125	-0.0149	0.0399	0.3689
Breathlessness	ROBUST	0.0132	-0.0116	0.0379	0.2964

Table 3.2.4 Slope Change in FEV_1 with C. *pneumoniae*

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
	MLR	-0.0028	-0.0059	0.0003	**0.0794**
	WLS	-0.0022	-0.0055	0.0011	0.1866
Age	ROBUST	-0.0030	-0.0064	0.0003	**0.0767**
	MLR	0.0368	-0.0324	0.1060	0.2948
	WLS	0.1339	0.0587	0.2091	**0.0006**
Gender	ROBUST	0.0197	-0.0520	0.0915	0.5902
	MLR	-0.0019	-0.0057	0.0020	0.3386
	WLS	-0.0088	-0.0135	-0.0040	**0.0004**
Height	ROBUST	-0.0015	-0.0054	0.0025	0.4723
	MLR	-0.0459	-0.1049	0.0132	0.1265
	WLS	-0.0898	-0.1625	-0.0171	**0.0160**
Current Smoking	ROBUST	-0.0419	-0.1037	0.0199	0.1836
	MLR	0.0141	-0.0106	0.0388	0.2605
	WLS	0.0134	-0.0138	0.0407	0.3300
Breathlessness	ROBUST	0.0130	-0.0120	0.0381	0.3083
	MLR	-0.0098	-0.0959	0.0763	0.8221
	WLS	-0.0584	-0.1283	0.0114	**0.1002**
C. *pneumoniae*	ROBUST	-0.0063	-0.0933	0.0807	0.8865

Table 3.3.1 Univariate Models for Baseline FEV_1

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Age	-0.0228	-0.0306	-0.0150	**<0.0001**
Gender	0.4411	0.2946	0.5876	**<0.0001**
Height	0.0333	0.0262	0.0403	**<0.0001**
Current Smoking	0.2746	0.1174	0.4317	**0.0007**
Pack Years of Smoking	-0.0008	-0.0049	0.0033	0.7022
Quit Years of Smoking	-0.0018	-0.0110	0.0075	0.7036
Flu Shot in Last Year	-0.1842	-0.4121	0.0437	0.1125
Number of Flu Shot in Last 5 Years	-0.0334	-0.0770	0.0101	0.1315
Pneumonia Shot	-0.2302	-0.3898	-0.0706	**0.0049**
Sputum	0.0657	-0.1318	0.2632	0.5127
Sputum Amount	0.0209	-0.0408	0.0825	0.5054
Current Cough	-0.0783	-0.2817	0.1251	0.4485
Productive Cough	0.0497	-0.1355	0.2350	0.5970
Breathlessness	-0.0112	-0.0869	0.0646	0.7719
C. *pneumoniae* at Baseline	-0.1001	-0.5303	0.3300	0.6467

Table 3.3.2 Univariate Models for Baseline FEV_1PP

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Current Smoking	4.9169	0.4805	9.3532	**0.0300**
Pack Years of Smoking	-0.0947	-0.2074	0.0180	**0.0990**
Quit Years of Smoking	0.0125	-0.2442	0.2692	0.9238
Flu Shot in Last Year	-2.3543	-8.7102	4.0015	0.4660
Number of Flu Shots in Last 5 Years	0.1205	-1.0945	1.3356	0.8451
Pneumonia Shot	-5.0551	-9.5176	-0.5927	**0.0266**
Sputum	-2.3432	-7.8202	3.1337	0.3999
Sputum Amount	-0.9240	-2.6320	0.7840	0.2873
Current Cough	-2.6025	-8.2428	3.0379	0.3640
Productive Cough	-2.4333	-7.5648	2.6982	0.3509
Breathlessness	-0.8269	-2.9255	1.2718	0.4381
C. *pneumoniae* at Baseline	-7.9181	-19.8085	3.9723	0.1906

Table 3.3.3 Baseline FEV_1 with Smoking

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Age	-0.0164	-0.0241	-0.0087	**<0.0001**
Gender	0.1541	-0.0185	0.3267	**0.0799**
Height	0.0242	0.0148	0.0335	**<0.0001**
Current Smoking	0.0675	-0.0796	0.2146	0.3666
Breathlessness	-0.0264	-0.0873	0.0345	0.3935

Table 3.3.4 Baseline FEV_1PP with Smoking

Variable	**Estimate**	**95% CI**		**P-Value**
		LCL	**UCL**	
Current Smoking	5.1728	0.7101	9.6354	**0.0233**
Breathlessness	-1.0924	-3.1817	0.9968	0.3037

Table 3.3.5 Baseline FEV_1 with C. *pneumoniae*

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Age	-0.0167	-0.0244	-0.0090	**<0.0001**
Gender	0.1593	-0.0136	0.3321	**0.0707**
Height	0.0241	0.0147	0.0334	**<0.0001**
Current smoking	0.0616	-0.0858	0.2091	0.4108
Breathlessness	-0.0247	-0.0856	0.0362	0.4246
C.*pneumoniae* at Baseline	-0.1855	-0.5272	0.1562	0.2856

Table 3.3.6 Baseline FEV_1PP with C. *pneumoniae*

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Current smoking	5.0446	0.5812	9.5081	**0.0270**
Breathlessness	-1.0294	-3.1193	1.0605	0.3325
C. *pneumoniae* at Baseline	-7.0820	-18.8952	4.7312	0.2385

Table 3.3.7 Comparisons of GEE Models with and without Outliers

Variable	with Outliers				without Outliers			
	Estimate	95% CI		P-Value	Estimate	95% CI		P-Value
		LCL	UCL			LCL	UCL	
Age	-0.0170	-0.0243	-0.0098	**<0.0001**	-0.0165	-0.0237	-0.0093	**<0.0001**
Gender	0.1632	-0.0183	0.3446	**0.0780**	0.2062	0.0394	0.3731	**0.0154**
Height	0.0246	0.0149	0.0344	**<0.0001**	0.0222	0.0133	0.0312	**<0.0001**
Current smoking	0.0240	-0.1006	0.1486	0.7055	0.0231	-0.1017	0.1480	0.7164
Breathlessness	-0.1271	-0.2203	-0.0339	**0.0075**	-0.1132	-0.2028	-0.0237	**0.0132**

Table 3.3.8 Univariate GEE Models for FEV_1 Measured at Multiple Times

Variable	Estimate	95% CI		P-Value	ρ
		LCL	UCL		
Age	-0.0240	-0.0321	-0.0158	**<0.0001**	0.9394
Gender	0.4557	0.3147	0.5968	**<0.0001**	0.9193
Height	0.0340	0.0273	0.0408	**<0.0001**	0.9034
Current smoking	0.2734	0.1156	0.4313	**0.0007**	0.9420
Pack Years of Smoking	-0.0006	-0.0049	0.0037	0.7884	0.9491
Quit Years of Smoking	-0.0025	-0.0122	0.0071	0.6067	0.9510
Flu Shot in Last Year	-0.2147	-0.4983	0.0688	0.1377	0.9463
Number of Flu Shots in Last 5 Years	-0.0358	-0.0821	0.0106	0.1309	0.9431
Pneumonia Shot	-0.3031	-0.4777	-0.1286	**0.0007**	0.9311
Sputum	0.0496	-0.1284	0.2276	0.5849	0.9493
Sputum Amount	0.0111	-0.0595	0.0818	0.7578	0.8828
Current Smoking	-0.0882	-0.2812	0.1049	0.3706	0.9488
Productive Cough	0.0327	-0.1333	0.1987	0.6995	0.9485
Breathlessness	-0.1565	-0.2750	-0.0380	**0.0096**	0.9519
C. *pneumoniae* at Baseline &Follow-up	-0.0813	-0.5191	0.3565	0.7159	0.9454

Table 3.3.9 Univariate GEE Models for FEV_1PP Measured at Multiple Times

Variable	Estimate	95% CI		P-Value	ρ
		LCL	UCL		
Current smoking	4.5148	0.1770	8.8525	**0.0414**	0.8903
Pack Years of Smoking	-0.0838	-0.1892	0.0216	0.1190	0.8929
Quit Years of Smoking	0.0078	-0.3077	0.3234	0.9611	0.8910
Flu Shot in Last Year	-1.3569	-8.3553	5.6415	0.7039	0.8841
Number of Flu Shots in Last 5 Years	0.4093	-0.7502	1.5689	0.4890	0.8852
Pneumonia Shot	-6.1710	-10.8633	-1.4788	**0.0099**	0.8797
Sputum	-3.4493	-8.8061	1.9074	0.2069	0.8841
Sputum Amount	-0.8615	-3.0014	1.2785	0.4301	0.9092
Current Smoking	-3.0441	-8.5696	2.4815	0.2802	0.8896
Productive Cough	-3.5432	-8.5184	1.4321	0.1628	0.8824
Breathlessness	-5.1998	-8.6186	-1.7809	**0.0029**	0.8923
C. *pneumoniae* at Baseline & Follow-up	-2.3344	-13.028	8.3592	0.6688	0.8878

Table 3.3.10 FEV_1 Measured over Time with Smoking

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	GEE	-0.0170	-0.0243	-0.0098	**<0.0001**
	MIXED	-0.0164	-0.0238	-0.0091	**<0.0001**
	RRE	-0.0163	-0.0236	-0.0089	**<0.0001**
Gender	GEE	0.1632	-0.0183	0.3446	**0.0780**
	MIXED	0.1487	-0.0282	0.3257	**0.0990**
	RRE	0.1466	-0.0177	0.3109	**0.0800**
Height	GEE	0.0246	0.0149	0.0344	**<0.0001**
	MIXED	0.0250	0.0160	0.0339	**<0.0001**
	RRE	0.0251	0.0161	0.0341	**<0.0001**
Current Smoking	GEE	0.0240	-0.1006	0.1486	0.7055
	MIXED	0.0167	-0.1226	0.1560	0.8138
	RRE	0.0149	-0.1244	0.1543	0.8340
Breathlessness	GEE	-0.1271	-0.2203	-0.0339	**0.0075**
	MIXED	-0.1341	-0.2218	-0.0463	**0.0028**
	RRE	-0.1361	-0.2240	-0.0482	**0.0020**

Table 3.3.11 FEV_1PP Measured at Multiple Times with Smoking

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Current Smoking	GEE	3.5433	-0.6963	7.7830	0.1014
	MIXED	3.1572	-1.1134	7.4278	0.1470
	RRE	3.0911	-1.1856	7.3679	0.1570
Breathlessness	GEE	-4.8432	-8.2838	-1.4026	**0.0058**
	MIXED	-5.1317	-8.1558	-2.1076	**0.0009**
	RRE	-5.1849	-8.2131	-2.1568	**<0.0001**

Table 3.3.12 FEV_1 Measured at Multiple Times with C. *pneumoniae*

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	GEE	-0.0170	-0.0243	-0.0097	**<0.0001**
	MIXED	-0.0164	-0.0238	-0.0090	**<0.0001**
	RRE	-0.0163	-0.0248	-0.0078	**<0.0001**
Gender	GEE	0.1614	-0.0183	0.3411	**0.0784**
	MIXED	0.1482	-0.0166	0.3130	**0.0778**
	RRE	0.1452	-0.0198	0.3102	**0.0850**
Height	GEE	0.0247	0.0149	0.0345	**<0.0001**
	MIXED	0.0250	0.0160	0.0340	**<0.0001**
	RRE	0.0251	0.0161	0.0341	**<0.0001**
Current Smoking	GEE	0.0229	-0.1022	0.1480	0.7196
	MIXED	0.0160	-0.1237	0.1557	0.8222
	RRE	0.0142	-0.1256	0.1540	0.8420
Breathlessness	GEE	-0.1286	-0.2210	-0.0361	**0.0064**
	MIXED	-0.1352	-0.2236	-0.0468	**0.0028**
	RRE	-0.1373	-0.2258	-0.0488	**0.0020**
C. *pneumoniae* at Baseline & Follow-up	GEE	-0.0449	-0.3036	0.2138	0.7339
	MIXED	-0.0338	-0.2847	0.2170	0.7911
	RRE	-0.0346	-0.2860	0.2168	0.7870

Table 3.3.13 FEV_1PP Measured at Multiple Times with C. *pneumoniae*

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Current Smoking	GEE	3.4500	-0.8494	7.7495	0.1158
	MIXED	3.1034	-1.1797	7.3865	0.1552
	RRE	3.0379	-1.2526	7.3285	0.1650
Breathlessness	GEE	-4.9481	-8.3638	-1.5325	**0.0045**
	MIXED	-5.2106	-8.2552	-2.1661	**0.0008**
	RRE	-5.2632	-8.3127	-2.2138	**<0.0001**
C. *pneumoniae* at Baseline & Follow-up	GEE	-3.2143	-12.4137	5.9851	0.4935
	MIXED	-2.2553	-10.9427	6.4322	0.6103
	RRE	-2.2446	-10.9541	6.4649	0.6130

Table 3.4.1 Univariate Models for Slope Change in FEV_1

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Log Matrix Metalloproteinase 9	MLR	-0.0972	0.0507	-0.1946	**0.0003**
	WLS	-0.1779	-0.2346	-0.1212	**<0.0001**
	ROBUST	-0.0872	-0.1787	0.0043	**0.0619**
Log C-reactive Protein	MLR	0.0291	-0.0409	0.0991	0.4103
	WLS	0.0798	0.0260	0.1337	**0.0042**
	ROBUST	0.0014	-0.0611	0.0638	0.9659
Log Ratio of MMP-9 to TIMP-1	MLR	-0.0821	-0.1560	-0.0081	**0.0301**
	WLS	-0.1305	-0.1712	-0.0898	**<0.0001**
	ROBUST	-0.0698	-0.1385	-0.0011	**0.0464**
Adenovirus or Herpesviriade	MLR	0.0271	-0.0260	0.0801	0.3126
	WLS	-0.0354	-0.0880	0.0173	0.1846
	ROBUST	-0.0296	-0.0772	0.0179	0.2223

Table 3.4.2 Slope Change in FEV_1 with MMP-9

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	MLR	0.0008	-0.0030	0.0045	0.6848
	WLS	-0.0022	-0.0050	0.0007	0.1407
	ROBUST	-0.0011	-0.0047	0.0026	0.5629
Gender	MLR	0.0972	0.0069	0.1875	**0.0352**
	WLS	0.0855	0.0015	0.1694	**0.0461**
	ROBUST	0.0582	-0.0253	0.1418	0.1720
Height	MLR	-0.0044	-0.0094	0.0005	**0.0796**
	WLS	-0.0059	-0.0107	-0.0011	**0.0167**
	ROBUST	-0.0037	-0.0082	0.0008	0.1090
Current Smoking	MLR	0.0368	-0.0371	0.1107	0.3234
	WLS	-0.0404	-0.1054	0.0246	0.2191
	ROBUST	-0.0070	-0.0768	0.0628	0.8440
Breathlessness	MLR	0.0068	-0.0221	0.0357	0.6415
	WLS	-0.0095	-0.0331	0.0141	0.4244
	ROBUST	0.0120	-0.0145	0.0385	0.3745
Log Matrix Metalloproteinase 9	MLR	-0.1033	-0.2012	-0.0054	**0.0389**
	WLS	-0.1814	-0.2473	-0.1156	**<0.0001**
	ROBUST	-0.0910	-0.1833	0.0014	**0.0535**

Table 3.4.3 Slope Change in FEV_1 with Ratio of MMP-9 to TIMP-1

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	MLR	0.0003	-0.0035	0.0040	0.8889
	WLS	-0.0026	-0.0055	0.0002	**0.0668**
	ROBUST	-0.0014	-0.005	0.0022	0.4442
Gender	MLR	0.0964	0.0065	0.1864	**0.0361**
	WLS	0.0778	-0.0041	0.1597	**0.0623**
	ROBUST	0.0561	-0.0274	0.1395	0.1878
Height	MLR	-0.0046	-0.0095	0.0004	**0.0708**
	WLS	-0.0057	-0.0104	-0.0009	**0.0199**
	ROBUST	-0.0037	-0.0082	0.0008	0.1088
Current Smoking	MLR	0.0373	-0.0364	0.1109	0.3164
	WLS	-0.0401	-0.1043	0.0241	0.2171
	ROBUST	-0.0061	-0.0759	0.0637	0.8640
Breathlessness	MLR	0.0063	-0.0225	0.0351	0.6641
	WLS	-0.0119	-0.0355	0.0116	0.3160
	ROBUST	0.0114	-0.015	0.0379	0.3959
Log MMP-9 to TIMP-1	MLR	-0.0833	-0.1586	-0.0081	**0.0305**
	WLS	-0.1322	-0.1786	-0.0858	**<0.0001**
	ROBUST	-0.0705	-0.1410	0.0001	**0.0503**

Table 3.4.4 Slope Change in FEV_1 with CRP

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	MLR	0.0015	-0.0025	0.0055	0.4481
	WLS	-0.0002	-0.0040	0.0036	0.9212
	ROBUST	0.0004	-0.0035	0.0043	0.8263
Gender	MLR	0.0843	-0.0099	0.1784	**0.0785**
	WLS	-0.0252	-0.1154	0.0650	0.5794
	ROBUST	0.0560	-0.0293	0.1413	0.1982
Height	MLR	-0.0028	-0.0081	0.0025	0.2994
	WLS	-0.0004	-0.0060	0.0051	0.8827
	ROBUST	-0.0028	-0.0075	0.0020	0.2543
Current Smoking	MLR	0.0200	-0.0546	0.0945	0.5946
	WLS	-0.1010	-0.1728	-0.0293	**0.0065**
	ROBUST	-0.0113	-0.0796	0.0571	0.7471
Breathlessness	MLR	0.0102	-0.0195	0.0399	0.4965
	WLS	0.0152	-0.0108	0.0413	0.2472
	ROBUST	0.0080	-0.0188	0.0349	0.5584
Log C-reactive Protein	MLR	0.0309	-0.0430	0.1049	0.4069
	WLS	0.0740	0.0151	0.1328	**0.0145**
	ROBUST	-0.0050	-0.0737	0.0637	0.8860

Table 3.4.5 Slope Change in FEV_1 with Viral Infection

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	MLR	0.0005	-0.0033	0.0044	0.7760
	WLS	-0.0036	-0.0070	-0.0002	**0.0396**
	ROBUST	-0.0006	-0.0043	0.0032	0.7647
Gender	MLR	0.0902	-0.0026	0.1830	**0.0565**
	WLS	0.0115	-0.0843	0.1073	0.8113
	ROBUST	0.0548	-0.0298	0.1394	0.2042
Height	MLR	-0.0041	-0.0092	0.0009	0.1095
	WLS	-0.0031	-0.0087	0.0025	0.2754
	ROBUST	-0.0033	-0.0079	0.0012	0.1550
Current Smoking	MLR	0.0291	-0.0461	0.1042	0.4429
	WLS	-0.1056	-0.1787	-0.0324	**0.0053**
	ROBUST	-0.0086	-0.0789	0.0616	0.8095
Breathlessness	MLR	0.0079	-0.0218	0.0375	0.5988
	WLS	0.0108	-0.0169	0.0384	0.4397
	ROBUST	0.0111	-0.0160	0.0381	0.4215
Adenovirus or Herpesviriade	MLR	-0.0292	-0.0831	0.0247	0.2828
	WLS	0.0239	-0.0283	0.0760	0.3643
	ROBUST	-0.0245	-0.0740	0.0251	0.3333

Table 3.4.6 Baseline FEV_1 with MMP-9

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Age	-0.0120	-0.0232	-0.0007	**0.0379**
Gender	0.1205	-0.1435	0.3845	0.3661
Height	0.0311	0.0166	0.0457	**<0.0001**
Current Smoking	0.0651	-0.1436	0.2738	0.5362
Breathlessness	0.0033	-0.1324	0.1391	0.9612
Log Matrix Metalloproteinase 9	-0.0775	-0.3692	0.2142	0.5983

Table 3.4.7 Baseline FEV_1PP with MMP-9

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Current Smoking	5.0085	-1.2431	11.2601	0.1148
Breathlessness	0.4639	-3.9725	4.9002	0.8356
Log Matrix Metalloproteinase 9	-3.6566	-13.3225	6.0093	0.4536

Table 3.4.8 Baseline FEV_1 with MMP-9_TIMP-1

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Age	-0.0122	-0.0235	-0.0009	**0.0342**
Gender	0.1186	-0.1455	0.3826	0.3739
Height	0.0312	0.0167	0.0458	**<0.0001**
Current Smoking	0.0626	-0.1462	0.2714	0.5521
Breathlessness	0.0029	-0.1329	0.1388	0.9660
Log Ratio of MMP-9 to TIMP-1	-0.0423	-0.2652	0.1806	0.7065

Table 3.4.9 Baseline FEV_1PP with Ratio of MMP-9 to TIMP-1

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Current Smoking	4.9968	-1.3009	11.2944	0.1182
Breathlessness	0.4085	-4.0346	4.8516	0.8552
Log Ratio of MMP-9 to TIMP-1	-1.9573	-9.3436	5.4289	0.5993

Table 3.4.10 Baseline FEV_1 with CRP

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Age	-0.0145	-0.0264	-0.0025	**0.0184**
Gender	0.1671	-0.1017	0.4360	0.2194
Height	0.0280	0.0128	0.0433	**0.0005**
Current Smoking	0.0459	-0.1599	0.2518	0.6578
Breathlessness	0.0137	-0.1220	0.1493	0.8412
Log C-reactive Protein	-0.1659	-0.3685	0.0368	0.1071

Table 3.4.11 Baseline FEV_1PP with CRP

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Current Smoking	4.5113	-1.6881	10.7107	0.1514
Breathlessness	0.6891	-3.6902	5.0685	0.7549
Log C-reactive Protein	-6.2396	-12.5096	0.0305	**0.0511**

Table 3.4.12 Baseline FEV_1 with Viral Infection

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Age	-0.0122	-0.0235	-0.0008	**0.0361**
Gender	0.1171	-0.1481	0.3824	0.3818
Height	0.0315	0.0171	0.0460	**<0.0001**
Current Smoking	0.0564	-0.1504	0.2631	0.5886
Breathlessness	0.0024	-0.1338	0.1385	0.9727
Adenovirus or Herpesviriade	-0.0177	-0.3063	0.2709	0.9031

Table 3.4.13 Baseline FEV_1PP with Viral Infection

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
Current Smoking	5.2463	-0.8709	11.3636	**0.0917**
Breathlessness	0.4631	-3.8884	4.8145	0.8328
Adenovirus or Herpesviriade	-4.6955	-9.6138	0.2228	0.0610

Table 3.4.14 Univariate GEE Models for FEV_1 Measured Over Time

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
AGE	-0.0215	-0.0346	-0.0085	**0.0012**
GENDER	0.5537	0.3585	0.7490	**<0.0001**
HEIGHT	0.0391	0.0292	0.0489	**<0.0001**
CURRENT SMOKING	0.3085	0.0702	0.5467	**0.0112**
BREATHLESSNESS	-0.0579	-0.2308	0.1150	0.5117
LOG MATRIX METALLOPROTEINASE 9	-0.2395	-0.5720	0.0930	0.1581
LOG C-REACTIVE PROTEIN	-0.2673	-0.5002	-0.0344	**0.0245**
LOG RATIO OF MMP9 TO TISSUE INHIBITOR OF METALLOPR0PTEINASE 1	-0.133	-0.3748	0.1089	0.2813
ADENOVIRUS OR HERPESVIRIADE	-0.2363	-0.4362	-0.0365	**0.0205**

Table 3.4.15 Univariate GEE Models for FEV_1PP Measured Over Time

Variable	Estimate	95% CI		P-Value
		LCL	UCL	
CURRENT SMOKING	4.1036	-1.6207	9.8278	0.1600
BREATHLESSNESS	-0.9017	-4.8984	3.0951	0.6584
LOG OF MATRIX METALLOPROTEINASE 9	-6.3888	-14.195	1.4174	0.1087
LOG OF C-REACTIVE PROTEIN	-7.5017	-12.9637	-2.0397	**0.0071**
LOG OF RATIO OF MMP9 TO TISSUE INHIBITOR OF METALLOPR0PTEINASE 1	-3.8400	-9.5689	1.8890	0.1889
ADENOVIRUS OR HERPESVIRIADE	-5.1020	-9.7881	-0.4160	**0.0328**

Table 3.4.16 FEV_1 Measured at Multiple-Time with MMP-9

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	GEE	-0.0132	-0.0251	-0.0012	**0.0312**
	MIXED	-0.0121	-0.0239	-0.0002	**0.0455**
	RRE	-0.0121	-0.0255	0.0013	**0.0760**
Gender	GEE	0.1552	-0.1327	0.4430	0.2907
	MIXED	0.1086	-0.1701	0.3873	0.4440
	RRE	0.1076	-0.1685	0.3836	0.4450
Height	GEE	0.0290	0.0129	0.0451	**0.0004**
	MIXED	0.0308	0.0155	0.0461	**<0.0001**
	RRE	0.0308	0.0146	0.0470	**<0.0001**
Current Smoking	GEE	0.0624	-0.1275	0.2523	0.5196
	MIXED	0.0705	-0.1494	0.2904	0.5286
	RRE	0.0701	-0.1667	0.3070	0.5620
Breathlessness	GEE	0.0019	-0.1179	0.1217	0.9751
	MIXED	-0.0065	-0.1494	0.1365	0.9290
	RRE	-0.0069	-0.1232	0.1094	0.9080
Log Matrix Metalloproteinase 9	GEE	-0.1802	-0.4189	0.0586	0.1392
	MIXED	-0.1508	-0.4576	0.1560	0.3344
	RRE	-0.1498	-0.4294	0.1298	0.2940

Table 3.4.17 FEV_1PP Measured at Multiple-Time with MMP-9

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Current Smoking	GEE	4.5469	-1.3699	10.4636	0.1320
	MIXED	4.6609	-1.8088	11.1305	0.1574
	RRE	4.6567	-1.5723	10.8856	0.1430
Breathlessness	GEE	-0.0708	-4.0122	3.8706	0.9719
	MIXED	-0.4947	-5.0807	4.0913	0.8321
	RRE	-0.5027	-4.6382	3.6328	0.8120
Log Matrix Metalloproteinase 9	GEE	-7.1560	-14.8068	0.4948	**0.0668**
	MIXED	-6.1423	-16.1454	3.8607	0.2280
	RRE	-6.1265	-15.3572	3.1041	0.1930

Table 3.4.18 FEV_1 Measured Over Time with Ratio of MMP-9 to TIMP-1

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	GEE	-0.0139	-0.0259	-0.0019	**0.0227**
	MIXED	-0.0127	-0.0246	-0.0008	**0.0364**
	RRE	-0.0127	-0.0259	0.0005	**0.0600**
Gender	GEE	0.1528	-0.1393	0.4448	0.3053
	MIXED	0.1062	-0.1728	0.3851	0.4546
	RRE	0.1052	-0.1718	0.3821	0.4570
Height	GEE	0.0290	0.0127	0.0452	**0.0005**
	MIXED	0.0308	0.0154	0.0462	**<0.0001**
	RRE	0.0308	0.0146	0.0471	**<0.0001**
Current Smoking	GEE	0.0601	-0.1291	0.2493	0.5337
	MIXED	0.0681	-0.1521	0.2882	0.5436
	RRE	0.0677	-0.1696	0.3050	0.5760
Breathlessness	GEE	0.0011	-0.1193	0.1215	0.9859
	MIXED	-0.0072	-0.1504	0.1359	0.9210
	RRE	-0.0076	-0.1243	0.1091	0.8980
Log Ratio of MMP-9 to TIMP-1	GEE	-0.1221	-0.2978	0.0535	0.1730
	MIXED	-0.0994	-0.3340	0.1352	0.4052
	RRE	-0.0986	-0.3070	0.1097	0.3530

Table 3.4.19 FEV_1PP Measured at Multiple-Time with MMP-9 to TIMP-1

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Current Smoking	GEE	4.6395	-1.2744	10.5533	0.1241
	MIXED	4.7184	-1.8048	11.2416	0.1557
	RRE	4.7138	-1.6021	11.0297	0.1440
Breathlessness	GEE	-0.1817	-4.1559	3.7926	0.9286
	MIXED	-0.5904	-5.1883	4.0075	0.8008
	RRE	-0.5981	-4.7543	3.5581	0.7780
Log Ratio of MMP-9 to TIMP-1	GEE	-4.7287	-10.3992	0.9418	0.1022
	MIXED	-3.8849	-11.5353	3.7655	0.3186
	RRE	-3.8729	-11.1048	3.3589	0.2940

Table 3.4.20 FEV_1 Measured at Multiple-Time with CRP

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	GEE	-0.0162	-0.0286	-0.0038	**0.0104**
	MIXED	-0.0152	-0.0276	-0.0027	**0.0170**
	RRE	-0.0152	-0.0302	-0.0002	**0.0480**
Gender	GEE	0.2125	-0.0838	0.5088	0.1598
	MIXED	0.1675	-0.1140	0.4491	0.2425
	RRE	0.1667	-0.1074	0.4409	0.2330
Height	GEE	0.0255	0.0082	0.0429	**0.0040**
	MIXED	0.0272	0.0112	0.0432	**0.0009**
	RRE	0.0272	0.0098	0.0446	**0.0020**
Current Smoking	GEE	0.0293	-0.1482	0.2068	0.7460
	MIXED	0.0376	-0.1774	0.2527	0.7308
	RRE	0.0373	-0.1894	0.2640	0.7470
Breathlessness	GEE	0.0140	-0.1016	0.1296	0.8123
	MIXED	0.0057	-0.1359	0.1473	0.9370
	RRE	0.0053	-0.1104	0.1211	0.9280
Log of C-reactive Protein	GEE	-0.2242	-0.3922	-0.0562	**0.0089**
	MIXED	-0.2231	-0.4345	-0.0116	**0.0387**
	RRE	-0.2234	-0.4452	-0.0017	**0.0480**

Table 3.4.21 FEV_1PP Measured at Multiple-Time with CRP

Variable	**Model**	**Estimate**	**95% CI**		**P-Value**
			LCL	UCL	
Current Smoking	GEE	3.6083	-2.1274	9.3440	0.2176
	MIXED	3.7174	-2.6499	10.0848	0.2516
	RRE	3.7145	-2.3367	9.7657	0.2290
Breathlessness	GEE	0.1388	-3.6941	3.9718	0.9434
	MIXED	-0.2545	-4.7478	4.2389	0.9114
	RRE	-0.2609	-4.2552	3.7335	0.8980
Log of C-reactive Protein	GEE	-7.4112	-12.8041	-2.0182	**0.0071**
	MIXED	-7.8154	-14.2592	-1.3717	**0.0176**
	RRE	-7.8218	-14.2293	-1.4143	**0.0170**

Table 3.4.22 FEV1 Measured at Multiple-Time with Viral Infection

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Age	GEE	-0.0140	-0.0255	-0.0026	**0.0161**
	MIXED	-0.0129	-0.0245	-0.0013	**0.0293**
	RRE	-0.0129	-0.0258	0.0000	**0.0490**
Gender	GEE	0.1225	-0.1611	0.4060	0.3974
	MIXED	0.0776	-0.1954	0.3506	0.5764
	RRE	0.0766	-0.1942	0.3474	0.5790
Height	GEE	0.0290	0.0137	0.0443	**0.0002**
	MIXED	0.0307	0.0158	0.0457	**<0.0001**
	RRE	0.0308	0.0153	0.0463	**<0.0001**
Current Smoking	GEE	0.0646	-0.1125	0.2417	0.4744
	MIXED	0.0738	-0.1399	0.2875	0.4975
	RRE	0.0735	-0.1518	0.2988	0.5230
Breathlessness	GEE	0.0055	-0.1168	0.1277	0.9302
	MIXED	-0.0026	-0.1426	0.1374	0.9710
	RRE	-0.0030	-0.1160	0.1100	0.9590
Adenovirus or Herpesviriade	GEE	-0.1742	-0.3026	-0.0459	**0.0078**
	MIXED	-0.1642	-0.3215	-0.0068	**0.0409**
	RRE	-0.1638	-0.3107	-0.0169	**0.0290**

Table 3.4.23 FEV_1PP Measured at Multiple-Time with Viral Infection

Variable	Model	Estimate	95% CI		P-Value
			LCL	UCL	
Current Smoking	GEE	4.6010	-0.9545	10.1564	0.1045
	MIXED	4.7263	-1.6444	11.0969	0.1454
	RRE	4.7235	-1.3770	10.8240	0.1290
Breathlessness	GEE	-0.1055	-4.1785	3.9675	0.9595
	MIXED	-0.5249	-5.0517	4.0020	0.8197
	RRE	-0.5327	-4.6202	3.5548	0.7980
Adenovirus or Herpesviriade	GEE	-5.4163	-9.8763	-0.9563	**0.0173**
	MIXED	-4.8400	-9.9611	0.2811	**0.0639**
	RRE	-4.8356	-9.7518	0.0806	**0.0540**

Table 3.5.1 Univariate GEE Models for FEV_1 with Imputed Data

Variable	Method	Estimate	95% CI		P-Value	ρ
			LCL	UCL		
Age	LOCF	-0.0235	-0.0318	-0.0153	**<0.0001**	0.9671
	MEAN	-0.0101	-0.0141	-0.006	**<0.0001**	0.3362
	HOTDECK	-0.0212	-0.0273	-0.0151	**<0.0001**	0.3927
	MI	-0.0141	-0.0184	-0.0099	**<0.0001**	0.2174
Gender	LOCF	0.4504	0.3063	0.5945	**<0.0001**	0.9684
	MEAN	0.2139	0.1451	0.2828	**<0.0001**	0.3261
	HOTDECK	0.4929	0.3949	0.5909	**<0.0001**	0.3340
	MI	0.2191	0.1365	0.3017	**<0.0001**	0.2320
Height	LOCF	0.0342	0.0273	0.0410	**<0.0001**	0.9624
	MEAN	0.0152	0.0113	0.0191	**<0.0001**	0.2929
	HOTDECK	0.0309	0.0260	0.0358	**<0.0001**	0.2988
	MI	0.0168	0.0125	0.0212	**<0.0001**	0.2074
Current Smoking	LOCF	0.2572	0.0968	0.4176	**0.0017**	0.9711
	MEAN	0.1196	0.0435	0.1957	**0.0021**	0.3537
	HOTDECK	0.2307	0.1087	0.3527	**0.0002**	0.4384
	MI	0.1563	0.0688	0.2439	**0.0006**	0.2399
Pack Years of Smoking	LOCF	-0.0009	-0.0052	0.0034	0.6828	0.9723
	MEAN	0.0000	-0.0023	0.0023	0.9792	0.3645
	HOTDECK	0.0011	-0.0025	0.0046	0.5524	0.4539
	MI	-0.0008	-0.0031	0.0014	0.4674	0.2511
Quit Years of Smoking	LOCF	-0.0021	-0.0118	0.0076	0.6755	0.9727
	MEAN	-0.0002	-0.0046	0.0042	0.9224	0.3640
	HOTDECK	-0.0015	-0.0082	0.0053	0.6679	0.4623
	MI	0.0014	-0.0034	0.0061	0.5704	0.2542
Flu Shot in Last Year	LOCF	-0.1925	-0.4686	0.0836	0.1718	0.9713
	MEAN	-0.0631	-0.1783	0.0521	0.2829	0.3629
	HOTDECK	-0.1992	-0.3873	-0.0111	**0.0379**	0.4498
	MI	-0.1931	-0.3702	-0.016	**0.0340**	0.2407
Number of Flu Shots in Last 5 Years	LOCF	-0.0283	-0.0724	0.0157	0.2076	0.9716
	MEAN	-0.0140	-0.0370	0.0090	0.2313	0.3618
	HOTDECK	-0.0506	-0.0847	-0.0166	**0.0036**	0.4432
	MI	-0.0411	-0.0664	-0.0159	**0.0019**	0.2381
Pneumonia Shot	LOCF	-0.3055	-0.4663	-0.1446	**0.0002**	0.9699
	MEAN	-0.1645	-0.2574	-0.0716	**0.0005**	0.3441
	HOTDECK	-0.0634	-0.1833	0.0564	0.2996	0.4527
	MI	-0.3685	-0.4516	-0.2854	**<0.0001**	0.1800

Table 3.5.1 Univariate GEE Models for FEV_1 with Imputed Data (Cont'd)

Variable	Method	Estimate	95% CI		P-Value	ρ
			LCL	UCL		
Sputum	LOCF	0.0683	-0.113	0.2495	0.4604	0.9724
	MEAN	0.0125	-0.0694	0.0943	0.7652	0.3646
	HOTDECK	0.0198	-0.1224	0.1619	0.7852	0.4550
	MI	-0.0031	-0.1513	0.1452	0.9663	0.2502
Sputum Amount	LOCF	0.0304	-0.0283	0.0891	0.3095	0.9725
	MEAN	0.0011	-0.025	0.0271	0.9351	0.3646
	HOTDECK	0.0335	-0.0064	0.0734	**0.0998**	0.4480
	MI	0.0059	-0.067	0.0788	0.8622	0.2496
Current Cough	LOCF	-0.0727	-0.2698	0.1245	0.4701	0.9720
	MEAN	-0.0222	-0.109	0.0646	0.6169	0.3641
	HOTDECK	-0.0476	-0.1866	0.0914	0.5023	0.4549
	MI	-0.0118	-0.1366	0.1131	0.8498	0.2507
Productive Cough	LOCF	0.0546	-0.1138	0.2229	0.5253	0.9724
	MEAN	0.0058	-0.0703	0.0819	0.8816	0.3658
	HOTDECK	0.0385	-0.0906	0.1676	0.5588	0.4561
	MI	0.0031	-0.1344	0.1405	0.9635	0.2504
Breathlessness	LOCF	-0.2564	-0.3958	-0.1169	**0.0003**	0.9710
	MEAN	-0.0395	-0.0793	0.0003	**0.0515**	0.3624
	HOTDECK	-0.0803	-0.1683	0.0076	**0.0734**	0.4519
	MI	-0.1414	-0.2423	-0.0405	**0.0103**	0.2228
C. *pneumoniae* at Baseline & Follow-up	LOCF	-0.0616	-0.5256	0.4023	0.7945	0.9722
	MEAN	-0.1308	-0.3212	0.0596	0.1781	0.3601
	HOTDECK	-0.0856	-0.4144	0.2433	0.6100	0.4543
	MI	-0.1116	-0.2714	0.0482	0.1702	0.2492

Table 3.5.2 Univariate GEE Models for FEV1PP with Imputations

Variable	Method	Estimate	95% CI		P-Value	ρ
			LCL	UCL		
Current Smoking	LOCF	4.5103	0.0726	8.9479	**0.0464**	0.9592
	MEAN	1.6891	-0.2091	3.5873	**0.0812**	0.3603
	HOTDECK	2.7761	0.2098	5.3425	**0.034**	0.2174
	MI	1.6769	-0.9357	4.2895	0.2027	0.2137
Pack Years of Smoking	LOCF	-0.0944	-0.2019	0.0131	**0.0851**	0.9588
	MEAN	-0.0295	-0.0834	0.0244	0.2828	0.3617
	HOTDECK	-0.0382	-0.1084	0.032	0.2862	0.2194
	MI	-0.0528	-0.1126	0.007	**0.0824**	0.2133
Quit Years of Smoking	LOCF	-0.0044	-0.3165	0.3076	0.9778	0.9596
	MEAN	0.0258	-0.1133	0.165	0.7158	0.3623
	HOTDECK	0.0653	-0.1031	0.2337	0.4473	0.225
	MI	0.0481	-0.1303	0.2264	0.5974	0.2163
Flu Shot in Last Year	LOCF	-1.7118	-8.6351	5.2115	0.628	0.9589
	MEAN	-0.2351	-3.1466	2.6764	0.8743	0.3627
	HOTDECK	-1.0809	-5.2052	3.0434	0.6075	0.219
	MI	-2.0312	-6.3653	2.3029	0.3444	0.2136
Number of Flu Shots in Last 5 Years	LOCF	0.3547	-0.7601	1.4696	0.5329	0.9593
	MEAN	0.145	-0.4136	0.7037	0.6109	0.3628
	HOTDECK	-0.0025	-0.6831	0.6782	0.9943	0.2192
	MI	0.0299	-0.7308	0.7906	0.9367	0.2156
Pneumonia Shot	LOCF	-5.4066	-9.8109	-1.0023	**0.0161**	0.9569
	MEAN	-2.9798	-5.2612	-0.6983	**0.0105**	0.3527
	HOTDECK	-2.6317	-5.2377	-0.0256	**0.0478**	0.2124
	MI	-6.5981	-9.4724	-3.7238	**<0.0001**	0.1837
Sputum	LOCF	-2.7963	-8.3177	2.7251	0.3209	0.9586
	MEAN	-1.9748	-4.3203	0.3707	**0.0989**	0.3579
	HOTDECK	-3.5147	-6.732	-0.2973	**0.0323**	0.2148
	MI	0.2142	0.2142	0.2142	0.2142	0.2142
Sputum Amount	LOCF	1.096	-0.6266	2.8186	0.2124	0.9598
	MEAN	-0.4552	-1.2484	0.338	0.2607	0.361
	HOTDECK	-0.5604	-1.3558	0.2351	0.1674	0.3601
	MI	-0.9828	-2.4678	0.5023	0.1773	0.2089

Table 3.5.2 Univariate GEE Models for FEV_1PP with Imputations (Cont'd)

Variable	Method	Estimate	95% CI		P-Value	ρ
			LCL	UCL		
Current Cough	LOCF	-2.764	-8.4387	2.9107	0.3398	0.9589
	MEAN	-1.0355	-3.421	1.3499	0.3949	0.3614
Current Cough	HOTDECK	-1.3142	-4.5973	1.969	0.4327	0.2199
	MI	-1.4272	-4.2343	1.3799	0.3155	0.2142
Productive Cough	LOCF	-2.8326	-7.9155	2.2503	0.2747	0.9586
	MEAN	-1.9763	-4.1188	0.1662	**0.0706**	0.3586
	HOTDECK	-2.2931	-5.2745	0.6882	0.1317	0.2192
	MI	-3.0241	-6.2533	0.205	**0.0654**	0.2082
Breathlessness	LOCF	-8.1837	-12.8023	-3.5651	**0.0005**	0.957
	MEAN	-1.2322	-2.6245	0.1601	**0.0828**	0.3596
	HOTDECK	-2.5673	-4.4761	-0.6586	**0.0084**	0.2122
	MI	-3.1746	-4.888	-1.4611	**0.0012**	0.1953
C. *pneumoniae* at Baseline & Follow-up	LOCF	-1.1745	-13.469	11.1199	0.8515	0.9592
	MEAN	-2.6934	-6.8146	1.4278	0.2002	0.3601
	HOTDECK	-1.4703	-7.8139	4.8733	0.6496	0.2201
	MI	-1.9461	-6.2069	2.3146	0.3688	0.215

Table 3.5.3 FEV_1 Measured Over Time with Smoking by Imputed Data

Variable	Method	Estimate	95% CI		P-Value	ρ
			LCL	UCL		
Age	LOCF	-0.0175	-0.0244	-0.0106	**<0.0001**	0.9544
	MEAN	-0.0074	-0.0110	-0.0038	**<0.0001**	0.2670
	HOTDECK	-0.0167	-0.0209	-0.0124	**<0.0001**	0.1948
	MI	-0.0113	-0.0157	-0.0070	**<0.0001**	0.1485
Gender	LOCF	0.1678	-0.0155	0.3511	**0.0728**	
	MEAN	0.0870	0.0116	0.1623	**0.0237**	
	HOTDECK	0.2960	0.1872	0.4047	**<0.0001**	
	MI	0.0971	-0.0051	0.1993	**0.0622**	
Height	LOCF	0.0242	0.0143	0.0340	**<0.0001**	
	MEAN	0.0103	0.0057	0.0149	**<0.0001**	
	HOTDECK	0.0168	0.0108	0.0229	**<0.0001**	
	MI	0.0097	0.0040	0.0154	**0.0011**	
Current Smoking	LOCF	0.0240	-0.0988	0.1469	0.7015	
	MEAN	0.0258	-0.0386	0.0903	0.4316	
	HOTDECK	0.0242	-0.0469	0.0953	0.5042	
	MI	0.0510	-0.0395	0.1414	0.2640	
Breathlessness	LOCF	-0.2256	-0.3477	-0.1034	**0.0003**	
	MEAN	-0.0325	-0.0681	0.0032	**0.0741**	
	HOTDECK	-0.0553	-0.1088	-0.0019	**0.0423**	
	MI	-0.1338	-0.2243	-0.0434	**0.0075**	

Table 3.5.4 FEV_1PP Measured Over Time with Smoking by Imputed Data

Variable	Method	Estimate	95% CI		P-Value	ρ
			LCL	UCL		
Current Smoking	LOCF	4.1442	-0.0577	8.3460	**0.0532**	0.9570
	MEAN	1.7658	-0.0978	3.6294	**0.0633**	0.3568
	HOTDECK	2.3176	-0.2172	4.8524	**0.0731**	0.2102
	MI	1.9832	-0.5636	4.5300	0.1236	0.1920
Breathlessness	LOCF	-8.0106	-12.6550	-3.3661	**0.0007**	
	MEAN	-1.2901	-2.6277	0.0476	**0.0587**	
	HOTDECK	-2.3431	-4.2709	-0.4152	**0.0172**	
	MI	-3.2172	-4.9321	-1.5023	**0.0011**	

Table 3.5.5 FEV_1 Measured Over Time with C. *pneumoniae* – Imputation

Variable	Method	Estimate	95% CI		P-Value	ρ
			LCL	UCL		
Age	LOCF	-0.0175	-0.0244	-0.0105	**<0.0001**	0.9545
	MEAN	-0.0073	-0.0110	-0.0036	**<0.0001**	0.2629
	HOTDECK	-0.0166	-0.0209	-0.0124	**<0.0001**	0.1947
	MI	-0.0113	-0.0156	-0.0070	**<0.0001**	0.1459
Gender	LOCF	0.1662	-0.0151	0.3475	**0.0724**	
	MEAN	0.0825	0.0058	0.1591	**0.0349**	
	HOTDECK	0.2949	0.1867	0.4030	**<0.0001**	
	MI	0.0923	-0.0100	0.1946	**0.0764**	
Height	LOCF	0.0242	0.0143	0.0340	**<0.0001**	
	MEAN	0.0103	0.0057	0.0150	**<0.0001**	
	HOTDECK	0.0168	0.0108	0.0229	**<0.0001**	
	MI	0.0098	0.0041	0.0154	**0.0009**	
Current Smoking	LOCF	0.0234	-0.0998	0.1465	0.7100	
	MEAN	0.0243	-0.0406	0.0893	0.4631	
	HOTDECK	0.0237	-0.0476	0.0950	0.5148	
	MI	0.0494	-0.0404	0.1392	0.2753	
Breathlessness	LOCF	-0.2274	-0.3483	-0.1066	**0.0002**	
	MEAN	-0.0346	-0.0713	0.0021	**0.0648**	
	HOTDECK	-0.0562	-0.1094	-0.0030	**0.0384**	
	MI	-0.1352	-0.2257	-0.0446	**0.0071**	
C. *pneumoniae* at Baseline & Follow-up	LOCF	-0.0404	-0.3154	0.2346	0.7732	
	MEAN	-0.1041	-0.2196	0.0114	**0.0772**	
	HOTDECK	-0.0254	-0.1808	0.1299	0.7482	
	MI	-0.1014	-0.2519	0.0491	0.1837	

Table 3.5.6 FEV_1PP Measured Over Time with C. *pneumoniae* – Imputation

Variable	Method	Estimate	95% CI		P-Value	ρ
			LCL	UCL		
Current Smoking	LOCF	4.0968	-0.1509	8.3445	**0.0587**	0.9571
	MEAN	1.7246	-0.1472	3.5964	**0.0710**	0.3540
	HOTDECK	2.2723	-0.2799	4.8246	**0.0810**	0.2094
	MI	1.9467	-0.5885	4.4819	0.1288	0.1903
Breathlessness	LOCF	-8.1418	-12.7449	-3.5387	**0.0005**	
	MEAN	-1.3490	-2.7250	0.0269	**0.0547**	
	HOTDECK	-2.4134	-4.3330	-0.4938	**0.0137**	
	MI	-3.2502	-4.9675	-1.5329	**0.0010**	
C. *pneumoniae* at Baseline & Follow-up	LOCF	-2.6916	-12.7748	7.3916	0.6008	
	MEAN	-2.8085	-6.4227	0.8058	0.1278	
	HOTDECK	-2.0007	-7.4127	3.4113	0.4687	
	MI	-2.5074	-6.8065	1.7918	0.2502	

Table 4.1.1 Comparisons of Models in Slope Change in FEV_1

Model	Models (R^2)		
	WLS	MLR	ROBUST
Slope Change in FEV_1 with Smoking	0.1880	0.0467	0.0288
Slope Change in FEV_1 with C. *pneumoniae*	0.2067	0.0471	0.0278
Slope Change in FEV_1 with MMP-9	**0.4450**	**0.1096**	**0.0877**
Slope Change in FEV_1 with RATIO of MMP9 to TIMP1	0.4569	0.1145	0.0894
Slope Change in FEV_1 with CRP	0.2694	0.0619	0.0493
Slope Change in FEV_1 with Adenovirus or Herpesviriade	0.2047	0.0771	0.1132

Table 4.1.2 Comparisons of Models in Lung Function Measured Over Time

Model	Log Likelihood		AIC	
	GEE	MIXED	GEE	MIXED
FEV_1 Measured Over Time with Smoking	-440.1966	-16.3	890.4	38.6
FEV_1 Measured Over Time with C. *pneumoniae*	-438.1417	-17.4	886.28	40.8
FEV_1 Measured Over Time with MMP-9	-233.2317	24.3	476.46	-42.6
FEV_1 Measured Over Time with Ratio of MMP-9 to TIMP-1	-233.6005	23.9	477.2	-41.8
FEV_1 Measured Over Time with CRP	-225.4489	25.8	450.90	-45.6
FEV_1 Measured Over Time with Adenovirus or Herpesviriade	-226.4738	25.3	452.94	-44.5

Table 4.2.1 Comparisons of Imputation Methods Using GEE Model

Model	Imputation Methods (Deviance/DF)				
	LOCF	MEAN	HOTDECK	MI	NOIMPU
FEV_1 Measured Over Time with Smoking	0.1918	0.1199	0.2287	0.2972	0.1959
FEV_1 Measured Over Time with C. *pneumoniae*	0.1919	0.1193	0.2288	0.2974	0.1951
FEV_1PP Measured Over Time with Smoking	235.3840	102.6993	246.8103	240.6851	231.4589
FEV_1PP Measured Over Time with C. *pneumoniae*	235.2488	102.3076	246.7114	240.6729	229.5970

Table 5.1 Summarized Conclusions of Primary Analysis

Relationship	Variable	Baseline (MLR)		Lung Function over Time(GEE)		Slope Change in FEV1 (WLS)	
		Estimate	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**
FEV_1 with Smoking	Age	-0.0164	**<0.0001**	-0.017	**<0.0001**	-0.0022	0.1858
	Gender	0.1541	**0.0799**	0.1632	**0.078**	0.1325	**0.0007**
	Height	0.0242	**<0.0001**	0.0246	**<0.0001**	-0.0092	**0.0002**
	Current Smoking	0.0675	0.3666	0.024	0.7055	-0.0869	**0.0203**
	Breathlessness	-0.0264	0.3935	-0.1271	**0.0075**	0.0125	0.3689
FEV_1 with C.*pneumoniae*	Age	-0.0167	**<0.0001**	-0.017	**<0.0001**	-0.0022	0.1866
	Gender	0.1593	**0.0707**	0.1614	**0.0784**	0.1339	**0.0006**
	Height	0.0241	**<0.0001**	0.0247	**<0.0001**	-0.0088	**0.0004**
	Current Smoking	0.0616	0.4108	0.0229	0.7196	-0.0898	**0.016**
	Breathlessness	-0.0247	0.4246	-0.1286	**0.0064**	0.0134	0.3300
	C. *pneumoniae*	-0.1855	0.2856	-0.0449	0.7339	-0.0584	0.1002

Table 5.2 Summarized Conclusions of Subgroup Analysis

Relationship	Variable	Baseline (MLR)		Lung Function Over Time (GEE)		Slope Change in FEV1 (WLS)	
		Estimate	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**
FEV_1 with MMP-9	Age	-0.012	**0.0379**	-0.0132	**0.0312**	-0.0022	0.1407
	Gender	0.1205	0.3661	0.1552	0.2907	0.0855	**0.0461**
	Height	0.0311	**<0.0001**	0.029	**0.0004**	-0.0059	**0.0167**
	Current Smoking	0.0651	0.5362	0.0624	0.5196	-0.0404	0.2191
	Breathlessness	0.0033	0.9612	0.0019	0.9751	-0.0095	0.4244
	LOG MMP-9	-0.0775	0.5983	-0.1802	0.1392	-0.1814	**<0.0001**
FEV_1 with CRP	Age	-0.0145	**0.0184**	-0.0162	**0.0104**	-0.0002	0.9212
	Gender	0.1671	0.2194	0.2125	0.1598	-0.0252	0.5794
	Height	0.028	**0.0005**	0.0255	**0.004**	-0.0004	0.8827
	Current Smoking	0.0459	0.6578	0.0293	0.746	-0.101	**0.0065**
	Breathlessness	0.0137	0.8412	0.014	0.8123	0.0152	0.2472
	LOG CRP	-0.1659	0.1071	-0.2242	0.0089	0.074	0.0145

Table 5.3 Summarized Conclusions of Sensitivity Analysis

Relationship	Variable	LOCF		Mean Estimation		Hot-deck		Multiple Imputation		Original Data	
		Estimate	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**	**Estimate**	**P-value**
$FEV_1$with Smoking	Age	-0.0175	**<0.0001**	-0.0074	**<0.0001**	-0.0167	**<0.0001**	-0.0113	**<.0001**	-0.0170	**<0.0001**
	Gender	0.1678	**0.0728**	0.0870	**0.0237**	0.2960	**<0.0001**	0.0971	**0.0622**	0.1632	**0.078**
	Height	0.0242	**<0.0001**	0.0103	**<0.0001**	0.0168	**<0.0001**	0.0097	**0.0011**	0.0246	**<0.0001**
	Current Smoking	0.0240	0.7015	0.0258	0.4316	0.0242	0.5042	0.0510	0.2640	0.0240	0.7055
	Breathlessness	-0.2256	**0.0003**	-0.0325	**0.0741**	-0.0553	**0.0423**	-0.1338	**0.0075**	-0.1271	**0.0075**
FEV_1 with *C. pneumoniae*	Age	-0.0175	**<0.0001**	-0.0073	**0.0001**	-0.0166	**<0.0001**	-0.0113	**<.0001**	-0.0170	**<0.0001**
	Gender	0.1662	**0.0724**	0.0825	**0.0349**	0.2949	**<0.0001**	0.0923	**0.0764**	0.1614	**0.0784**
	Height	0.0242	**<0.0001**	0.0103	**<0.0001**	0.0168	**<0.0001**	0.0098	**0.0009**	0.0247	**<0.0001**
	Current Smoking	0.0234	0.7100	0.0243	0.4631	0.0237	0.5148	0.0494	0.2753	0.0229	0.7196
	Breathlessness	-0.2274	**0.0002**	-0.0346	**0.0648**	-0.0562	**0.0384**	-0.1352	**0.0071**	-0.1286	**0.0064**
	C. pneumoniae	-0.0404	0.7732	-0.1041	**0.0772**	-0.0254	0.7482	-0.1014	0.1837	-0.0449	0.7339

6.3 Related Codes

6.3.1 Primary Analysis

```
/* Weighted Least Squares Model */

 Proc rag data=primary;

Model slope = age gender height cur's breath / club p r;
Weight weights;
Odds output Parameter Estimates=regest1;
Output     out=reside
           p=prod
           r=reship;
Run;


/* Q-Q Plot */
odds PDF file="H:/pdf_wls_sm.pdf";
proc univariate data=resid noprint;
qqplot reschi/normal(mu=est sigma=est color=red) noframe;
title'Slope Change in FEV1 with Smoking -WLS';
run;
ods pdf close;

/* Select Outputs*/
data f1;
set regest1;
where Variable NE 'Intercept';
keep Variable Estimate Stderr LowerCL UpperCL varianceInflation
     Probt ;
drop Label;
run;


/* Multiple Linear Regression */
proc reg data=primary;
model slope = age gender height cur_sm breath / clb vif p r;
ods output ParameterEstimates=regest2;
output     out=resid
           p=pred
           r=reschi;
run;

ods pdf file="H:/pdf_mlr_sm.pdf";
proc univariate data=resid noprint;
qqplot reschi/normal(mu=est sigma=est color=red) noframe;
title'Slope Change in FEV1 with Smoking -MLR';
run;
ods pdf close;

data f2;
```

```
set regest2;
where Variable NE 'Intercept';
keep Variable Estimate Stderr LowerCL UpperCL varianceInflation
     Probt ;
drop Label;
run;


/* Robust Linear Regression */
proc robustreg data=primary method=MM;
model slope = age gender height cur_sm breath;
output out=robout
         r=resid
         p=pred;
ods output ParameterEstimates = robest3;
run;

ods pdf file="H:/pdf_robust_sm.pdf";
proc univariate data=robout noprint;
qqplot resid/normal(mu=est sigma=est color=red) noframe;
title'Slope Change in FEV1 with Smoking -Robust';
run;
ods pdf close;

data f3;
set robest3;
where Parameter NE 'Intercept';
keep Parameter Estimate Stderr LowerCL UpperCL ProbChisq;
drop Label;
run;


/* OUTPUTS */
ods rtf  file='slope change in Fev1' startpage=no;
title 'Final linear model with Fev1';
proc print data=f1;
run;
proc print data=f2;
run;
proc print data=f3;
run;
ods rtf close;
```

6.3.2 Secondary & Subgroup Analysis

```
/* GEE Model */
proc genmod data=secondary;
class pat_id ;
model  fev1 = age gender height cur_sm breath;
repeated  subject=pat_id /corrw type=ar(1);
```

```
 output        out=resid
               pred=geepred
               reschi=geereschi;
 ods output    ModelFit=geeFit
               GEEEmpPEst=geeEPEst
               GEEWCorr=geeWCorr;
 run;

/* Select Outputs*/
data final1_1;
set geeEPEst;
where Parm NE 'Intercept';
keep Parm Estimate Stderr LowerCL UpperCL Probz;
run;

data final1_2;
set geeFit ;
where Criterion EQ 'Log Likelihood';
keep Value;
run;

data final1_3;
set geeWCorr;
where RowName EQ 'Row1';
keep Col2;
run;

data final1 ;
merge final1_1 final1_2 final1_3;
run;

data final_1;
set final1(RENAME=(Parm=Variable Value=loglihood col2 = Wcorr
          LowerCL=LCL UpperCL = UCL ProbZ=Pvalue));
run;

/* Q-Q Plot */
 ods pdf file="H:/pdf_gee_sm.pdf";
proc univariate data=resid;
qqplot geereschi;
title'QQPLOT: Final GEE Model for Secondary: Fev1 vs Smoking';
run;
ods pdf close;


/* MIXED MODEL */
proc mixed data=secondary covtest noclprint NOITPRINT  NOINFO ;
class pat_id visit;
model fev1 =  age gender height cur_sm breath /solution cl
      influence;
random pat_id;
repeated visit / subject=pat_id type=AR(1) RCorr;
ods output     RCorr=mixrcorr
               FitStatistics = mixFit
```

```
                SolutionF = mixSolutionF
                Influence = residual;
run;

ods pdf file="H:/pdf_mixed_sm.pdf";
proc univariate data=mixresid noprint;
qqplot residual/normal(mu=est sigma=est color=red) noframe;
title 'FEV with Smoking - Mixed';
run;
ods pdf close;

data final2_1;
set mixSolutionF;
where Effect NE 'Intercept';
keep Effect Estimate StdErr Lower Upper Probt ;
run;
data final2_2;
set mixFit ;
where Descr EQ '-2 Res Log Likelihood';
keep Value;
run;

data final2_3;
set mixrcorr;
where Row EQ 1;
keep Col2;
run;

data final2;
merge final2_1 final2_2 final2_3;
run;

data final_2;
set final2(RENAME=(Effect=Variable Lower=LCL Upper=UCL Probt=Pvalue
Value=NRloglihood col2 = Wcorr));
run;

ods rtf  file='Final_Secondary_gee_mixed.rtf' startpage=no;
title 'Final GEE and MIXED model for Secondary: Fev1 vs smoking';
proc print data=final_1;
run;
proc print data=final_2;
run;
ods rtf close;
```

6.3.3 Sensitivity Analysis

- **Imputation Methods**

```
/* Group Mean Estimation (SAS CODE)*/
proc sort data=primary;
by visit;
```

```
run;

proc standard data=primary out=primary_out replace;
by visit;
var flu_shot flu_num pneu_5y pneu_sho sput_amt breath fev1 fev1pp;
run;

proc sort data=primary_out out=primary_mean;
by pat_id;
run;
```

/* LOCF (S-PLUS/R CODE) */

```
locf<-function (x)
{
   n<-nrow(x)
    pat<-matrix(c(x[,1]),ncol=1)
   years<-matrix(c(x[,2]),ncol=1)
   visit<-matrix(c(x[,3]),ncol=1)
   fev1pp<-matrix(c(x[,4]),ncol=1)
   fev1<-matrix(c(x[,5]),ncol=1)

   for (i in 2:n)
   {
     if(pat[i]==pat[i-1])
    {
      if(is.na(years[i]))
      {
        years[i]= years[i-1]
        fev1[i] = fev1[i-1]
        fev1pp[i] = fev1pp[i-1]
        visit[i]= visit[i-1]+1
        }
      }
    }
    y<-cbind(pat, years,visit, fev1, fev1pp)
    y
}
dat<-read.csv("H:\\data \\locf.csv")
y<-locf(dat)
write.csv(y, "H:\\data \\primary_locf.csv")
```

/* Multiple Imputation (SAS CODE)*/

```
proc sort data=primary out=prim_sort;
by pat_id date;
run;

proc mi data=prim_sort seed=21355417 nimpute=10 out=primary_mi;
 mcmc chain=multiple displayinit initial=em(itprint);
var  flu_shot flu_num pneu_sho sput_amt breath fev1 fev1pp ;
run;
```

```
proc sort data=primary_mi;
by pat_id _imputation_ visit;
run;
```

```
/* Hot-deck Method (STATA 9.1) */

/* Impute Missing Data in Outcomes */
insheet using "H:\data \primary.csv"
hotdeck fev1 fev1pp using imp1, store by (gage gender gheight)
       impute(1) keep(pat_id visit)

/* Impute Missing Data in Predictors */
insheet using "H:\data\predictor.csv"
hotdeck breath using imp2, store by (gage cur_sm quit_yrs)
        impute (1) keep(pat_id)

/* Transfer .dta file to .xls file */
use "C:\Program Files\stata 9\imp1.dta"
outsheet pat_id visit fev1 fev1pp using hotdeck1.xls, replace

use "C:\Program Files\stata 9\imp2.dta"
outsheet pat_id breath using hotdeck2.xls, replace
```

•. GEE Models Using Multiple Imputation Data Sets

```
/* GEE MODEL */
 proc genmod data=mimcmc;
 class pat_id ;
 model fev1= age gender height cur_sm breath;
 repeated subject=pat_id / corrw covb type=ar(1);
 by _imputation_;
 output          out=resid
                 pred=geepred
                 resraw=geeresraw
                 reschi=geereschi;
 ods output      ParameterEstimates=geeparms
                 ParmInfo=geepinfo
                 GEENCov=geencov
                 GEEWCorr=geewcorr
                 ModelFit=geeFit  ;
run;

proc mianalyze parms=geeparms COVB=geencov parminfo=geepinfo;
modeleffects Intercept age gender height cur_sm breath ;
ods output ParameterEstimates = geefinalparms;
run;
```

Bibliography

1. Barnes PJ. Chronic obstructive pulmonary disease. New England Journal of Medicine 2000; 343(4): 269-80
2. Murray CJ, Lopez AD. Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study. Lancet 1997; 349(9064):1498-504.
3. http://www.lung.ca/CCA/refterms.html CCA Terms of Reference. February 1, 2006
4. Mannino DM, Buist AS, Petty TL, Enright PL, Redd SC. Lung function and mortality in the United States: data from the First National Health and Nutrition Examination Survey follow up study. Thorax 2003; 58(5): 388-93.
5. http://www.on.lung.ca/lungdisease.html. The Ontario Lung Association, last accessed 2002-05-31. Your Lungs: What is Chronic Obstructive Pulmonary Disease? 2002.
6. Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease. American Journal of Respiratory and Critical Care Medicine 1996; 154: 701-6.
7. Lung Facts 1994 Update. Canadian Lung Association 1993.
8. Sethi JM, Rochester CL. Smoking and chronic obstructive pulmonary disease. Clinical Chest Medicine 2000; 21(1) 67-86, viii.
9. http://www.priory.com/cmol/causesof.htm The causes of COPD and who is at risk? February, 2006.
10. Von Hertzen L, Kaprio J, Koskenvuo M, Isoaho R, Saikku P. Humoral immune response to Chlamydia pneumoniae in twin discordant for smoking. J Intern Med 1998; 244(3) 227-34.
11. Hahn DL. Chlamydia pneumoniae, asthma, and COPD: what is the evidence? Ann Allergy Asthma Immunol 1999; 83(4): 271-88, 291.
12. Blasi F, Legnani D, Lombardo VM, Negretto GG, et al. Chlamydia pneumoniae infection in acute exacerbations of COPD. European Respiratory Journal 1993;

6(1): 19-22.

13. 13. Von Hertzen L. Chlamydia pneumoniae and its role in chronic obstructive pulmonary disease. Ann Med. 1998; 30(1): 27-37.
14. Smieja M, Leigh R, Petrich A, Chong S, Kamada D, Hargreave FE et al. Smoking, season, and detection of Chlamydia pneumonia DNA in clinically stable COPD patients. BMC Infection Diseases 2002; 2(1):12.
15. Aikawa M, Rabkin E, Sugiyama S, Voglic SJ, Fukumoto Y, Furukawa Y et al. An HMGCoA reductase inhibitor, cerivastatin, suppresses growth of macrophages expressing matrix metalloproteinases and tissue factor in vivo and in vitro. Circulation 2001; 103(2): 276-83.
16. Danesh J, Wheeler JG, Hirschfield GM, Eda S, Eiriksdottir G, Rumley A. C-reative protein and other circulating markers of inflammation in the prediction of coronary heart disease. New England Journal of Medicine 2004; 350(14): 1387-97.
17. Blankenberg S, Rupprehct HJ, Poirier O, Bickel C, Smieja M, Hafner G. Plasma concentrations and genetic variation of matrix metalloproteinase 9 and prognosis of patients with cardiovascular disease. Circulation 2003; 107(12):1579-85.
18. Donner A. Sample size requirements for stratified cluster randomization design. Statistics in Medicine 1992; 11: 743-50.
19. Babyak MA. What you see may not be what you get: a brief, non-technical introduction to overfitting in regression-type models. Psychosomatic Medicine 2004; 66(3): 411-21.
20. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology 1996; 49(12): 1373-9.
21. McCullagh P, Nelder J.A. Generalized Linear Models, Second Edition; 1989. Chapman & Hall, London.
22. McCullagh P. Quasi-likelihood functions. Annals of Statistics 1983; 11: 59-67.
23. Liang K.Y, Zeger L. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 44: 121-30.
24. Liang K.Y. & Zeger L. Models for Longitudinal Data: A Generalized Estimating

Equation Approach. Biometrics 1988; 44 (4): 1049- 60.

25. James H, Joseph H. Generalized Linear Models and Extensions. Stata Press, 2001.
26. Breslow, N.E., Clayton D.G. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 1993; 88: 9-25.
27. Qian G, Kunsch H. On model selection in robust linear regression. Journal of Statistical Planning & Inference. In Press, 1998.
28. Colin C. Robust regression and outlier detection with the ROBUSTREG procedure. SAS Institute Inc. SUGI 27: 265-27.
29. http://www.stata.com/support/faqs/stat/mle_vs_gmm.html Random-effect Model 2006.
30. Horton NJ. Goodness-of-fit for GEE: an example with mental health service utilization. Statistics in Medicine 1999; 18: 213-22.
31. http://supportsas.com/onlinedoc/913/docMainpage.jsp Mixed Model Theory, 2006.
32. Roderick JA, Donald BR. Statistical analysis with missing data. Second Edition, 2002.
33. http://supportsas.com/onlinedoc/913/docMainpage.jsp. Multiple Imputation, 2006
34. Rubin D.B. Multiple Imputation for Non-response in Surveys. J. Wiley & Sons, New York , 1987.
35. Lening Z., Richard H.J., James R. M Comparison of Models for Longitudinal Data with Unequally Spaced Binary Outcomes (Abstract).