

AN EVALUATION OF TWO EXISTING METHODS FOR ANALYZING  
LONGITUDINAL RESPIRATORY SYMPTOM DATA

by

VICTORIA HELEN ARRANDALE

B.Sc., Simon Fraser University, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Occupational & Environmental Hygiene)

THE UNIVERSITY OF BRITISH COLUMBIA

December 2006

© Victoria Helen Arrandale 2006

## **Abstract**

Due to the complexities of analyzing repeated binary outcomes, changes in respiratory symptoms over time are rarely studied. In fact, most respiratory epidemiology studies to date have not taken full advantage of longitudinal symptom data.

This thesis evaluated discrete mixture models (SAS® Proc Traj) and generalized linear mixed models (SAS® Proc Glimmix) with respect for their applicability to six basic respiratory symptom research questions. These methods are both capable of handling repeated binary outcome data and permit inclusion of time varying covariates. These two techniques were then applied in a case study.

Results from the evaluation of the methods indicated that Proc Glimmix can model the predictors of respiratory symptoms as well as population trends in symptom reporting over time. But Proc Glimmix is not suitable for modeling pattern or shape of change over time. In contrast, Proc Traj models patterns of change over time, and identifies multiple subgroups within the population. Proc Traj is not capable of modeling overall population trends. Both methods have statistical limitations that researchers need to understand; to help with this a simple guide describing both techniques was compiled.

The case study utilized longitudinal data from a population of marine workers and focused on the outcome breathlessness, or dyspnea. Results from both Proc Traj and Proc Glimmix models indicated that the probability of reporting dyspnea changed over time in this population. Proc Traj models identified two distinct patterns of change in the population (one increasing over time, one steady over time). Proc Glimmix models identified several factors that were associated with dyspnea reporting; older age, childhood asthma, smoking and being female were associated with more dyspnea, whereas better lung function and current exposure to respiratory irritants were associated with less dyspnea.

The overall conclusion was that both Proc Traj and Proc Glimmix models are suitable for analyzing repeated binary respiratory symptom data and researchers are encouraged to consider their use. Proc Glimmix is best for modeling the predictors of reporting a symptom at the population level, while Proc Traj is suited for modeling multiple subgroups in the population and their patterns of change over time.

# Table of Contents

Abstract.....	ii
Table of Contents .....	iii
List of Tables .....	vi
List of Figures.....	viii
Acknowledgments .....	ix
Dedication .....	x
<b>1 Introduction .....</b>	<b>1</b>
<b>2 Background &amp; Literature Review.....</b>	<b>2</b>
2.1 Respiratory Symptoms.....	2
2.2 Literature Review: Longitudinal Studies of Respiratory Symptoms.....	3
2.2.1 Symptoms as a Predictor Variable .....	8
2.2.2 Symptoms as an Outcome Variable .....	14
2.2.3 Strengths & Limitations of Previous Studies .....	17
2.3 Literature Review: Statistical Methods for Longitudinal Data Analysis .....	20
2.3.1 SAS® Proc Traj .....	21
2.3.2 SAS® Proc Glimmix .....	23
<b>3 Results I: Evaluation of Proc Traj and Proc Glimmix for use in Respiratory Epidemiology.....</b>	<b>25</b>
3.1 Introduction .....	25
3.2 Overview of the Methods .....	26
3.2.1 SAS® Proc Glimmix: Generalized Linear Mixed Models .....	26
3.2.2 SAS® Proc Traj .....	28
3.3 Evaluation of Generalized Linear Mixed Models and Proc Traj Models .....	32
3.3.1 What Factors Predict Reporting a Respiratory Symptom?.....	33
3.3.2 Do Respiratory Symptoms Change Over Time?.....	34
3.3.3 What are the Patterns of Change in Respiratory Symptoms Over Time? .....	35
3.3.4 What Factors Predict Respiratory Symptom Change Over Time?.....	36
3.3.5 Do Different Respiratory Symptoms Change with Similar Patterns Over Time?.....	38
3.3.6 How Does a Time-varying Covariate Affect the Pattern of Respiratory Symptom Reporting?.....	40
3.4 Summary .....	40
<b>4 Results II: Case Study .....</b>	<b>45</b>
4.1 Introduction .....	45
4.2 Study Population.....	46
4.3 Methods: Data Collection .....	47
4.3.1 ATS Questionnaire.....	47
4.3.2 Spirometry .....	47
4.3.3 Skin Prick Testing.....	47
4.3.4 Previous Analysis.....	48

<b>4.4 Methods: Current Analysis.....</b>	<b>48</b>
4.4.1 Study Population.....	48
4.4.2 Outcome.....	49
4.4.3 Definitions .....	49
4.4.4 Descriptive Analysis .....	50
4.4.5 Traditional Fixed Effects - Proc Logistic.....	52
4.4.6 Proc Glimmix.....	52
4.4.7 Proc Traj .....	53
<b>4.5 Results.....</b>	<b>54</b>
4.5.1 Descriptive Analysis .....	54
4.5.2 Traditional Fixed Effects - Proc Logistic.....	60
4.5.3 Proc Glimmix.....	64
4.5.4 Proc Traj .....	70
<b>4.6 Discussion .....</b>	<b>77</b>
4.6.1 Traditional Fixed Effects- Proc Logistic.....	78
4.6.2 Proc Glimmix.....	78
4.6.3 Proc Traj .....	80
4.6.4 Relevance to Previous Literature .....	82
<b>4.7 Conclusions .....</b>	<b>83</b>
<b>5 Perspectives .....</b>	<b>84</b>
<b>5.1 Strengths.....</b>	<b>86</b>
<b>5.2 Limitations .....</b>	<b>87</b>
<b>5.3 Future Research.....</b>	<b>88</b>
<b>References.....</b>	<b>89</b>
<b>Appendix A. How to use SAS® Proc Traj and SAS® Proc Glimmix in Respiratory Epidemiology .....</b>	<b>94</b>
<b>A.1 Introduction.....</b>	<b>94</b>
<b>A.2 Goal.....</b>	<b>94</b>
<b>A.3 How to use this document .....</b>	<b>94</b>
<b>A.4 SAS® Trajectory Procedure .....</b>	<b>95</b>
A.4.1 Overview.....	95
A.4.2 Requirements .....	95
A.4.3 Data Organization .....	96
A.4.4 Dummy Variables .....	97
A.4.5 Missing Data .....	97
A.4.6 Types of Research Questions .....	97
A.4.7 Syntax .....	97
A.4.8 Selecting the Best Model .....	99
A.4.9 Output .....	102
A.4.10 User Information .....	103
A.4.11 Cautions .....	103
A.4.12 Reference Texts.....	104
<b>A.5 SAS® Glimmix Procedure.....</b>	<b>105</b>
A.5.1 Overview.....	105
A.5.2 When to Use Mixed Effects .....	105
A.5.3 Requirements .....	106

A.5.4	Data Organization .....	106
A.5.5	Dummy Variables .....	107
A.5.6	Missing Data .....	107
A.5.7	Types of Research Questions .....	107
A.5.8	Syntax .....	108
A.5.9	Selecting the Best Model .....	109
A.5.10	Output .....	110
A.5.11	User Information .....	110
A.5.12	Cautions .....	111
A.5.13	Reference Texts.....	111
<b>Appendix B.</b>	<b>Cross-tabulations of Personal Risk Factors for Dyspnea .....</b>	<b>112</b>

## List of Tables

<b>Table 1</b> Summary of previous literature examining respiratory symptoms over time.....	4
<b>Table 2</b> Range of respiratory symptom pattern frequencies in previous studies .....	8
<b>Table 3</b> Summary of findings on the utility of Proc Glimmix and Proc Traj for each outlined research question .....	42
<b>Table 4</b> Summary of variables considered as risk factors for reporting dyspnea .....	49
<b>Table 5</b> Summary of the categories describing dyspnea change over time .....	51
<b>Table 6</b> Demographics of the entire marine transportation workers cohort .....	55
<b>Table 7</b> Demographics of the subset used in respiratory symptom analyses.....	55
<b>Table 8</b> Crude prevalence rate for dyspnea by visits year, stratified by sex.....	56
<b>Table 9</b> Association between dyspnea and known indicators of respiratory disease.....	57
<b>Table 10</b> Association between dyspnea and risk factors for reporting dyspnea .....	58
<b>Table 11</b> Means (SD) for continuous risk factors for dyspnea, using measures from subjects' first visit.....	58
<b>Table 12</b> Prevalence of dyspnea change categories .....	59
<b>Table 13</b> FEV1 annual change stratified by dyspnea change over time .....	60
<b>Table 14</b> Initial model for predictors of dyspnea using only first visit responses. Results from fixed effects logistic regression.....	61
<b>Table 15</b> Final model for predictors of dyspnea using only first visit responses. Results from fixed effects logistic regression.....	61
<b>Table 16</b> Final fixed logistic regression model, adjusted for lung function, using only first visit responses .....	62
<b>Table 17</b> Initial model for predictor of dyspnea using all visits, and all subjects. Results from fixed effects logistic regression.....	63
<b>Table 18</b> Final model for predictor of dyspnea using all visits, and all subjects. Results from fixed effects logistic regression.....	63
<b>Table 19</b> Final fixed effects logistic regression, adjusted for lung function, using all data.....	64
<b>Table 20</b> Initial mixed model using all data (n=2472) to determine predictors of reporting dyspnea at any point in time .....	65
<b>Table 21</b> Final mixed model using all data (n=2472) to determine predictors of reporting dyspnea at any point in time .....	65
<b>Table 22</b> Final mixed model describing predictors of dyspnea, adjusted for lung function. ....	66

<b>Table 23</b> Model for the effect of visit year on reporting dyspnea including random intercept term, all subjects (n=2472) .....	68
<b>Table 24</b> Model for the effect of visit year on reporting dyspnea including random intercept term, Men (n=2221).....	68
<b>Table 25</b> Model for the effect of visit year on reporting dyspnea including random intercept term, Women (n=251) .....	68
<b>Table 26</b> Saturated model demonstrating the effect of each visit year on reporting dyspnea. ....	69
<b>Table 27</b> Final mixed model using all data (n=2473), risk factors, time covariates and an adjustment for lung function (% predicted FEV1) .....	69
<b>Table 28</b> Model of the effect of calendar year (continuous) on reporting dyspnea, all subjects (n=2472) .....	70
<b>Table 29</b> Model fit statistics for stepwise iterations of Proc Traj model to determine number of groups using all subjects' data (n=925 subjects) .....	71
<b>Table 30</b> Model fit statistics for stepwise iterations of Proc Traj model to determine number of groups only male subjects with complete data.....	71
<b>Table 31</b> Model fit statistics for stepwise iterations of Proc Traj model to determine trajectory shape over time using all subjects' data (n=925 subjects) .....	72
<b>Table 32</b> Model fit statistics for stepwise iterations of Proc Traj model to determine trajectory shape over time using subjects with complete data (n=148 subjects).....	72
<b>Table 33</b> Final two-group model describing the change in the probability of dyspnea over time, using all subjects' data (n=925) .....	72
<b>Table 34</b> Final two-group model describing the change in the probability of dyspnea over time, using subjects with complete data (n=148) .....	73
<b>Table 35</b> Description of subgroups using posterior group assignments from Proc Traj output dataset (entire dataset, n=925) .....	76
<b>Table 36</b> Category of dyspnea change over time stratified by sex and group membership.....	77
<b>Table 37</b> Mock data set up for analysis with Proc Traj .....	96
<b>Table 38</b> Description of variables in mock data (Table 37).....	96
<b>Table 39</b> Interpretation of logged Bayes factor ( $2 \cdot \Delta \text{BIC}$ ) for model selection .....	101
<b>Table 40</b> Interpretation of Bayes Factor ( $e^{\text{BIC}_i - \text{BIC}_j}$ ) for model selection.....	101
<b>Table 41</b> Mock data set up for analysis with Proc Glimmix.....	106
<b>Table 42</b> Description of variables in mock data .....	107
<b>Table 43</b> Cross-tabulations and Chi-square p-values for personal risk factors (Men) .....	113
<b>Table 44</b> Cross-tabulations and Chi-square p-values for personal risk factors (Women) ....	114

## List of Figures

<b>Figure 1</b> Schematic representation of the patterns of symptom change over time and the corresponding categories. ....	51
<b>Figure 2</b> Prevalence of dyspnea for men and women between 1988 and 1999.....	56
<b>Figure 3</b> Estimated correlation matrix (correlation between repeated measures of dyspnea) for men with complete data (n=148).....	67
<b>Figure 4</b> Graphical output from Proc Traj showing final two-group model with group membership probability and shape of change over time using all subjects' data (n=925) .....	74
<b>Figure 5</b> Graphical output from Proc Traj showing final two-group model with group membership probability and shape of change over time using men with complete data (n=148).....	75
<b>Figure 6</b> Output from basic Proc Traj model with no covariates.....	103
<b>Figure 7</b> Sample output from mixed model using Proc Glimmix.....	111



## **Acknowledgments**

First and foremost I would like to acknowledge my thesis supervisor, Dr. Susan Kennedy, for her support throughout this process. Susan was a firm believer in my abilities even when I was not, but was also quick to point out my deficits when I was too naïve to see them. This balance made for a challenging and extremely rewarding learning environment, and also stimulated tremendous personal growth for me over the last two years. I am exceedingly grateful for your guidance and mentorship – Thank you.

I would also like to acknowledge my thesis committee members, Drs. Mieke Koehoorn and Ying MacNab, who were probably more involved in the development and execution of this thesis than either initially envisioned. Together they provided frank feedback and were always open to my endless (sometimes repetitive) questions. Thank you for your patience with me, and your commitment to my work.

Finally, I would like to acknowledge the community that is the School of Occupational and Environmental Hygiene (SOEH). I have shared countless memories with my fellow students, as well as the staff and faculty, over the last two years and I will carry these with me for years to come. I thank you all for your support and encouragement.

*To my Mum, Dad and Sister:*

*Thank-you for your unconditional love & continued support.*

# 1 Introduction

This thesis is motivated by the hypothesis that patterns of respiratory symptom change over time may be useful in predicting (or describing) subsequent pulmonary function deterioration in general populations of working adults and by the fact that, to date, most respiratory epidemiology studies have not taken full advantage of longitudinal symptom data.

The focus of this thesis is an exploration of existing statistical methods for the investigation of respiratory symptoms in longitudinal epidemiologic studies of lung health.

The thesis objectives are:

- To review what analytic approaches have been used to date in longitudinal studies of respiratory symptoms (Chapter 2);
- To evaluate the potential for application of two existing statistical approaches that make full use of the longitudinal nature of respiratory symptom data almost always collected, but seldom used (Chapter 3);
- To apply and compare both ‘traditional’ and ‘newer’ approaches in one case study example, using data from a longitudinal study of marine transportation workers (Chapter 4); and
- To prepare a guidance document to facilitate respiratory epidemiologists using these newer approaches (Appendix A).

## **2 Background & Literature Review**

### **2.1 Respiratory Symptoms**

The American Thoracic Society (ATS) Questionnaire was originally developed in 1978 as part of the Epidemiology Standardization Project (1). The goal of the Epidemiology Standardization Project was to develop standardized criteria for respiratory disease survey questionnaires, tests of pulmonary function and chest radiographs (1). The Epidemiology Standardization Project used both the Medical Research Questionnaire and the National Heart and Lung Institute respiratory questionnaires as the basis for developing a new standardized questionnaire (1). The original ATS Questionnaire (ATS-DLD-78) was born of this process and contained thirty-four questions pertaining to respiratory symptoms (cough, phlegm, wheeze, dyspnea). Since the development of the ATS Questionnaire most studies of respiratory disease have included the questionnaire, or a variation of it, as part of their study protocol.

Symptoms are important because they are what people experience as part of the disease process. Symptoms are also what people report to their physicians or other trusted health practitioner. The fact that respiratory symptoms, especially the changes over time in respiratory symptoms, have not been extensively studied makes them even more enticing as the focus of this thesis work. In terms of occupational disease, symptoms are key to successful occupational surveillance as they are what a worker will report to their physician, even when the worker themselves does not make the connection to their workplace (2, 3). For these reasons, it would benefit occupational health professionals to have further understanding of the longitudinal patterns of respiratory symptoms and their relationship to both workplace exposures as well as disease processes.

Longitudinal studies of respiratory health using the ATS Questionnaire often result in symptom data with the following characteristics:

- a. repeated measures on the same individuals
- b. unequal spacing between repeated measures
- c. dichotomous outcomes (yes/no)
- d. unbalanced data (different number of observations for different individuals in the population)

e. missing data

The main reason researchers have avoided studies of longitudinal respiratory symptoms is due to the challenges involved in the analysis of repeated, correlated dichotomous data. Two methods for handling data of this type are addressed in this thesis; (1) SAS® Trajectory Procedure (Proc Traj), and (2) SAS® Generalized Linear Mixed Model Procedure (Proc Glimmix) (4, 5).

## **2.2 Literature Review: Longitudinal Studies of Respiratory Symptoms**

Studies have used longitudinal symptom data as a predictor of lung function outcomes (6-17) but neither symptoms nor the pattern of symptom change over time have been studied thoroughly as an outcome. The literature examining symptoms at one point in time and their ability to predict lung function, as well as the literature examining cross sectional exposure levels and respiratory symptoms, is extensive and is not reviewed here.

A thorough literature review was completed for articles that studied longitudinal respiratory symptoms; articles were included if they measured respiratory symptoms at multiple time points and used repeated measures (over time) of respiratory symptoms in the analysis. Nineteen articles fitting these requirements were located in the peer-reviewed literature. These articles are summarized in Table 1.

The literature search began with searching for studies of longitudinal respiratory symptoms. PubMed and Web of Science were utilized. The reference lists from located articles were used to identify additional relevant literature; this proved valuable as the keywords that successfully located studies focused on longitudinal respiratory symptoms were varied. Initial search terms included specific symptoms (e.g. cough, phlegm, dyspnea, and wheeze) or diseases (e.g. asthma, chronic obstructive pulmonary disease or COPD) combined with variations on the longitudinal theme (e.g. “over time”, “onset”, “incidence”, “cohort”).

**Table 1 Summary of previous literature examining respiratory symptoms over time**

Lead Author	Year	Study Population	# of Visits		Follow-up time	Outcome	Predictor(s) of Interest	Analysis Method
			Total	# Used				
Symptoms as a Predictor Variable:								
Sharp (18)	1973	1263 white men from an electric company	2	2	7 years	Lung function	Respiratory symptom change over time	Student's t-test, chi square test
Jedrychowski (19)	1988	1747 randomly selected individuals from Cracow, Poland	3	3	13 years	FEV1 decline during follow-up	Pattern of symptom change between first 2 visits	Multiple linear regression
Jaakkola (20)	1993	1044 young white adults, age 15-40 years	3	2	7.7 years (mean)	Rate of lung function change over time	Longitudinal pattern of symptom reporting	Multiple linear regression
Brodkin (21)	1996	446 men in the Seattle Asbestos Lung Cancer Chemoprevention Trail with >3years of follow-up	annual visits	2	2.9-5.2 years	Annual loss of FEV1 and FVC	Longitudinal pattern of symptom reporting	Multiple linear regression
Krzyzanowski (22)	1990	Subjects from the Cracow and Tuscon studies with >2 visits C: 740 Men, 1024 women, T: 266 Men, 374 Women	C: 3 T: 9	3	C: 13 years T: 12 years	Lung function (FEV1, FVC, FEV1/FVC ratio)	Respiratory symptom pattern over time	Multiple linear regression
Krzyzanowski (23)	1992	Subjects from Cracow and Tuscon studies with >1 visits C: 1265 Men, 1818 women. T: 613 Men, 839 Women	C: 3 T: 9	2	C: 13 years T: 12 years	Respiratory symptom pattern between two time points	Age, smoking, gender, city (Tuscon or Cracow)	Log-linear regression and logistic regression
Krzyzanowski (24)	1993	Subjects from the Cracow and Tuscon studies who were smokers at baseline C: 815 Men, 439 women, T: 234 Men, 234 Women	C: 3 T: 9	2	C: 13 years T: 12 years	Incidence and prevalence of respiratory symptoms	City, gender, age, smoking at first visit, age started smoking	Log-linear regression and logistic regression

**Table 1 Cont'd**

Lead Author	Year	Study Population	# of Visits		Follow-up time	Outcome	Predictor(s) of Interest	Analysis Method
			Total	# Used				
Christiani (25)	2001	447 cotton textile workers 472 silk textile workers (controls)	4	4	15 years	Rate of FEV1 and FVC change over time	Consistency of symptom reporting, number of symptoms reported	Marginal model with generalized estimating equations
Wang (26)	2002	240 newly hired female workers at 3 state-owned cotton mills, age 16-29 years)	3	3	1 year	FEV1 at all visits	"Symptomatic" - reporting cough wt phlegm or dry cough at 3 month visit	Marginal model with generalized estimating equations
Sherrill (27)	1993	633 males and 891 females (>55yrs) from the Tuscon Study of Airways Obstructive Disease	6	6	0-10 years	FEV1, FVC, FEV1/FVC ratio from all visits	Respiratory symptoms at each visit	Mixed random effects model
<u>Symptoms as a Binary Outcome Variables:</u>								
Pahwa (28)	1998	1848 asymptomatic male grain elevator workers	5	5	9-15 years	Onset of new wheeze	Lung function, smoking, years in industry	Survival analysis
Carta (11)	1996	1078 Sardinian coal miners	7	7	11 years	Onset of respiratory symptoms	Exposure to coal dust	Logistic regression
Kongerud (7)	1991	1013 aluminum potroom workers	2	2	4 years	Development of respiratory symptoms	Smoking, fluoride exposure	Turnbull algorithm (similar to proportional hazards model)
Boutet (29)	2006	769 apprentices	3	3	5 years	Onset of respiratory symptoms	Airway hyper-responsiveness	Logistic regression
Gunnbjornsdottir (30)	2006	16,190 adults, age 20-44 years at baseline	2	2	5-11 years	Change in respiratory symptoms	Self-reported indoor dampness	Logistic regression

**Table 1 Cont'd**

Lead Author	Year	Study Population	# of Visits		Follow-up time	Outcome	Predictor(s) of Interest	Analysis Method
			Total	# Used				
<u>Symptoms as a Score, Scale or other Outcome Variable:</u>								
Mahler (31)	1995	76 male COPD patients recruited from outpatient clinics	5	5	2 years	Transition dyspnea index (TDI)	Lung function (FEV1, FVC, inspiratory pressure)	ANCOVA
Lareau (32)	1999	34 male subjects with COPD	5	5	5.3 years (mean)	Change in dyspnea score and lung function measures over time		Linear correlation
Hodgev (33)	2004	19 male COPD patients	2	2	at least 2 years	Change in dyspnea score and lung function measures over time		Linear correlation
Wu (34)	2004	764 workers from a steelworks surveillance program in Australia	6	6	4.6 years (mean)	Rate of symptom occurrence	Age, smoking, work location, work duration	Binomial logistic regression



Several hundred abstracts were located and reviewed. The process was challenging because most longitudinal studies reported on symptoms but very few actually used multiple measures of symptoms in the analysis portion of the research. Adding to the challenge was the wide variety of terminology used to describe longitudinal studies of respiratory symptoms mentioned previously.

The majority of excluded literature was longitudinal studies of respiratory disease that actually did measure symptoms at multiple time points, but only used symptoms measured at one time point in the analysis.

Several articles were located that studied respiratory symptoms using daily diaries. These articles were excluded because they were from panel studies and the time series analysis techniques employed are not relevant for the repeated measures distributed over a longer period of time, such as in occupational studies of respiratory health.

Most literature that reported on respiratory symptoms over time used symptom reporting at multiple time points as a predictor of lung function (18-23, 25-27) or investigated the correlation between symptom change and lung function (32, 33). Others attempted to use the pattern of respiratory symptom change as an outcome (7, 11, 28-30). With the exception of three studies (31-33), all other studies used the ATS questionnaire or a modified version of it (or its predecessors) to measure respiratory symptoms.

Among studies using respiratory symptom change as a predictor of lung function, five studies categorized symptom change over time into patterns of change and used these patterns as the predictor variable (18-22). Each of these studies categorized symptom responses at two or three time points into four patterns: never reporting a symptom, always reporting a symptom (persistent), developing a new symptom and resolution (remission) of a symptom. The range of prevalence of these respiratory symptom patterns in previous studies is shown in Table 2.

**Table 2 Range of respiratory symptom pattern frequencies reported in six previous studies\***

Symptom	Category of Change over time	Mean	Min	Max
Cough	Never	73.6	56	84.6
	Resolved (remission)	9.4	4.3	16.5
	New (development)	9.9	5.1	20.2
	Persistent	7.2	2.1	12
Phlegm	Never	70.3	47.5	83.4
	Resolved (remission)	10.2	4.8	16.9
	New (development)	10.8	3.8	21.4
	Persistent	8.8	2.9	20
Wheeze	Never	74.6	50.9	93.9
	Resolved (remission)	7.1	1.7	11
	New (development)	9.8	3.6	19.7
	Persistent	8.5	0.9	22
Dyspnea	Never	74.5	50.4	89.9
	Resolved (remission)	7.0	2.4	14
	New (development)	10.9	3.6	20.7
	Persistent	7.6	1.8	16.3

\*(18-23)

### 2.2.1 Symptoms as a Predictor Variable

An early study by Sharp and colleagues (18) explored the reversibility of respiratory symptoms, the incidence of new respiratory symptoms and the relationships between progression/regression of respiratory symptoms and lung function using data from a cohort of electric company workers (18). Using the four categories described previously (never, persistent, developed new, remission) Sharp et al grouped the workers for each symptom of interest and then related symptom change to categories of lung function change over time (worsened lung function, improved lung function and no change in lung function). Results from this analysis indicated that subjects without respiratory symptoms were more likely to have observable improvements in their lung function at follow-up. At the time of publication (1973) Sharp et al did not have access to the advanced modeling techniques available today and their analysis was constrained by this fact.

Jedrychowski et al (19) published on respiratory symptoms using data from the Cracow prospective study on chronic obstructive lung disease. The Cracow study was a random sample of the population in Cracow, Poland and consisted of three survey visits: 1968, 1973 and 1981.

Jedrychowski et al (19) categorized subjects based on their responses to the Medical Research Questionnaire respiratory symptom responses at the first two survey dates. Using these two responses resulted in the four categories mentioned previously (never, persistent, developed new and remission).

A multiple linear regression analysis using lung function as the outcome and the symptom category as the primary independent variable was completed (19). Two iterations of the analysis were completed, one using the decline in lung function between the first two visits (five years), and one using lung function from the first and third survey (thirteen years).

The results from the analyses including men only, showed that for chronic wheezing, shortness of breath and chronic cough the effect of reporting a persistent symptom on FEV1 decline is equal to smoking 40-50 cigarettes daily (19). This finding was not consistent in the female analysis.

Unlike most other publications in this area, Jedrychowski et al (19) did not use linear regression of the FEV1 change over time as the outcome variable in the analyses, instead they calculated a metric they refer to as the FEV1 decrease index (FDI). The FDI metric does not account for the time period elapsed since the last visit date. The FDI measurement is the absolute change in FEV1 divided by the sum of both FEV1 measurements (from the two dates in question). This metric does not appear to have been used in any other publications.

Jaakkola et al (20) published on the relationship between respiratory symptoms and pulmonary function decline over 8 years in a population of 391 young adults recruited from Montreal, Canada.

Jaakkola et al (20) used least squares regression to calculate the decline in FEV1 during follow-up. All available measurements from an individual were used to calculate

the slope of FEV1 decline. The calculated FEV1 decline per year was used as the outcome variable in the analyses involving respiratory symptoms.

The authors categorized subjects based on their response to the ATS questionnaire at baseline and at the final follow-up visit, ignoring the visits in between (20). A symptom was new if “it was absent at baseline but present at follow-up”(20). A symptom was classified as persistent if “it was present at both examinations”, and “remission of a symptom was defined if it was present at baseline but not present” at the final follow-up visits. Subjects were categorized as never having a symptom if the symptom was not reported at either baseline or follow-up.

Using the constructed symptom category as an independent variable, Jaakkola et al constructed a multiple linear regression model with FEV1 annual decline as the outcome variable and the category of symptom change over time as a predictor variable (20). Results indicated that particularly in non-smokers, development of new symptoms was associated with an increased loss of FEV1 during follow-up. Among former and current smokers the development of new symptoms as well as the persistence of symptoms was generally associated with more rapid FEV1 decline. These trends are of particular interest because the population was young (age 18-40 at baseline) and healthy, highlighting the importance of symptoms in relation to FEV1 decline even at an early age.

Although Jaakkola et al used a fixed effects regression model, they did attempt to incorporate the longitudinal changes in both symptoms (categories of change) and smoking behavior into their smoking variable used in the model (20). The smoking variable classified persistent and new smokers as “smokers”, ex-smokers either at baseline or follow-up as “ex-smokers” and never smokers at both baseline and follow-up were classified as “never smokers”. This accounts for a portion of the longitudinal changes in smoking behavior but does not take into account the timing of a change in smoking behavior.

Brodkin et al (21) studied respiratory symptoms in a cohort of men enrolled in the Seattle Asbestos Lung Cancer Chemoprevention Trial. Symptoms were classified based on responses at baseline and most recent follow-up, again ignoring the visits in between,

resulting in the same categories as Jaakkola et al (20): asymptomatic, persistent, development and resolution (remission).

The annual rate of decline in both FEV1 and FVC were used as outcome variables in separate fixed effects multiple linear regression analyses controlling for smoking, age, height, race, baseline spirometry measurements and asbestos exposure. From these analyses Brodtkin et al found that development of new symptoms (compared with consistent symptom reporting or resolution of symptoms) over the follow-up period was strongly associated with decrease in pulmonary function (21).

The men enrolled in this cohort attended annual follow-up visits where they reported on their symptoms, but only the first and last visits were used in analysis possibly preventing detection of annual changes in symptoms (21).

Because Brodtkin et al (21) incorporated the longitudinal nature of the data into the outcome and predictor variables (annual decline as outcome, longitudinal pattern of symptom change as predictor) their model did not allow for time-varying covariates. For example, changes in smoking or body weight during follow-up were not accounted for. For these variables Brodtkin et al used the baseline values in the model.

In 1990, Krzyzanowski et al (22) published the first of three analyses of respiratory symptoms using the Cracow and Tuscon longitudinal studies of obstructive lung disease data. The goal of this first paper was to determine whether the longitudinal changes in respiratory symptoms were related to the baseline lung function values or the decline in lung function over time, and whether these relationships were consistent between populations.

All subjects with three complete visits were included in the analysis (22). Subjects' pattern of respiratory symptoms were classified based on their reported symptoms (at all three visits) into the four categories previously described (never, persistent, new onset, remission) and this variable was used as a dummy variable in a multiple linear regression. Theoretically, subjects could report a pattern of Yes/No/Yes or No/Yes/No for a symptom at three visits. In order to have only four symptom change categories Krzyzanowski et al decided to group the Yes/No/Yes subjects in the Remission group

and the No/Yes/No subjects in the Persistent group. The baseline, final and annual decline in lung function were the outcome variables in three separate models.

Krzyzanowski et al (22) showed that symptom change over time is related to lung function, and that the relationship is consistent across two different populations. In particular the presence of any dyspnea or asthma syndrome (two of wheeze, attacks of dyspnea or asthma diagnosis) were related to both lower FEV1 at baseline and FEV1 annual decline.

Krzyzanowski used the Tuscon and Cracow data for two other publications on respiratory symptoms and lung function (23, 24), these analyses were limited to subjects with at least two visits and subjects who reported smoking at baseline and completed two visits, respectively. In the 1992 paper (23), subjects were categorized based on their symptoms reported at baseline and most recent follow-up. A dummy variable describing symptom change over time was used as a variable in a log linear model along with the covariates age, smoking, gender and city. Briefly, a log linear model is the equivalent of a multiple linear regression for categorical variables, and no one variable is considered the outcome; all variables are referred to as response variables (35). Covariates with significant effects (main and interaction) were then used in logistic regression models to estimate odds ratios for the symptom patterns (yes/no). Models indicated that current smokers were more likely to report persistent or incident symptoms and lifetime never smokers were more likely to report remission of symptoms. In two separate logistic regression models, Krzyzanowski compared persistent and incident symptoms to the group with no symptoms, and the group with resolved symptoms to the persistent symptoms group. This subsetting of the population to facilitate a logistic regression results in an exclusion of subjects and may lead to biased results.

In the 1993 Krzyzanowski publication (24) log linear models were used again, this time to determine the relationship between smoking cessation and other covariates. The significant effects were input into a logistic regression model to explain the effect of smoking cessation or persistence on the incidence or persistence of symptoms. Results indicated that persistent smokers have higher rates of persistent and incident symptoms compared to quitters.

Christiani et al (25) studied Chinese cotton textile workers over a 15-year period (1981 – 1996); silk textile workers were used as a control group. Respiratory questionnaires and pulmonary function tests were completed at four visits during the follow-up period. Respiratory symptoms reported on the questionnaires were used to construct two different types of symptom variables. The first variable described the number of times a subject reported the symptom during follow-up; this variable was a scale variable with possible values from zero to 4. The second symptom variable was a binary variable that described whether a subject ever reported a symptom.

These symptom variables were used in separate longitudinal linear regression models as predictors of change in the lung function parameters FEV<sub>1</sub> and FVC (25). Christiani et al used generalized estimating equations (GEE) to estimate the parameters in their regression model. Results from the analyses indicated that both cotton and silk workers who consistently reported respiratory symptoms at work (at three or four of the test visits) had significantly greater FEV<sub>1</sub> decline during the follow-up period (25).

In 2002, Wang et al (26) published a study of newly hired textile workers in Shanghai, China. Occupational exposure, lung function and respiratory symptoms were measured at baseline (before starting work) and at two follow-up surveys (one at three months and one at one year). As in Christiani (2001) (25), Wang et al (26) used longitudinal regression models (with GEE) to model the predictors of a change in lung function over follow-up. Results indicated that “symptomatic” workers, those who reported cough with phlegm at the three month follow-up, had greater loss of lung function.

The only publication to use a random effects to model respiratory symptoms at multiple points was published by Sherrill et al (27) in 1993. Sherrill and colleagues used a mixed effects model to describe the relationship between respiratory symptoms, smoking and lung function in the Tuscon cohort. There were six complete survey visits available in the Tuscon data and with the mixed effects model all six data points could be included (27) (unlike the other analyses of the same data that used only two or three visits (22-24)). In addition, time varying covariates (variables that are expected to change over time: symptoms, smoking, weight, job title) can be included for each visit date. Their model used lung function measures (FEV<sub>1</sub>, FVC and FEV<sub>1</sub>/FVC as three separate

models) as the outcome variable. Results from the models indicated that subjects reporting wheeze and dyspnea had lower lung function parameters (significant in all three models).

## **2.2.2 Symptoms as an Outcome Variable**

Several studies have used a measure of respiratory symptoms as an outcome in analyses, this is an important difference because there may be factors not traditionally considered in studies of lung function that are relevant to respiratory symptoms. In the literature there were many different metrics for modeling respiratory symptoms as an outcome: the onset of new symptoms using hazard models (7, 28) or logistic regression (11, 29), directional change in symptom using logistic regression (30), the rate of symptom incidence or prevalence using binomial logistic regression (34) and a continuous dyspnea scale in a linear regression (31).

### **2.2.2.1 Binary Outcomes**

Pahwa et al (28), Carta (11), Kongerud (7) and Boutet (29) all studied the development or onset of respiratory symptoms. Pahwa et al used Cox proportional hazards models, Kongerud et al (7) used the Turnbull algorithm (similar to Cox proportional hazards) and Carta and Boutet (11, 29) used logistic regression.

Pahwa et al reported on risk factors for developing new wheeze, using data from a health surveillance program among Canadian grain elevator workers (28). The surveillance program consisted of pulmonary function tests, respiratory symptom questionnaires and chest x-rays at three-year intervals. Surveillance began in 1978 and continued until 1990-93 (cycle 5); due to incomplete data, cycle 2 (1981) was considered baseline in the analysis. The analysis by Pahwa in 1998 (28) included only individuals who were asymptomatic with normal chest x-ray at baseline. The outcome of interest was the development of new wheeze at any of the follow-up cycles. Survival analysis (Cox's proportional hazards) was used to identify the factors predictive of the onset of new wheeze during follow-up. The model adjusted for age, height and smoking and included years of exposure (categorical variable) and FEV1/FVC (at baseline) ratio as possible predictors of new wheeze.



Pahwa et al had data on subjects for at least two visits and as many as four visits, subjects were followed at each visit attended and follow-up ended at the last visit attended or at the time wheeze was first reported, whichever occurred first. Despite the ability to use all of the symptom data (at all visits) the survival analysis method did not allow for time varying covariates to be included, instead Pahwa et al used the repeated measures of smoking to influence the creation of a single smoking variable accounting for changes in smoking behavior between baseline and last follow-up.

Results indicated that risk factors for the development of wheeze during follow-up were current smoking and decreased FEV1/FVC ratio at baseline. This seems to indicate that in this population of grain elevator workers, decreased pulmonary function measures preceded the development of wheeze.

Kongerud et al (7) studied the development of dyspnea and wheeze in a group of aluminum pot room workers who were asymptomatic at baseline. The Turnbull algorithm, similar to a Cox proportional hazard model was used to model the probability of developing wheeze and dyspnea. Risk factors considered were sex, age, allergy, workplace exposure to fluorides and smoking metrics. Results indicated that fluoride exposure and smoking were related to the development of wheeze and dyspnea.

Both Carta and Boutet (11, 29) used logistic regression to compare subjects who developed symptoms to asymptomatic subjects. Carta (11) studied the onset of dyspnea, chronic bronchitis, wheeze and "any symptom" in separate logistic models. Results indicated that workers in higher quartiles of exposure were more likely to develop symptoms. Boutet (29) investigated the predictors of reporting 2 or more symptoms at any time during follow-up; these subjects were classified as symptomatic and compared to asymptomatic subjects. Logistic regression results showed that symptomatic subjects were more likely to have bronchial hyper-responsiveness at baseline and also to have a personal history of rhinitis and a family history of asthma.

Gunnbjornsdottir et al (30) reported on data from the Respiratory Health in Northern Europe (RHINE) study, a follow-up on subjects who participated in the European Respiratory Health Study (ECRHS). Subjects were originally selected at random from the population as part of the ECRHS; the RHINE study then focused on ECRHS subjects

living in Iceland, Norway and Sweden. Subjects completed a postal questionnaire as part of the ECRHS and then completed a follow-up questionnaire as part of the RHINE study. The questionnaire contained questions on respiratory symptoms and indoor dampness among other factors (i.e. smoking, body mass index, socioeconomic status).

As in the case of Jaakkola (20) and Jedrychowski (19), Gunnbjornsdottir(30) categorized subjects symptom pattern over time based on their responses to the two questionnaires. This resulted in the same four patterns of symptom change as in previous papers: consistent symptoms, onset (new) symptoms, remission (resolution) of symptoms and symptomatic individuals. Subjects were also categorized as living in a 'damp' home or in a 'dry' home based on their questionnaire responses. Gunnbjornsdottir et al were interested in the association between living in a damp home and the pattern of respiratory symptom change over time. To answer this research question the authors subsetted the population based on the pattern of symptom change over time and ran two separate logistic regression models. First, the effect of a damp home on the onset of respiratory symptoms was modeled, with subjects experiencing a symptom onset being compared to asymptomatic subjects. Second, the effect of a damp home on the remission of symptoms was explored by comparing subjects experiencing a resolution (remission) of symptoms to subjects with persistent symptoms.

The results of the two logistic regression models indicated that living in a damp home was a risk factor for developing new symptoms and also that living in a damp home prevented symptom resolution (30). The separate models support the same conclusion: that living in a damp home increases the risk of symptom development and decreases the chance of symptom resolution.

#### **2.2.2.2 Scores, Scales and other Outcomes**

Mahler et al (1995) studied the longitudinal changes in dyspnea, general health and lung function in a cohort of COPD patients. A repeated measures analysis of covariance model was used to model dyspnea as a continuous dependent variable. Dyspnea was measured by a clinical tool that resulted in a dyspnea scale that could be considered as a continuous variable. Repeated measures ANCOVA analysis is a technique for accounting for the modifying effects of categorical independent variables on interval

dependent variables (in this case the dyspnea score) (36). The output from a repeated measure ANCOVA analysis can be interpreted in a similar way as regression output. Results from Mahler et al (31) indicated that patients with better lung function also had better dyspnea scores.

Another method for measuring the association between lung function and symptoms when symptoms are measured on a continuous scale is to calculate a linear correlation between the two variables. Lareau (32) and Hodgev (33) reported on the correlation between the longitudinal change in dyspnea and annual lung function decline. First an individual linear regression was run for each subject on their repeated lung function and their repeated dyspnea measurements. The calculated coefficients from each regression (lung function and dyspnea) were used as input for a linear correlation between dyspnea change and lung function change over time.

Results from Hodgev (33) indicated that dyspnea scores did decrease over time, and that this decrease was significantly correlated with FEV1 decline over the same period. Conversely, Lareau et al (32) did not observe a change in dyspnea during the study period and as a result found no correlation between dyspnea and FEV1 change over time. Both study populations were comprised of COPD patients.

Wu (34) used data from a surveillance program on steelworkers to study the rate of positive symptom responses using a binomial logistic regression. The outcome in this model was the total number of positive responses divided by the total number of visits (a rate). Because of how the rate outcome variable was constructed, Wu was able to include even subjects who had only one visit in the analysis. Included covariates were age, smoking, work location and work duration. Results indicated that working near the coke ovens (exposure source) was a risk factor for reporting a higher rate of symptoms.

## **2.2.3 Strengths & Limitations of Previous Studies**

### **2.2.3.1 Limitations**

Previous respiratory epidemiology studies have made many attempts to deal with the longitudinal data they have collected, but these approaches have not always been ideal. The use of coefficient estimates as outcome variables, the categorization of symptoms

over time, the exclusion of intermediate data points and the subsetting of datasets to ease analyses have all been reported and all have the potential to bias results.

Many respiratory epidemiology studies use the decline in lung function over time as an outcome in their analyses, and the studies involving symptoms are no different (19-22). This decline is usually the estimated coefficient from an individual linear regression of lung function on time. Other studies have also used this estimated coefficient as a variable in linear correlation between two variables of interest (32, 33). By using this estimated coefficient researchers are ignoring the error associated with the estimated coefficient. Using all of the lung function measurements and applying a mixed effects or longitudinal regression model would better address the change over time.

When considering symptom change over time, it is ideal to have as many measurements as possible. In six of the reviewed papers there was more respiratory symptom data available than was used in the analyses (19-24). In these papers, subjects were generally categorized based on their symptom responses at two time points, often the first and last visit, and all intermediate responses were ignored. Theoretically this means that someone could report no symptoms at the first and last visit, but have reported symptoms at every intermediate visit, and they would be classified as asymptomatic for the entire follow-up.

Conversely, in the 1990 study by Krzyzanowski (22), data from three visits was used to categorize subjects. Using three visits resulted in symptom response patterns (Yes/No/Yes, and No/Yes/Yes) that did not fit easily into the previously reported Never, New, Persistent and Remission symptom categories. The authors decided to put Yes/No/Yes subjects in the Persistent group and the No/Yes/No subjects in the Remission group. This decision may bias the results, but the magnitude of bias would be dependent on the number of subjects reporting these patterns, which was not reported.

When the number of repeated visits goes beyond two, the number of possible patterns increases exponentially. Three visits and a binary outcome results in eight possible patterns, four visits results in 16 possible patterns. Krzyzanowski (22) avoided the problem of having numerous categories of symptom change by forcing some categories into the four original categories reported in the literature.

In the study by Gunnbjornsdottir(30) the data was subsetting to allow for two separate logistic regression models to be run. In this case, the authors subsetting the data by symptom category. In the subsequent logistic regression models New symptoms were compared to Never symptoms, and Remission of symptoms were compared to Persistent symptoms. By stratifying the study population in order to complete logistic regression, they may have biased their results, or at least limited the applicability of their results. Logistic regression is based on the assumption that the probabilities of the two possible outcomes sum to one in the population. In the case of splitting the population, the probabilities of the two outcomes (in each group) do indeed sum to one, but with respect to the original population, they do not.

### **2.2.3.2 Strengths**

Despite the limitations of previous studies of longitudinal respiratory symptoms, there are also examples of 'better' approaches to addressing the longitudinal nature of the symptom data. Some studies have constructed new variables using information from repeated measures, used more appropriate statistical methods with lung function data (generalized estimating equations and random effects models) and considered novel symptom outcomes.

When constructing single variables to describe behaviors like smoking, studies have taken the repeated measures into account (20). For example, if a subject begins as a smoker but then quit during follow-up they are classified as a quitter or ex-smoker. In contrast, if the authors had used smoking status at baseline, the same subject would have been categorized as a smoker (21).

Christiani (25) and Wang (26) both used generalized estimating equations (GEE) (37) to estimate the parameters of marginal regression models. GEE is a method for estimating parameters in models with correlated data, as in the case of repeated measure on individuals. A marginal model with GEE handles longitudinal data, accounts for the correlation between repeated measures and allow for time varying covariates. But GEE requires the researcher to explicitly specify the structure of the correlation between repeated measures (i.e. is the correlation constant, or does it decrease over time etc.). The main benefit of GEE is that the estimation process produces unbiased coefficients

estimates even when the researchers assumptions about the correlation structure are incorrect.

Perhaps even better than a marginal model with GEE, is the approach of Sherrill et al (27) who applied random effects models to their study of lung function. Random effects models do not require an explicit assumption about the correlation structure between repeated measures. Random effects models also, as the name implies, allow the inclusion of random effects where the effect of the random variable on the outcome is estimated for each subject (in a longitudinal model). This approach is fairly new to the area of respiratory epidemiology and does not appear to have been used with symptoms as the outcome.

The approach of Wu et al (34) is also interesting because it makes use of all available data by calculating a rate of symptom reporting for each subject (# of symptom reported/ possible # of symptoms reported). This approach ensures that you are considering the individual rather than only the population prevalence, and results in a rate outcome for each subject in the dataset (34). If the pattern of symptoms is seen to be unimportant or perhaps too variable, using the rate as an outcome is a good alternative because the estimated coefficients will provide insight into the predictors of reporting symptoms more often.

Pahwa (28), Kongerud (7), Carta (11) and Boutet all used symptom development as the outcome in their analysis. Subjects were limited to asymptomatic individuals and followed for the onset of symptoms. These results can only inform about the development of new symptoms and not the resolution or persistence of symptoms, which may also be important symptom changes with respect to respiratory health.

## **2.3 Literature Review: Statistical Methods for Longitudinal Data Analysis**

Two SAS® procedures that can model repeated, correlated, binary data are reviewed and evaluated in this thesis:

- SAS® Trajectory Procedure (Proc Traj)
- SAS® Generalized Linear Mixed Model Procedure (Proc Glimmix)

Both of these procedures run on base SAS® v9.1, but are not included in the shipped program. They must be downloaded from the web, Proc Traj from the developer's website<sup>1</sup> and Proc Glimmix from the SAS® download centre<sup>2</sup>.

### **2.3.1 SAS® Proc Traj**

SAS® Proc Traj is a discrete mixture modeling procedure that is designed to model multiple patterns of change over time within a population (5). Bobby Jones, Daniel Nagin and Kathryn Roeder, from Carnegie Mellon University, designed Proc Traj (5). Unlike a traditional growth curve or regression model, which models only one group (the population mean), a mixture model identifies multiple distinct subgroups within the population and models the mean of each group. As in the case of the group mean in a traditional model, Proc Traj estimates an intercept and regression coefficients for each group in the mixture model. In comparison to the random/mixed effects (see Glimmix below), Proc Traj does not provide any estimates of individual deviation from the group mean. Perhaps, most importantly, Proc Traj models linear and non-linear trajectories of change. This means that research questions relating to the pattern of change over time in the outcome variable can be explored using this procedure.

Mixture models may be considered when the researcher expects there to be multiple trajectories over time in the population based on substantive knowledge, or perhaps, when there is little knowledge about how a certain outcome changes over time and one wants to model the unobserved heterogeneity in the data (5).

Proc Traj is capable of modeling longitudinal data with a binomial distribution (dichotomous outcomes), Poisson distribution (count outcomes) and normally distributed censored outcomes (a scale variable). There are a series of steps in the process of fitting a model using Proc Traj; these steps and other tips for using Proc Traj are explored in Appendix A.

---

<sup>1</sup> [www.andrew.cmu.edu/user/bjones/](http://www.andrew.cmu.edu/user/bjones/)

<sup>2</sup> [www.support.sas.com/rnd/app/da/glimmix.html](http://www.support.sas.com/rnd/app/da/glimmix.html)

Proc Traj has been almost exclusively applied in the social sciences, particularly in the criminology and psychology literature. Although the research using Proc Traj is not extensive, most have used Proc Traj to model count data (i.e. number of criminal conviction) or scale data (i.e. psychometric scale data – normally distributed but censored at zero). Fewer studies have used Proc Traj to model dichotomous outcomes.

In the original Proc Traj publication (5) Jones, Nagin and Roeder briefly describe an application of the logit Proc Traj model using the Cambridge Study of Delinquent Development data. In this example, the logit model was used to model the presence/absence of offenses, rather than a count of offenses. The data support a three group model where the majority (88%) of the subjects follow a trajectory of never offending. The remainder of the population follows either a high or low prevalence of offending during adolescence.

Mustillo et al (38) used a Proc Traj logit model to describe the presence/absence of obesity in a study of the development of psychiatric disorder in rural youth. Mustillo et al applied the SAS® Trajectory procedure because they “suspected that individuals do not vary continuously on obesity, but rather that there are a distinct number of obesity related trajectories”. The logit model was used because they modeled the presence/absence of obesity, rather than a scale or continuous measure of obesity. The dependent variable was obesity (yes/no) and the predictor variable was age. The data supported a best fit model with four groups: one with children who were never/rarely obese, a second with children who developed obesity over time, a third with children who were chronically obese and fourth group with children who moved from obese to normal during adolescence. Mustillo and colleagues used the results from the basic Proc Traj logit model (four groups with probability of membership in each group) for further analyses; these capacities of Proc Traj will be described further in the Chapter 3 and Appendix A.



Proc Traj has been almost exclusively applied in the social sciences, particularly in the criminology and psychology literature. Although the research using Proc Traj is not extensive, most have used Proc Traj to model count data (i.e. number of criminal conviction) or scale data (i.e. psychometric scale data – normally distributed but censored at zero). Fewer studies have used Proc Traj to model dichotomous outcomes.

In the original Proc Traj publication (5) Jones, Nagin and Roeder briefly describe an application of the logit Proc Traj model using the Cambridge Study of Delinquent Development data. In this example, the logit model was used to model the presence/absence of offenses, rather than a count of offenses. The data support a three group model where the majority (88%) of the subjects follow a trajectory of never offending. The remainder of the population follows either a high or low prevalence of offending during adolescence.

Mustillo et al (38) used a Proc Traj logit model to describe the presence/absence of obesity in a study of the development of psychiatric disorder in rural youth. Mustillo et al applied the SAS® Trajectory procedure because they “suspected that individuals do not vary continuously on obesity, but rather that there are a distinct number of obesity related trajectories”. The logit model was used because they modeled the presence/absence of obesity, rather than a scale or continuous measure of obesity. The dependent variable was obesity (yes/no) and the predictor variable was age. The data supported a best fit model with four groups: one with children who were never/rarely obese, a second with children who developed obesity over time, a third with children who were chronically obese and fourth group with children who moved from obese to normal during adolescence. Mustillo and colleagues used the results from the basic Proc Traj logit model (four groups with probability of membership in each group) for further analyses; these capacities of Proc Traj will be described further in the Chapter 3 and Appendix A.

### 2.3.2 SAS® Proc Glimmix

The Glimmix procedure was recently added to the SAS® platform in June 2006. Prior to this, researchers interested in these modeling techniques had to run a Glimmix macro embedded in the SAS® linear mixed models procedure (Proc Mixed).

Proc Glimmix is a SAS® procedure for constructing generalized linear models (39). Proc Glimmix is capable of running general and generalized linear models with and without random effects where the outcome has a normal, Poisson or binary distribution. Most important for this work is the capability of Proc Glimmix to model generalized linear mixed models (including random effects). Random effects are variables for which we are not particularly concerned with the fixed effect of each category, but we are interested in the variability between categories.

Proc Glimmix estimates the model parameters (intercept, regression coefficients) as well as the subject specific deviation from the parameters for each identified random effect. In comparison to Proc Traj, Proc Glimmix and generalized linear models in general, should be used when the trajectory of change is expected to be similar throughout the population, and also when the research question may require the inclusion of random effects.

Searches for Glimmix in databases of peer-reviewed literature produced few results. Web of Science located fifteen publications, while PubMed produced only nine – eight of which were common between the two databases. A further search using Google Scholar produced more results, likely due to two reasons: one, Google Scholar searches outside the medical and health science research searched by PubMed and Web of Science, and second because Google Scholar appears to be able to search the full text of articles rather than simply the abstract text. All but one of the publications located used the macro predecessor of Proc Glimmix.

Redpath et al (40) used the new Proc Glimmix procedure to model prey items delivered to the nests of harriers (a type of bird). In the model the outcome is a proportion, the proportion of prey items delivered that are considered large divided by the total number of prey item delivered to the nest. The logit link function in Proc Glimmix was used to apply the binomial distribution to the data and model the data.

Of particular interest to this thesis because of its research topic, is an article by Mendell et al (41) which investigated the relationship between the ventilation system in office buildings and the respiratory symptoms reported by the building occupants. Mendel and colleagues used the Glimmix macro and Proc Logistic in SAS® to construct a generalized mixed logistic regression model (logit link function).

In the mixed logistic model, the outcome was “work related symptom” (models were run for each symptom of interest) and the explanatory variables were related to the ventilation the office building where the subject worked (41). The random effect was the office building, so that the correlation between multiple measures of work related symptoms within the same office building were accounted for.

Mendell et al (41) also tested the same models using Proc Logistic alone, without any random effects (a fixed effect logistic regression) and compared the results to the mixed logistic regression models. As they expected, the fixed effect and mixed effect model produced similar point estimates for the model parameters, but with wider confidence intervals in the mixed effects model.

## **3 Results I: Evaluation of Proc Traj and Proc Glimmix for use in Respiratory Epidemiology**

### **3.1 Introduction**

As discussed in Chapter 2, repeated measures of respiratory symptoms are rarely analyzed and those studies that have attempted to investigate the pattern of symptom change over time have tended to use categorization techniques. The next section of this thesis will evaluate two statistical methods that are potentially useful for analyzing repeated respiratory symptom data: SAS® Proc Traj and SAS® Proc Glimmix. These two methods permit the study of repeated binary symptom data as the outcome and the inclusion of time-varying covariates.

Both SAS® Proc Traj and SAS® Proc Glimmix will be evaluated based on their applicability to six basic respiratory symptom research questions:

1. What factors predict reporting a respiratory symptom?
2. Do respiratory symptoms change over time?
3. What are the patterns of respiratory symptom change over time?
4. What factors predict respiratory symptom change over time?
5. Do different respiratory symptoms (e.g. cough and phlegm) change with similar/different patterns over time?
6. How does event occurrence affect the pattern of respiratory symptom change over time?

The ability of both Proc Glimmix and Proc Traj to answer the each question is outlined. In addition, the output generated from each model is briefly described and the relative strengths and limitations are discussed. For the purposes of comparison to a more common approach, and because simplicity is sometimes the best option, fixed effects logistic regression is compared and contrasted in some situations.

In addition to the information in this chapter, Appendix A is provided as a user guide for both Proc Traj and Proc Glimmix. Where necessary, reference to the Appendix is made, but information is also duplicated for the purposes of understanding.

## **3.2 Overview of the Methods**

### **3.2.1 SAS® Proc Glimmix: Generalized Linear Mixed Models**

By definition, a mixed model contains both fixed and random effects. Fixed effects are variables that are of particular interest to the research question, and for which all levels of interest are represented in the population (42). Random effects are generally variables for which only a sample of possible values is included and the interest is in the variability between levels and not in the particular effect of each level on the outcome (42). For example, in a longitudinal model the subject is specified as a random effect because we are not interested in the particular effect of each subject on the outcome, but we are interested in estimating the variability between individuals.

In a longitudinal mixed effects regression model the autocorrelation between repeated measures on individual subjects is accounted for, and one overall group mean is modeled. A mixed effect model runs in two stages, first the group mean is modeled and second the variation around this group mean is modeled for each subject. The output contains regression coefficients for the fixed effects and an estimate of the overall measure of the variability around the mean for each random variable.

When applying mixed effects regression in a longitudinal model, subject is specified as a random effect. The mixed model will account for the correlation between the repeated measures on each subject. When the intercept is specified as a random variable, the model is called a random intercept model. If variables in the model are specified as random variables, the model is referred to as a random coefficient model.

The SAS® platform has three procedures capable of running mixed models: Proc Mixed, Proc Glimmix, and Proc Nlmixed. Proc Mixed runs linear mixed model with a continuous normally distributed outcome variable. Proc Nlmixed runs non-linear mixed models for outcome variables belonging to a wide variety of distributions. And thirdly, Proc Glimmix runs generalized linear mixed models for both continuous and discrete outcome variables (39).

Longitudinal studies of respiratory symptoms result in repeated binary (yes/no) data. The binary nature of the outcome variable means that Proc Mixed cannot be used. Proc Nlmixed

is designed for advanced non-linear mixed models and is programmatically complex. Proc Glimmix is suitable for constructing a mixed logistic regression model, also called a generalized linear mixed model with logit link function.

Proc Glimmix is a fast, flexible procedure capable of running linear models (fixed effects), generalized linear models (fixed effects), linear mixed models (fixed and random effects) as well as generalized linear mixed models (fixed and random effects). The focus of Proc Glimmix for this research is the generalized linear mixed model capability. Proc Glimmix does not ship with the SAS® v.9, but the add-on and the documentation are both available for download on the SAS® support website<sup>3</sup>.

Proc Glimmix uses a link function to approximately linearize the model and then estimate the parameters as if it was a general linear model. Generalized linear mixed models with binary outcomes can result in biased parameter estimates when the variance components are large. This can happen when the number of subjects is small, the number of repetitions on each subject is low and the repeated measures are highly correlated (43). This discussion is beyond the scope of this thesis, but researchers should consult a statistician for advice in these circumstances (44). The potential for biased estimates in the Proc Glimmix models is one of the major limitations of the procedure, and researchers should keep this in mind when considering its application.

Users of Proc Glimmix should be aware that the modeling procedure makes an assumption that the random effect is normally distributed with mean equal to zero. This means that the estimated deviations for each individual from the estimated mean parameters (intercept and coefficients) will belong to a normal distribution.

When modeling using Proc Glimmix, the link function must be specified in the model statement (see Appendix A). The link function is what communicates the type of data you are modeling. With reference to respiratory symptoms (yes/no) the link function will always be logit. The user must also specify the categorical variables, the model structure and the random effects using the SAS® syntax (see Appendix A).

---

<sup>3</sup> [www.support.sas.com/rnd/app/da/glimmix.html](http://www.support.sas.com/rnd/app/da/glimmix.html)

The output from a generalized linear mixed model using Proc.Glimmix includes fit statistics, covariance estimates for the random effects and coefficient estimates for the fixed effects. (A description of the output and basics on interpretation can also be found in Appendix A.)

The covariance parameter estimates are a measure of the variability from the modeled group mean and are provided for each variable specified as random in the model syntax. If the variance parameter is significantly different from zero this indicates that there are significant differences in the subject level estimates for the random variable. For example, if a random intercept is included in the model and the estimated covariance parameter for the random intercept is significantly different from zero, this is interpreted to mean that subjects in the population have different intercept values, or in other words, they have different starting values. The covariance parameter estimates can also be thought of as the variability in the random variables that is not captured by the fixed effects.

The fixed effects parameter estimates are analogous to the output from a fixed effect regression model and are interpreted in the same manner. In the case of a logistic regression for respiratory symptoms, the coefficient estimates can be thought of as the change in risk of reporting a symptom for a one-unit increase in the associated independent variable (when the independent variable is continuous).

Deciding on model fit is more challenging because the output fit statistics do not include a traditional log-likelihood, in its place pseudo-likelihood is calculated by Proc Glimmix. This pseudo-likelihood cannot be used to compute a likelihood ratio test, instead, the researcher is advised to guide their modeling using substantive knowledge and other alternate methods for ensuring best model fit. Two examples are Akaike's information criterion (AIC) and Bayesian information criterion (BIC), in both of these fit indices smaller values indicate better model fit. Appendix A describes the syntax for requesting these fit indices in a SAS® Glimmix model.

### **3.2.2 SAS® Proc Traj**

Proc Traj is a specialized mixture model (45), and as a mixture model Proc Traj models multiple groups within the population, in contrast to a traditional regression or growth curve

model that models only one mean within the population. Designed by researchers, Proc Traj is not part of the base SAS® program and must be downloaded from Dr. B. Jones' website<sup>4</sup>. The paper which serves as the documentation for the program is also available for download at this website (5).

Proc Traj is designed to address research questions focused on describing the trajectory, or pattern, of change over time in the dependent variable, specifically questions concerned with multiple distinct patterns of change over time and modeling unobserved heterogeneity in the data.

Proc Traj models the number of distinct patterns of change over time in the dependent variable and the shape of each modeled pattern of change. Proc Traj estimates a regression model for each discrete group within the population. These parameters are modeled using maximum likelihood estimation, where the probability of the estimates is maximized based on the model structure. The significance of the estimated intercept and regression coefficients is tested using the Wald test.

Modeling in Proc Traj is step-wise and iterative (see Appendix A for further discussion). First, the number of trajectories of change over time in the population must be determined. Then the shape of each group's change over time must be specified. Before beginning, the researcher must use substantive knowledge to set reasonable limits on the modeling process.

For example, when thinking about respiratory symptoms the published literature indicates that when two time points are used there are four possible patterns of symptom change, and when there are more time points included there may be more groups. In this case, five might be the maximum number of groups modeled. To begin, a one-group model is computed then a two-group model et cetera until the a priori maximum number of groups is modeled.

Model selection in Proc Traj uses the change in the Bayesian Information Criterion (BIC) between two models to measure the weight of evidence against the null model. (For details of model fit, see Appendix A). Of important note is that Proc Traj models result in negative BIC values because the developers of Proc Traj use a slightly different equation for

---

<sup>4</sup> [www.andrew.cmu.edu/user/bjones](http://www.andrew.cmu.edu/user/bjones)



calculating the BIC (46). Therefore, in Proc Traj the best fit model is the one with the smallest negative BIC value.

Again, in reference to the example with respiratory symptoms, if we tested five models (one group up to five groups) we would have five BIC values to review. The comparisons are completed in a step-wise manner so that the two-group model is compared to the one-group model, and the three-group model to the two-group model and so on. The change in BIC from one model to the next is a measure of the evidence for one model versus the other.

The next step in fitting a model using Proc Traj is selecting the shape of each group's trajectory over time. Proc Traj can model up to a fourth order polynomial and can model both linear and non-linear trajectories within the same model. This can be done using substantive knowledge (i.e. we expect one group to never report symptoms so this group's trajectory will be a zero-order equation, or a straight line) or it can be done using the change in the BIC ( $\Delta BIC$ ). It seems ideal to use a combination of substantive knowledge and statistical inference to make the decision regarding the shape of each group's trajectory.

Two methods of using the BIC to make model fit decisions are described in the Proc Traj literature. The first describes the approximation of the logged Bayes factor using  $2 \cdot \Delta BIC$  (5). The second, known as Jeffreys's scale of the evidence describes approximating the Bayes Factor itself,  $B_{ij}$ , using  $e^{BIC_i - BIC_j}$  (47). Both provide a measure of the evidence for/against the complex model and both provide the same results. The crude scale (logged Bayes factor) is better suited to select the number of groups when the change in BIC between models is generally large. Jeffreys's scale of the evidence has a finer scale and seems to be better suited for selecting the shape of group trajectories (see Appendix A for the description of model fitting in Proc Traj).

The output from a Proc Traj model contains information on the number of groups, the shape of their trajectories, as well as the probability of group membership for the study population. In addition, posterior group membership probabilities and group assignments for each subject can be obtained from the output dataset. These output variables can be used to describe the population, or in further analyses as a dependent variable.

The macro included in the Proc Traj syntax creates a graphical output that visually describes the different groups, their trajectory of change over time and summarizes the group membership probabilities.

Time stable and time varying covariates can be included in the Proc Traj model using separate command lines (see Appendix A for specific details). Included covariates must be binary variables (values of 0/1) or continuous. Proc Traj does not support the inclusion of categorical covariates that have not been transformed into dummy variables.

In the case of time varying covariates the output does not change as drastically. If only time varying covariates are input, Proc Traj estimates a regression coefficient for the impact of the covariate on the outcome in each group. Group membership probabilities are still provided.

A post-model procedure can also be used to graphically model the effect of a change in the time varying variable on the predicted trajectory. The benefit to modeling the time varying covariates is that you can visualize the change in trajectory at the point in time where the value of the time varying covariate changes. The drawback is that you can only model one pattern of change (in the covariate) at a time.

Although the subject specific group assignment variable can be exported for use in further (separate) analyses, using a built-in Proc Traj option you can relate group membership to a subsequent outcome that occurs beyond the time frame included in the trajectory analysis. For example, a researcher could model the probability of a specific respiratory disease diagnosis later in life based on symptom trajectory group during the study period (before the diagnosis).

SAS® Proc Traj can also model multiple outcomes simultaneously (details in Appendix A). This is called a dual trajectory model; each outcome is modeled separately, then the two outcomes are modeled together. The output from these analyses is more complex but begins by presenting simple group membership probabilities for each outcome independently (as if each outcome was modeled independently), then shows conditional probabilities for each outcome conditional on the other, and finally contains joint probabilities for the dual trajectory. This capability may prove particularly useful for modeling related outcomes that you do not want to combine into one variable (e.g. cough and phlegm)

Researchers using Proc Traj should be careful not to reify the estimated groups. The groups themselves do not actually exist in reality; they are merely estimates of change over time. Because there is no random effect component in Proc Traj there is no estimate of subject level variability around the estimated group means.

Proc Traj should be considered as an analysis tool when you are interested in, or expect that there are, multiple patterns of change over time in your dependent variable.

### **3.3 Evaluation of Generalized Linear Mixed Models and Proc Traj Models**

Both Proc Traj and Proc Glimmix are theoretically capable of dealing with the repeated binary data generated from the ATS questionnaire. However, the specific research questions that may be answered by each technique must be clarified. The next section of this thesis describes how Proc Traj and Proc Glimmix can, or cannot, be used to answer six basic research questions relating to respiratory symptoms and how they change over time. The six research questions investigated are:

1. What factors predict reporting a respiratory symptom?
2. Do respiratory symptoms change over time?
3. What are the patterns of respiratory symptom change over time?
4. What factors predict respiratory symptom change over time?
5. Do different respiratory symptoms change with similar/different patterns over time?
6. How does event occurrence affect the pattern of respiratory symptom change over time?

An approach for answering each research question is described for both Proc Traj and Proc Glimmix. For comparison purposes, an approach using traditional fixed effects regression (SAS® Proc Logistic) is described where appropriate. The structure of each model is described and the interpretation of the output is discussed. In addition, the strengths and limitations of each approach for each research question are identified.

A summary of the analysis approaches described for each research question is shown in Table 3. This table is intended as a guide for determining which procedure, Proc Traj or Proc Glimmix, is most reasonable for a particular research question.

### **3.3.1 What Factors Predict Reporting a Respiratory Symptom?**

The factors that predict reporting a respiratory symptom at any point in time can be identified using a regression model, either fixed or random effects including risk factors of interest. The basic fixed effects model would be a cross sectional analyses of one visit from a longitudinal data set to determine what factors predict reporting a symptom at one point in time. This method does not make use of the repeated measures on individuals, but may be of interest if the research question itself is focused on one point in time (i.e. baseline measures, at last follow-up etc.).

A marginal logistic regression model can also be used to model the predictors of reporting a symptom. In this case one would use all of the repeated measures on all subjects and would specify the correlation structure between these repeated measures. Using a marginal regression model and specifying the correlation structure is a feasible method for modeling the longitudinal data, but may not be as robust as using a random effects model because the coefficient estimates may be biased if the correlation structure is misspecified and do not include random effects.

A generalized linear mixed model (i.e. Proc Glimmix) will account for the autocorrelation between repeated measure on subjects and will also allow each subject to vary around the group mean. This is valuable because the output will contain a measure of variability around the mean probability of reporting a symptom for each individual. Both marginal models and generalized linear mixed models will permit the inclusion of time-varying covariates.

Proc Traj is a mixture model with the goal of identifying multiple groups within a population, so this technique is not the ideal method for answering what predicts reporting a respiratory symptom. In addition, Proc Traj is concerned mainly with change over time and not on the likelihood of an outcome at any time. However, by looking at the predictors of group membership for a trajectory group that consistently reports symptoms, some insight

into the predictors of reporting symptom could be gained. However, a random effects logistic regression model is really a better choice for this research question.

### **3.3.2 Do Respiratory Symptoms Change Over Time?**

If the researcher is interested in whether respiratory symptoms are changing over time, the first analysis may be to calculate crude prevalence rates at each visit date in the data. This will provide population level estimates of how prevalent a symptom is at each visit and whether these rates are changing over time, but this will not provide any information on how, or whether, the symptoms are changing at the individual level. For example, the population prevalence could remain steady even though an equal number of individuals in the population are gaining symptoms as losing them. If the research question is focused at a population level and is only interested in whether the prevalence is changing at the population level, crude prevalence rates may suffice.

However, if the research question is concerned with the change over time at the individual level further analyses must be undertaken. The correlation (or lack of correlation) between repeated measures on individuals over time can be modeled using a mixed effects model. This can be achieved by setting up a model with each visit date as a covariate and requesting that the correlation between the outcome at each visit date be estimated. With the output you receive a correlation matrix of correlation coefficients that measures the correlation between the covariates (each visit date), and thus the correlation between visit 1 and visit 2 (and each pair wise comparison between visits) can be obtained from the output.

If the correlation between two visits is close to zero, the visits are not very highly correlated and there was probably a change in respiratory symptoms between these two measurements, but if the correlation is close to one the measurements are similar and it is not likely that a change occurred between the two visits. Although this can provide some insight into whether the respiratory symptoms are changing over time, there is very little information on the direction of change or pattern of change over time.

Using a generalized linear mixed model can account for the autocorrelation and provide an estimate of the variation around the modeled mean for the population. Similar to the fixed effects regression, the visit dates would be included in the model as covariates, but the

correlation structure does not need to be specified in any form. Because the visit dates are included in this manner the output will include the regression coefficients for each visit date; these regression coefficients indicate the probability of reporting a symptom at each visit date (recall that in logistic regression the coefficient is the probability of a '1' or 'yes' for a one unit increase in the value of the associated covariate). Interpretation of the output will allow for inferences to be made about whether the probability of reporting a symptom is the same or changing between visits, and unlike the fixed effects regression output, there will some ability to say whether reporting a symptom is more or less likely at one visit compared to another. Time can also be modeled as the continuous variable calendar year using a generalized linear mixed model. These results would provide an estimate of the change in the probability of the outcome per one year change in calendar year.

Proc Traj is specifically designed to identify multiple patterns of change over time within the population. Proc Traj is also capable of modeling non-linear change in the outcome variable. In order to conclude that a symptom was not changing over time the output from Proc Traj would have to support a single group model where the shape of the trajectory was flat (zero-order equation) or a multiple group model where all the individual group trajectories were flat. This can be determined by fitting the basic model to the data to determine how many groups are supported by the data and the shape of each group's pattern of change over time. Only in the circumstances described previously (all groups having zero-order patterns of change) could you conclude that a symptom was not changing over time. In other situations, Proc Traj will model multiple groups each with their own pattern of change over time and this will provide a great deal of information regarding whether symptoms are changing over time, and, how they are changing over time within the population.

### **3.3.3 What are the Patterns of Change in Respiratory Symptoms Over Time?**

The question of how respiratory symptoms change over time is a difficult question to answer when thinking about using a regression model to estimate the change over time in a binary outcome such as respiratory symptoms. The output from a regression model includes the regression coefficients and the intercept - the coefficients provide information on the

probability of an outcome occurring based on the value of the independent variable, and the intercept tells us the probability of an outcome occurring if all included covariates are equal to zero (so that the coefficients are nullified). From this output it is not possible to determine the shape or pattern of change in the outcome over time. In research question two, it was described how you can use the output from a logistic regression (or generalized linear mixed model) to determine whether the probability of the outcome occurring is changing over time using the regression coefficients, but it is not possible to determine how the outcome is changing over time.

Alternatively, Proc Traj is a powerful tool for answering questions of how an outcome changes over time. In addition, as a mixture model, Proc Traj is specifically designed to answer the question “how many patterns of change over time exist in the population?” In general, Proc Traj should not be applied when there is only one expected pattern of change (even if there was change expected). When only one group with one pattern of change is the expected in the population, the research questions are likely better answered with the use of a traditional growth curve models (not discussed in this thesis).

As in research question two, a Proc Traj model can be fitted to the data so that the number of distinct groups and the shape (linear or non-linear) of the change over time are described for all groups. The probability of group membership (for each identified group) will also be included in the output. These results will allow the researcher to describe the different patterns of respiratory symptom change over time, and also to discuss the prevalence of each pattern.

It is important to remember that group membership is fixed for each subject. Because Proc Traj does not model the subject level deviation from the group mean (as a mixed effects model does) each individual in the group is considered to follow the identical trajectory of change over time.

### **3.3.4 What Factors Predict Respiratory Symptom Change Over Time?**

The basic Proc Traj model can easily be adapted to include covariates, and thus help answer research questions about the predictors of group membership. When covariates are included in the Proc Traj model, the output changes slightly. If covariates are time varying

you will receive a regression coefficient for each covariate. For time stable covariates you will receive an estimate of the effect of the covariate on group membership. If you hypothesized that being a current smoker meant an individual was more likely to consistently report respiratory symptoms, you would expect the output to indicate that a current smoker is more likely to belong to a group that consistently reports respiratory symptoms (rather than a group that never reports a symptom during follow-up).

Another alternative, still using Proc Traj, would be to run a basic model with only the symptom of interest and time included to determine the number and shape of distinct groups within the population. The posterior group assignments contained in the output dataset can then be used to descriptively determine which factors differ between the groups. Additionally, the group assignment variable can be used as an outcome in a determinants of group assignment model.

Although these analyses seem similar, it is possible that they would return different results. The first option (Proc Traj model with covariates) uses the included covariates to predict not only group membership, but also the shape of each group's trajectory of change over time. The second option, exporting the group membership data and modeling it as a static outcome, is concerned simply with the predictors of membership in each group.

For research question three, it was concluded that neither random nor fixed effects logistic regression could be used to study the pattern of change in a dependent variable. However, it is possible to study the predictors of a change in a binary outcome variable such as a respiratory symptom using these methods.

In the case of generalized linear mixed models a new variable can be created to describe a change in symptoms. The new variable could describe a simple change, or it could describe a directional change in symptoms, for example:

- Yes symptom changed since last visit, no symptom did not change;
- Yes subject developed a new symptom, no subject did not gain a new symptom;
- Yes subject resolved a symptom, no subject did not resolve a symptom.

The generalized linear mixed model will determine the predictors of a change in respiratory symptoms at any point in time because there will be a measure of change for each



subject at each visit (except the first visit). The model continues to account for the autocorrelation between repeated measures of change (a subject with  $>2$  visits will have at least two measures of “symptom change”) and will still provide an estimate of the individual deviation around the mean parameter estimates.

One drawback to this method is the limitation of the outcome. By specifying only a “change” as the outcome, you cannot infer about the direction of this change. And, conversely, by specifying a directional change you are required to make a comparison against the remainder of the population - individuals who experienced a change in the opposite direction as well as individuals who experienced no change. Neither of these is an ideal situation, but both will provide some information on the predictors of a change in respiratory symptoms (either directional or not).

This research question can also be answered in another manner using mixed effects logistic regression. However this method is more observational and iterative than the models described previously.

In order to determine what factors predict respiratory symptom change over time using this mixed effects model you must include the visit dates as covariates (continuous or categorical) and time varying covariates for each visit date. The covariates in

your final model are the time varying variables you hypothesize to predict symptom change over time. Iteration is key, you must include different covariates in different models and compare the coefficients estimates and the model fit statistics to determine which model is best. If a time varying covariate input into the model completely removes the effect of time, then we may be able to conclude that this variable is predicting change over time. Model fit can be assessed by BIC (smaller is better). It is important to realize that using this model will only provide you with evidence suggesting covariates that may be predicting change over time.

### **3.3.5 Do Different Respiratory Symptoms Change with Similar Patterns Over Time?**

As discussed in research question three, logistic regression (either fixed effects or in a generalized linear mixed model) cannot be used to describe the pattern of change over time

in the outcome variable, so it is not possible to determine whether multiple respiratory symptoms change with the same pattern using any type of logistic regression.

A simple descriptive analysis of respiratory symptom change over time to investigate the similarities or differences between two or more symptoms could begin with calculating the correlation between pairs of symptoms over time. Correlation coefficients could be calculated for each visit date to determine whether the correlation between the two symptoms was changing over time. This approach is simple and straight forward, but would be limited to comparisons between pairs of symptoms. The output would also not provide any insight into similarities or differences in the shape of change over time, only the similarity between two symptom responses at specific time points.

With Proc Traj however, there are two ways to determine whether different respiratory symptoms change with similar patterns. One is to qualitatively compare the number and shape of each subgroup identified for each respiratory symptom, the second is to run a dual trajectory model for two respiratory symptoms.

Qualitatively comparing the shape and probability of the group trajectories for different respiratory symptoms is a simple way to determine whether different symptoms appear to be changing in the same manner over time. For example, is there the same number of distinct subgroups in the population with respect to different symptoms? Do these distinct subgroups experience the same pattern of change over time? This approach is simple, and will allow for comparisons to be made between multiple respiratory symptoms, however this approach will not allow for determining joint group membership probabilities for multiple symptoms, nor will it allow for statistical significance to be determined.

The alternative to a qualitative comparison is to run a Proc Traj model for two symptoms, also called a dual trajectory analysis. The results from a dual trajectory analysis include the single trajectory results for each of the two included symptoms (parameter estimates and membership probabilities), the probabilities of group membership for each symptom group conditional on the other symptom, and the joint probabilities for the dual trajectories.

The main limitation of this approach is that you are limited to modeling only two respiratory symptoms (as compared to the qualitative comparisons). However, this approach will identify significantly different subgroups of a dual trajectory.

### **3.3.6 How Does a Time-varying Covariate Affect the Pattern of Respiratory Symptom Reporting?**

As has been discussed in reference to research questions three and five, it is not possible to study the pattern of change in a binary outcome variable using logistic regression. Therefore it is not possible to address research question six using either fixed effects logistic regression or a generalized linear mixed model. Each of these models could describe the effect of a change in a time varying covariate on the probability of reporting a respiratory symptom (research question one) by creating a dummy variable to denote the change. A dummy variable can be created for each visit date to denote a change since last visit (only in a mixed model), or a single dummy variable can be created to denote a change ever. However, no information on how the event changes the pattern of change over time can be gained.

Time varying covariates can be incorporated into a Proc Traj model and the effect of a change in this covariate on the shape of each group's trajectory can be determined. The pattern of change in the time-varying covariate must be specified in the syntax. Proc Traj will complete the basic model then the effect of the time varying covariate on the modeled trajectories. The result is a graph showing the basic trajectory shape overlaid with the trajectory shape taking into account the time varying covariate. If the time varying covariate is affecting the symptom trajectory, these two lines will diverge at the point in time where the time varying covariate changed. Although this is a useful tool, it is limited to investigating the effect of one pattern of change in the covariate at a time.

## **3.4 Summary**

In conclusion, both generalized linear mixed models (Proc Glimmix) and finite mixture models (Proc Traj) are potentially very useful for analyzing respiratory symptom data from longitudinal studies. Each has strengths and limitations, but both have been underutilized in the literature to date.

Proc Traj is best suited for addressing research questions interested in the pattern or shape of change over time, and questions interested in the heterogeneity of this change over time. This finite mixture model can determine the number of distinct patterns (trajectories) of

change over time and the shape of each. In addition, the predictors of these groups and patterns can be explored using time stable and time varying covariates. Results can be exported and used in separate analyses or, within Proc Traj, to predict outcomes beyond the time studied in the trajectory analysis. Proc Traj is not suited for answering research questions that are interested in the population mean or research questions that require the inclusion of random effects. Researchers must be careful to remember that the groups estimated by Proc Traj are not real entities; they are the estimates of the multiple patterns of change within the population.

Generalized linear mixed models are best suited for research questions where the goal is to account, or adjust, for the autocorrelation inherent in repeated measures data and explore research questions interested in the overall population mean. In repeated measures data, generalized linear mixed models are valuable tools for investigating the subject level deviation around the group mean (for the specified random effects). These mixed models are able to answer questions relating to whether a change over time is occurring and are more useful for determining the predictors of the change, but are not able to model the pattern of the change over time. There are also known limitations in the estimation technique employed in Proc Glimmix that may result in biased coefficient estimates. Researchers must keep this possible bias in mind and consult a statistician for help to ensure that the bias is minimized.

The findings from the previous section are summarized and organized by research question in Table 3. Table 3, in addition to Appendix A, is intended to help to guide research in the study of respiratory change over time using generalized linear mixed models and finite mixture models.

**Table 3 Summary of findings on the utility of Proc Glimmix and Proc Traj for each outlined research question**

Res. Question	Technique	Description	Output	Strengths	Limitations
1. What factors predict reporting a respiratory symptom?	Proc Traj	Proc Traj is not the ideal technique for answering this research question. However, the factors predicting membership in a group (see Research Question #3) that persistently reports a symptom could partially answer this question.			
	Proc Glimmix	Use a mixed effects logistic regression model with covariates of interest to determine which are predictive of reporting a symptom	Regression coefficients for each independent variable Measure of each individual deviation from the group mean	Able to easily incorporate covariates into model Account for autocorrelation between repeated measures on individuals	Unable to identify subgroups within the population
	Proc Logistic	Use a fixed effects logistic regression to model the probability of reporting a symptom Either use only one visits data (one measure per subject) or use all measures per subject and specify a correlation structure	Regression coefficients describe the probability of reporting symptom given the independent variable associated with each coefficient	Can answer what predicts reporting a symptom for one point in time (i.e. baseline), which may be important	Using one measure per subject is a simple cross sectional analysis Estimated when using all of the data you are dependent on the correlation structure
2. Do respiratory symptoms change over time?	Proc Traj	Fit basic model to describe multiple patterns of symptom change over time as well as the shape of the change for each group	Number and shape of distinct patterns of change over time Probability of membership in each group	Models distinct groups Models non-linear as well as linear change	Requires a larger sample size than a mixed effects regression model A priori decision about number of groups required
	Proc Glimmix	Use mixed effects logistic regression model each visit date as a covariates in the model, compare these coefficients for each visit date to determine if risk is changing.	Regression coefficients describe probability of reporting a symptom at each visit date Measure of each individual's deviation from the modeled mean	Models the probability of reporting a symptom Accounts for correlation between repeated measures Models the variability around the mean for each individual	No ability to determine the pattern of change or the presence/absence of multiple groups Potential for statistical power problems when running multiple models
	Proc Logistic	Use a marginal regression model to model the correlation structure between the covariates "visits date" for each visits	Modeled correlation structure gives estimates of correlation (from 0->1) between visits dates	Gives a measure of the probability of a change in symptom response from one visit to another	Only reasonable to model the correlation structure with short chains of repeated measures (<5 repeats) No mixed effect component because you are modeling the correlation structure

**Table 3 Cont'd**

Res. Question	Technique	Description	Output	Strengths	Limitations
3. What are the patterns of symptom change over time?	Proc Traj	Fit basic model to describe multiple patterns of symptom change over time as well as the shape of the change for each group	Number and shape of distinct patterns of change over time Probability of membership in each group	Describes multiple patterns of change within the population Probability of membership provides a measure of the size of each group	No measure of variability around the mean of each group
	Proc Glimmix	The pattern of change over time cannot be modeled with a logistic regression model			
	Proc Logistic				
4. What factors predict respiratory symptom change over time?	Proc Traj	Introduce covariates into the basic model, determine the probability of group membership conditional on the covariate	Number of patterns of change over time Shape of each pattern Probability of membership in each group	Able to determine which independent variables are predictive of membership in each symptom change group	No measure of variability around the mean of each group (individuals assigned to the group are assumed to each follow an identical pattern of change)
		Run the basic model including only time, then export group membership data and use this variables as the outcome in further analyses	Probability of membership in each group given the value of the included covariates		
	Proc Glimmix	Use a logistic regression model to estimate the probability of a symptom change over time, include covariates in the model Outcome variable: symptom change - yes/no new symptom - yes/no resolved symptom - yes/no	Regression coefficients for each independent variable estimate the probability of the outcome Estimate of each individual's deviation from mean	Able to determine the magnitude and direction of effect for each independent variables on the probability of the outcome occurring	Potential for bias is results Challenge to interpret results due to heterogeneous comparison group
		Used mixed effects model to explore the effect of time varying risk factors on symptom outcome	Regression coefficients Estimated variances	Simple approach - compare the estimated coefficients and model fit to determine the best model	Research question is answered based on comparison between two models

**Table 3 Cont'd**

Res. Question	Technique	Description	Output	Strengths	Limitations
5. Do different respiratory symptoms change with similar patterns over time?	Proc Traj	Compare output for two different symptoms (number of groups and shape of trajectory)	For <b>each</b> symptom: Number and shape of patterns of change over time, Probability of membership in each group	Can compare multiple single trajectories in an observational manner	Observational approach, no way to test whether trajectories are statistically different
		Model dual trajectory for two symptoms	Number and shape of distinct patterns of change over time for each symptom alone Conditional probabilities for group membership Joint probabilities of group membership	Able to observe joint trajectories for two symptoms together Determine statistical significance of trajectories	Limited to modeling dual trajectories (cannot model 3 or more symptoms together)
	Proc Glimmix	The pattern of change over time cannot be modeled with a logistic regression model, either mixed or fixed effects			
	Proc Logistic				
6. How does a time-varying covariate affect the pattern of respiratory symptom reporting over time?	Proc Traj	Run a basic model including time and any time stable covariates, then incorporate a time varying covariate and the pattern of the time varying covariate you are interested in	Number and shape of distinct patterns of change over time Probability of membership in each group Probability of trajectory change (and shape of change) for the time varying covariate pattern specified in the model	Able to determine whether, and how, a time varying covariate affects the trajectory of change	Important to include any independent variables which may have an impact on, or be affected by, the event occurrence
	Proc Glimmix	The pattern of change over time cannot be modeled with a logistic regression model			
	Proc Logistic	However, you could get some information about the influence of a time varying covariate on the probability of reporting a respiratory symptom (research question #1). To do this you could include the time varying covariate at each time point and use the estimated coefficient for each time point to determine how the occurrence of the event affects the probability of reporting a respiratory symptom.			

## 4 Results II: Case Study

### 4.1 Introduction

As summarized in the literature review (Chapter 2), previous studies have not taken full advantage of repeated respiratory symptom data and few published studies have used symptoms as an outcome. Those that have studied symptoms as an outcome have categorized symptom change over time (30), studied only the onset of new symptoms (7, 11, 28, 29) or used a symptom scale or rate as the outcome (31, 34). We are interested in the study of the respiratory symptoms as an outcome, particularly in reference to the six research questions outlined in Chapter 3.

Chapter 3 reviewed two methods, Proc Traj and Proc Glimmix, for handling repeated binary outcomes and repeated measures of covariates. The current chapter is a Case Study of these methods using previously collected data from an occupational surveillance program.

For the Case Study, we will explore the first three research questions from Chapter 3:

1. What factors predict reporting a respiratory symptom?
2. Do respiratory symptoms change over time?
3. What are the patterns of change over time in respiratory symptoms?

Until this point, this thesis has discussed respiratory symptoms in general. In the interest of limiting the analyses to a manageable body of work, this Case Study will focus exclusively on the respiratory symptom dyspnea, or breathlessness.

Dyspnea is most commonly measured as a binary symptom using the ATS questionnaire (7, 11, 18-30, 34) but can be measured as a continuous outcome using less common clinical tools (31-33). Dyspnea has been shown to be a predictor of mortality (48, 49) but the association between dyspnea and lung function has been less conclusive (32, 33). Other studies have investigated the relationship between dyspnea and occupational exposure in working populations; conclusions regarding these associations were mixed (7, 50-52). Overall, there is a need for further study of dyspnea.



The Case Study is a key piece in evaluating Proc Traj and Proc Glimmix for use in respiratory epidemiology. The results from the different models will be compared, their limitations described and conclusions regarding the usefulness of each will be made.

## **4.2 Study Population**

A surveillance study of marine transportation workers in British Columbia was initiated in 1987 to monitor workers with a previous asbestos exposure, but soon expanded to include workers without exposure. The program was conducted by the Occupational and Environmental Lung Diseases Research Unit at the University of British Columbia. The surveillance program data currently spans twelve years and five visit years: 1987, 1989, 1991, 1994 and 1999. The study continues to date, and a sixth testing period is currently underway.

The marine transportation workers in the study are involved in the maintenance and operation of the coastal ferries in British Columbia. This population includes maintenance, engineering, catering, cleaning, ticketing and traffic control workers. Historically some of these workers have experienced asbestos exposures as part of their work activities. Workers may also have current or historical exposures to combustion by-products, car exhaust, solvents from cleaning products and other respiratory hazards. These exposures were not quantitatively measured as part of the study protocol, but detailed work history was collected from all subjects.

Initially, contact information for the participants was collected from the employer. A letter inviting workers to participate in the study was sent to the home addresses obtained from the employer. Participants gave written consent when returning the letter, and also at each test date. The study protocol was approved by the UBC Clinical Research Ethics Board.

The data collection has been repeated on five occasions: 1987, 1989, 1991, 1994 and 1999. Subjects were recruited in both 1987 and 1989, and no subject has recorded visits in both 1987 and 1989. The maximum number of repeat visits a participant may have is four (either 1987 or 1989, and 1991, 1994 and 1999). For descriptive analysis, the visits

in 1987 and 1989 were combined into a 1988 visit year, because no subjects had data at both visits.

The same standardized testing procedures were followed during each test period. At each test date subjects provided answers to the ATS respiratory questionnaire, performed spirometry, provided a chest radiograph and underwent skin prick testing.

## **4.3 Methods: Data Collection**

### **4.3.1 ATS Questionnaire**

At each visit trained personnel administered the ATS questionnaire in person. At the beginning of the questionnaire subjects were read the following instructions:

“These are questions mainly about your health. Please answer Yes or No. If in doubt about the answer, please answer No.”

This instruction is intended to increase the probability of a true positive response (increased sensitivity). Subjects responded to a series of questions relating to cough, phlegm, wheeze, dyspnea, chest tightness, history of asthma and allergy, smoking activity as well as general health questions.

### **4.3.2 Spirometry**

Spirometry included measurement of forced expiratory volume in one second (FEV<sub>1</sub>) and forced vital capacity (FVC). Spirometric measurements were made using a 10 liter dry rolling seal spirometer (S&M Instruments, Doylestown PA) according to standard techniques recommended by the American Thoracic Society (53). Subjects were tested while seated, wearing nose-clips. At least three acceptable forced expiratory blows were obtained from each participant and expiration was continued until a visible one-second volume plateau was achieved. The best values for FVC and FEV<sub>1</sub> were used; flow rates were taken from best blow.

### **4.3.3 Skin Prick Testing**

At each testing visits, subjects completed skin prick testing. Three allergens were tested on each subject: cat dander, dust mites and grasses, along with a positive control

(histamine) and a negative control (saline). The diameter of each subject's wheel (skin response) for each allergen and control was measured in millimeters (mm) and a positive allergy result was recorded if any of the allergens resulted in a wheel greater than 3mm in diameter.

#### **4.3.4 Previous Analysis**

Workers were categorized into asbestos risk exposure groups based on reported job history, specifically jobs that were held for at least two years more than 10 years prior to the testing occasion. The time lag for jobs was applied because of the latency period associated with asbestos related disease. The risk groups were developed using information on the use of asbestos on marine vessels and after consultation with senior employees and asbestos inspection personnel.

Results from this process indicated that workers engaged in ship repair at the maintenance dock, workers in the engine room and workers on the live aboard vessels were considered to have high historical asbestos exposure. Workers in all other areas were assumed to have a low historical asbestos exposure.

Predicted values for FEV1 were calculated using the prediction equations described by Crapo et al (54). Predicted values were adjusted downwards by 15% for subjects who neither white nor native Indian, to account for natural differences in lung function.

At each visit smoking status was assigned based on self-reported use of cigarettes, pipes and cigars. Subjects who reported ever using cigarettes, pipes or cigars, but reported no current use were categorized as former smokers. Subjects who reported never using any of these tobacco products were classified as never smokers, and those reporting current use were considered current smokers.

### **4.4 Methods: Current Analysis**

#### **4.4.1 Study Population**

The population for the current Case Study is limited to subjects with repeated measures of respiratory symptoms. Subjects with only one test visit were excluded

(n=776). Demographic information was summarized for the entire population (n=1701) and the subset used in analyses (n=925).

#### 4.4.2 Outcome

The focus of this analysis is dyspnea, specifically subjects' responses to the first dyspnea question in the ATS questionnaire: "Are you troubled by shortness of breath when hurrying on the level or walking up a light hill? – yes/no".

#### 4.4.3 Definitions

Several potential risk factors for reporting dyspnea were explored: age, atopy, race, smoking, number of complete visits, high historical asbestos exposure, current respiratory irritant exposure, history of childhood asthma and percent predicted FEV1. Each potential risk factor variable is described in Table 4.

**Table 4 Summary of variables considered as risk factors for reporting dyspnea**

Variable	Type	Time-varying?	Effect	Description
Age	Continuous	Yes	Fixed	Age reported at each visit
FEV1, percent predicted	Continuous	Yes	Fixed	FEV1 percent of predicted based on (54)
Atopy	Categorical, yes/no	Yes	Fixed	Positive allergic response to any of the three allergens tested
Race	Categorical, white/nonwhite	No	Fixed	Self reported racial heritage
Smoking Status	Categorical, never/former/current	Yes	Fixed	Smoking status at each visit.
Number of Visits	Categorical, two/three/four	No	Fixed	Number of complete tests for each subject
High Historical Asbestos Exposure	Categorical, yes/no	No	Fixed	Asbestos exposure risk group
Current Exposure to Respiratory Irritants	Categorical, yes/no	Yes	Fixed	Exposure to respiratory irritants in current job at each visit
Childhood Asthma	Categorical, yes/no	No	Fixed	History of doctor diagnosed asthma resolving before adulthood

Subjects who reported doctor diagnosed asthma that resolved before the age of nineteen were classified as having childhood asthma. Race was categorized as white and nonwhite based on subjects' self reported racial heritage. Age was reported at each visit.

The variable describing current exposure to respiratory irritants was created using the work area and job title data for each visit. Subjects working in the passenger areas, on the bridge and anywhere at the terminals were assumed to have little or no exposure to respiratory irritants. Workers at the maintenance dock, on the car deck, in the engine room and in the kitchen were assumed to have exposure to respiratory irritants. Job titles for the exposed subjects were reviewed to permit exclusions of administrative positions in each work area.

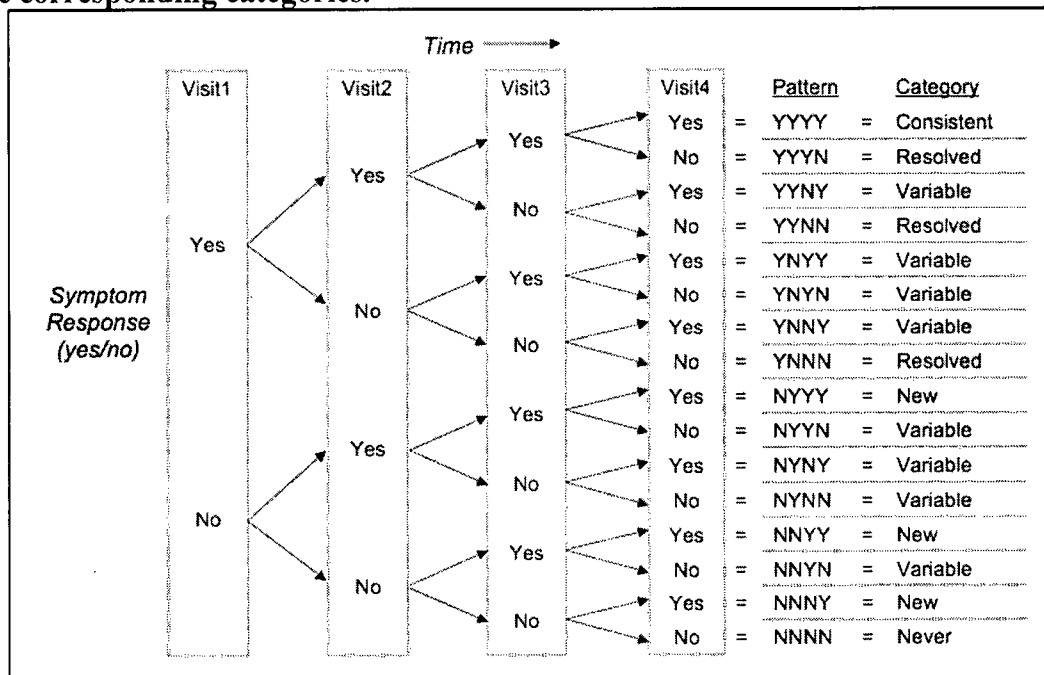
#### **4.4.4 Descriptive Analysis**

The entire surveillance population and the subset used in analyses ( $\geq 2$  visits) are described with respect to dyspnea risk factors. The prevalence of dyspnea at each visit date was calculated for men and women, and the trend over time was described for male subjects with data at all four visits.

Using responses to the dyspnea question at each visit date, a variable describing the pattern of dyspnea responses was constructed. An extension of the categorization described previously in the literature (19-23) was required to accommodate individuals in the data with as many as four repeat visits (previous studies have dealt with only two or three repeat visits).

This resulted in sixteen possible patterns of symptom change over time that are shown schematically in Figure 1. After determining the pattern of symptom change over time for each individual, the patterns were grouped into five categories similar to the categories in the published literature except that there was an additional group, which was called "Variable". The Variable group includes patterns of symptom change over time that did not fit into the four groups previously defined given the number of repeat measures in this study.

**Figure 1 Schematic representation of the patterns of symptom change over time and the corresponding categories.**



**Table 5 Summary of the categories describing dyspnea change over time**

Category	Definition
Asymptomatic	Never reported any of cough, phlegm, wheeze or dyspnea
Developed New Dyspnea	Did not report dyspnea at first visit, reported dyspnea at a subsequent visit, continued to report dyspnea
Resolved Dyspnea	Reported dyspnea at first visit, stopped reporting dyspnea at subsequent visits and continued to not report dyspnea
Always Reported Dyspnea	Reported dyspnea at every visit
Variable Reporting Pattern	A pattern of positive and negative responses that does not fit into the above categories

After constructing the variable describing the pattern of dyspnea change over time, the new categories were used as grouping variables and were related to smoking, number of visits and lung function using simple descriptive statistics. All descriptive analyses were stratified by sex.

Relationships between categorical risk factor variables and dyspnea were investigated using chi-square tests. Associations between continuous risk factors and dyspnea were explored using Student's t-test. Age was considered as both a categorical and continuous variable. Histograms for continuous risk factor variables were created to confirm normally distributed values. Chi-square tests between risk factor variables were completed to ensure highly correlated variables were not entered into the models together (results shown in Appendix B).

The association between dyspnea and known indicators of respiratory disease (FEV<sub>1</sub>, current asthma and chronic bronchitis) was also completed.

#### **4.4.5 Traditional Fixed Effects - Proc Logistic**

SAS® Proc Logistic was used to construct fixed effect logistic regression models. Two models were constructed. The first modeled the predictors of reporting dyspnea at baseline using every subject's first visit data. The second model was a 'flawed' model for the predictors of dyspnea, which used all of the data for every subject and ignored the correlation between repeated measures. Both models were repeated including an adjustment for percent predicted FEV<sub>1</sub>.

Risk factors that resulted in  $p < 0.2$  in chi-square tests with dyspnea were considered for entry into the fixed logistic model. Once in the model, variables with coefficients  $p < 0.10$  remained in the model. Models were manually constructed in a backwards stepwise process.

#### **4.4.6 Proc Glimmix**

Risk factors that resulted in  $p < 0.2$  in chi-square tests with dyspnea were considered for entry into the mixed model. Once in the model, variables with coefficients  $p < 0.10$  remained in the model. Models were manually constructed in a backwards stepwise process.

A marginal (fixed) model was run to estimate the correlation structure between repeated measures of dyspnea. Subject was identified as the unit on which repeated measures were made. The model included dyspnea as the outcome and each visit date as

a separate dummy predictor variable. The estimated correlation matrix was used to infer information about the correlation between measures of dyspnea at different visit dates.

Next, a mixed model including risk factors for reporting dyspnea was constructed using Proc Glimmix. A random intercept was included. Subject was included as a random effect to account for the correlation between repeated measures on subjects. No random variables were included. Initial and final models are reported where appropriate to demonstrate the modeling process.

Finally, mixed models (random intercept) using only visit dates as dummy covariates were constructed to estimate the odds ratios for reporting dyspnea at each visit date. A model was first run using 1988 as the reference visit, and another saturated model was run without an intercept to allow comparison between each visit year. An additional model was run using time as a continuous variable to estimate the population change in the probability of dyspnea per calendar year. Subject was included as a random effect in all models. These results were used to infer on whether the probability of reporting dyspnea was changing over time.

#### **4.4.7 Proc Traj**

Proc Traj models the change in dyspnea over time. Time was input as a two-digit visit year variable (i.e. 1988 = '88'). Based on previous literature using two time points, it is clear that there are four possible patterns of change over time in the dyspnea outcome (Never, New, Resolved, Persistent). With the addition of more data points (as in the marine workers cohort) a fifth pattern of change was considered likely. For these reasons a five group model was considered the maximum number of groups permitted. All groups were allowed to follow a quadratic (second order) equation while fitting the number of groups as recommended by Nagin (47).

In previous published studies, the "never reported the symptom" group is generally the largest in the population. This knowledge lead to a decision to force one of the group's trajectories to a zero order (flat line) equation to represent the subjects who never report dyspnea throughout follow-up. This limitation was only enforced on the model after the number of groups had been decided.



As described in Chapter 3 and Appendix A, the Bayesian Information Criterion (BIC) was used to assess model fit and ultimately decide on the number of groups, as well as the trajectory shape for each group.

Proc Traj models were completed for the entire dataset (n=925) as well as a subset of men with complete data (n=148) to ensure that the missing data did not influence the results of the model.

## **4.5 Results**

### **4.5.1 Descriptive Analysis**

The entire cohort of marine workers consisted of 1701 individuals with at least one complete visit. A visit was considered complete if the ATS questionnaire was answered and the subject recorded acceptable FEV1 and FVC measurements.

For the analysis of symptom change over time only subjects with two or more visits were included (n=925), a total of 2472 visits were included in this subset sample.

The demographics of the entire population (Table 6) and the subset (Table 7) indicate that the two groups are similar. Specifically, the subset population is primarily male with men and women being approximately the same age at baseline. The percent of predicted lung function variable shows that on average, subjects had normal lung function at enrollment.

Among subjects included in the analysis, approximately 30% of men and women reported themselves as never smokers at baseline. Among subjects self-reporting as current smokers, the reported pack-years of smoking were similar between men and women. However, men had higher rates of atopy at baseline than women (34% vs. 25%). Men tended to have more history of asbestos exposure and were more likely to experience current exposure to respiratory irritants.

**Table 6 Demographics of the entire marine transportation workers cohort**

Variable		Men (n=1473)		Women (n=228)	
Age at first visit, mean (SE)		46.97	0.28	46.17	0.81
FEV1 (L) at first visit, mean (SE)		3.70	0.02	2.74	0.04
FEV1 % Predicted at first visit, mean (SE)		95.7	0.39	97.2	1.07
Pack Years of Smoking, mean (SE)		24.84	0.64	22.37	1.44
Asthma at first visit, n (%)		103	7%	26	11%
Atopy at first visit, n (%)		485	33%	56	25%
Smoking, n (%)	Never	408	28%	64	28%
	Former	624	42%	70	31%
	Current	441	30%	94	41%
Race, n (%)	White	1225	83%	222	97%
	Non-white	248	18%	6	2%
High Historical Asbestos Exposure, n (%)	No	581	39%	209	92%
	Yes	892	61%	19	8%

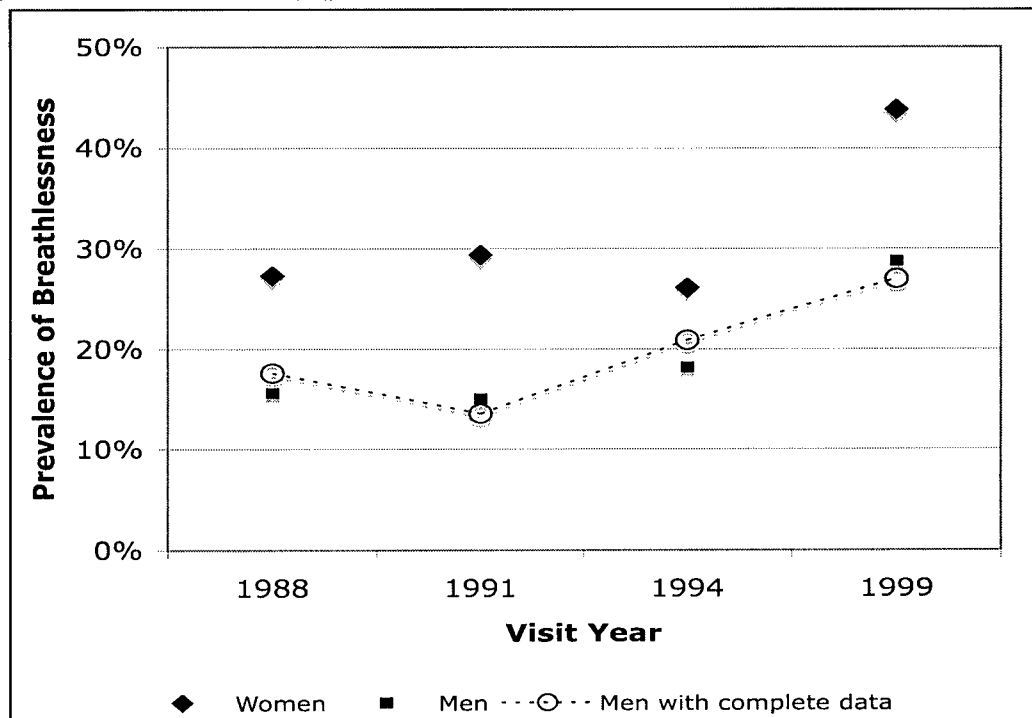
**Table 7 Demographics of the subset used in respiratory symptom analyses (workers with  $\geq 2$  test dates).**

Variable		Men (n=824)		Women (n=101)	
Age at first visit, mean (SE)		45.84	0.34	46.81	1.09
FEV1 (L) at first visit, mean (SE)		3.79	0.03	2.69	0.05
FEV1 % Predicted at first visit, mean (SE)		96.7	0.48	96.2	1.43
FEV1 decline (mL/yr), mean (SE)		49.4	2.11	25.6	5.32
Pack Years of Smoking, mean (SE)		23.99	0.85	25.58	2.30
Asthma at first visit, n (%)		54	7%	7	7%
Atopy at first visit, n (%)		277	34%	25	25%
Smoking, n (%)	Never	248	30%	27	27%
	Former	340	41%	32	32%
	Current	236	29%	42	42%
Race, n (%)	White	709	86%	99	98%
	Non-white	115	14%	2	2%
Number of Visits, n (%)	Two	399	48%	60	59%
	Three	277	34%	33	33%
	Four	148	18%	8	8%
High Historical Asbestos Exposure, n (%)	No	272	33%	90	89%
	Yes	552	66%	11	11%
Current Exposure to Respiratory Irritants, n (%)	No	242	29%	92	91%
	Yes	582	71%	9	9%

Dyspnea prevalence rates for each visits date are shown in Table 8 and Figure 2. In general, women reported more dyspnea throughout the study and there was a trend

towards higher prevalence of dyspnea at the later visits in both sexes. The trend over time in dyspnea prevalence is shown for men with complete data (n=148) and this trend is similar to the population level trend observed.

**Figure 2 Prevalence of dyspnea for men and women between 1988 and 1999.**



**Table 8 Crude prevalence rate for dyspnea by visits year, stratified by sex.**

	Percentage of Subjects Reporting Dyspnea		
	Men	Women	p-value*
1988	15.6%	27.3%	0.006
1991	14.9%	29.4%	<0.001
1994	18.2%	26.1%	0.2
1999	28.8%	43.8%	0.08

\*Chi-square for differences between men and women by year

The results from chi-square analysis of the relationship between dyspnea and indicators of respiratory disease indicated that FEV1 and chronic bronchitis were associated with dyspnea in men (Table 9). The lack of association in women is likely limited by sample size because several cells had less than five subjects.

Relationships between dyspnea and potential personal risk factors were also investigated using chi-square tests (Table 10). All risk factors investigated except for atopy and race showed at least a weak association with dyspnea in men. Again, the analyses involving women only were limited by small sample size.

Dyspnea was associated ( $p < 0.05$  in Student's t-test) with lower FEV1 percent predicted as well as older age in both women and men (Table 11).

Both age and percent predicted FEV1 were considered as continuous variables in the modeling process. Histograms for the age and FEV1 percent predicted variables were constructed and both variables were approximately normally distributed (results not shown).

**Table 9 Association between dyspnea and known indicators of respiratory disease (based on responses at first visit)**

		% with Dyspnea			
		Men (n=824)		Women (n=101)	
		Yes	p*	Yes	p*
FEV1, percent predicted	>80%	14%	<0.0001	24%	0.02
	<80%	31%		60%	
Current Asthma	No	16%	0.3	28%	0.9
	Yes	21%		28%	
Current Chronic Bronchitis	No	12%	<0.0001	26%	0.4
	Yes	41%		35%	

\*Chi-square for differences between respiratory disease categories

**Table 10 Association between dyspnea and risk factors for reporting dyspnea (based on responses at first visit)**

		% Reporting Dyspnea			
		Men (n=824)		Women (n=101)	
		%	p*	%	p*
Atopy	No	16%	0.6	29%	0.6
	Yes	15%		24%	
Race	White	16%	0.4	27%	0.5
	Non-white	13%		50%	
Smoking Status	Never	9%	0.002	18%	0.4
	Former	19%		34%	
	Current	19%		28%	
Number of Visits	Two	18%	0.1	30%	0.5
	Three	12%		21%	
	Four	18%		38%	
High Historical Asbestos Exposure	No	15%	0.8	28%	0.9
	Yes	16%		27%	
Current Exposure to Respiratory Irritants	No	22%	0.002	29%	0.2
	Yes	13%		11%	
Childhood Asthma	No	16%	0.09	28%	0.7
	Yes	26%		20%	

\*Chi-square for differences between risk factor categories

**Table 11 Means (SD) for continuous risk factors for dyspnea, using measures from subjects' first visit.**

		Dyspnea		p-value*
		No	Yes	
Men	Mean Age (years)	48.4	52.9	<0.0001
	Mean FEV1, percent predicted	96.4	86.8	<0.0001
Women	Mean Age (years)	47.0	54.5	<0.0001
	Mean FEV1 percent predicted	97.6	91.7	0.005

\*p-value for Student's t-test between subjects with and without dyspnea.

The categories constructed to describe respiratory symptom change over time for dyspnea are represented in the first column of Table 12. Table 12 summarizes the distribution of dyspnea change over time by smoking and by number of visits for men and women, respectively. Asymptomatic men and women were subjects who never reported any symptom (none of cough, phlegm, wheeze or dyspnea) during the study

period. Of the groups reporting dyspnea, the New (developed) symptom group was the largest for both men and women.

Of importance is the observation that the more visits a subject completed, the more likely they were to report a change in dyspnea during follow-up. Subjects in both the asymptomatic and the persistent groups had the same responses throughout follow-up, and these groups have high proportion of subjects with only two visits. Subjects who have more visits (three or four) tended to belong to the categories that describe a dyspnea change over time (New, Resolved, Variable). Table 13 describes the change in FEV1 over time stratified by the pattern of dyspnea change over time. In women, the presence of dyspnea seems to have a greater impact on the rate of lung function change. Whereas in men, the effect of reporting dyspnea or a change in dyspnea, is harder to determine using this descriptive analysis.

**Table 12 Prevalence of dyspnea change categories by smoking status and number of visits**

	Frequency		Current		Former		# of Visits, n					
	n	%	n	%	n	%	2	%	3	%	4	%
Asymptomatic Men	184	22%	21	11%	78	42%	108	59%	56	30%	20	11%
Dyspnea (Men)												
Persistent	56	7%	17	30%	31	55%	39	70%	11	20%	6	11%
Variable	46	6%	15	33%	20	43%	0	0%	25	54%	21	46%
New	104	13%	20	19%	53	51%	44	42%	35	34%	25	24%
Resolved	54	7%	13	24%	35	65%	30	56%	12	22%	12	22%
Asymptomatic Women	15	15%	-	-	6	40%	11	73%	4	27%	-	-
Dyspnea (Women)												
Persistent	12	12%	3	25%	7	58%	9	75%	2	17%	1	8%
Variable	3	3%	1	33%	1	33%	0	0%	2	67%	1	33%
New	20	20%	6	30%	8	40%	10	50%	8	40%	2	10%
Resolved	14	14%	6	43%	7	50%	9	64%	4	29%	1	7%

**Table 13 FEV1 annual change stratified by dyspnea change over time**

	N	FEV1 slope (ml/yr)	
		Mean	SE
Asymptomatic Men	184	-51.00	4.56
Dyspnea (Men)			
Persistent	56	-67.63	11.16
Variable	46	-48.20	4.86
New	104	-53.63	5.51
Resolved	54	-39.27	5.91
Asymptomatic Women	15	-12.11	9.43
Dyspnea (Women)			
Persistent	12	-8.86	21.69
Variable	3	-30.83	12.57
New	20	-35.18	14.20
Resolved	14	-29.38	15.48

#### **4.5.2 Traditional Fixed Effects - Proc Logistic**

Fixed effects models were constructed for comparison purposes. All risk factors with  $p < 0.20$  in descriptive analysis were offered to the model: age, sex, smoking, work area, exposure frequency and number of visits.

First a model predicting dyspnea was run using only data from the first visit. The final model results are shown in Table 15 (Table 14 shows the initial model before non-significant variables were removed). The final model indicates that at the first visit subjects who are older and female are more likely to report dyspnea. Subjects with current respiratory exposure are more likely to report dyspnea. This finding may be a result of the healthy worker effect, whereby workers experiencing negative health effects are moving out of high exposure work areas.

**Table 14 Initial model for predictors of dyspnea using only first visit responses (n=925). Results from fixed effects logistic regression.**

	Estimate	SE	
Intercept	-3.34	0.54	
	OR	95% Wald CL	
Age	1.03	1.01	1.05
Female	1.54	0.88	2.69
Childhood asthma	1.63	0.79	3.36
Never Smoker	ref		
Former Smoker	1.92	1.18	3.11
Current Smoker	2.18	1.32	3.60
High Historical Asbestos Exposure	1.20	0.80	1.82
Current Respiratory Irritant Exposure	0.55	0.36	0.82
Two Visits	ref		
Three Visits	0.74	0.49	1.12
Four Visits	1.25	0.76	2.06
Model AIC	821		

**Table 15 Final model for predictors of dyspnea using only first visit responses (n=925). Results from fixed effects logistic regression.**

	Estimate	SE	
Intercept	-3.18	0.50	
	OR	95% Wald CL	
Age	1.03	1.01	1.05
Never Smoker	ref		
Former Smoker	1.93	1.20	3.11
Current Smoker	2.16	1.32	3.54
Current Respiratory Irritant Exposure	0.53	0.37	0.75
Model AIC	820		



**Table 16 Final fixed logistic regression model, adjusted for lung function, using only first visit responses**

	Estimate	SE	
Intercept	0.34	0.91	
	OR	95% Wald CL	
FEV1, percent predicted	0.97	0.96	0.98
Age	1.02	1.00	1.04
Never Smoker	ref		
Former Smoker	1.81	1.12	2.93
Current Smoker	1.80	1.08	2.98
Current Respiratory Irritant Exposure	0.54	0.38	0.77
Model AIC	854		

After adjusting for FEV1 percent predicted (Table 16), there were no changes in the risk factors. Lower FEV1 percent predicted was associated with reporting dyspnea. Smoking and older age remained risk factors and exposure to respiratory irritants in the current job remained a protective factor.

For comparison purposes only, a fixed effects model of the predictors of dyspnea was also run using all the data points and ignoring the correlation between repeated visits. The repeated visits on individual subjects were treated as independent observations for this model. This model is wrong from a statistical point of view, but the results are presented so that we may compare the results to those of the mixed model adjusting for the autocorrelation. The initial and final models using all the data (ignoring autocorrelation) are shown in Table 17 and Table 18 respectively. The models using all of the data indicated that older age, being female, reporting childhood asthma, smoking and having more complete visits were all risk factors for reporting dyspnea. Substantially more risk factors were identified in this model than in the model using only subjects' first visit data. As in the fixed effects model using only baseline data, exposure to respiratory irritants in the current job was protective for reporting dyspnea.

**Table 17 Initial model for predictor of dyspnea using all visits, and all subjects (n=2472). Results from fixed effects logistic regression.**

	Estimate	SE	
Intercept	-3.78	0.32	
	OR	95% Wald CL	
Age	1.05	1.04	1.06
Female	1.52	1.08	2.15
Childhood asthma	2.84	2.01	4.01
Never Smoker	ref		
Former Smoker	1.10	0.85	1.43
Current Smoker	1.72	1.29	2.28
High Historical Asbestos Exposure	1.08	0.84	1.39
Current Respiratory Irritant Exposure	0.62	0.50	0.79
Two Visits	ref		
Three Visits	0.76	0.60	0.97
Four Visits	1.11	0.85	1.46
Model AIC	2273		

**Table 18 Final model for predictor of dyspnea using all visits, and all subjects (n=2472). Results from fixed effects logistic regression.**

	Estimate	SE	
Intercept	-3.76	0.31	
	OR	95% Wald CL	
Age	1.05	1.04	1.06
Female	1.48	1.06	2.05
Childhood asthma	2.85	2.02	4.02
Never Smoker	ref		
Former Smoker	1.11	0.86	1.44
Current Smoker	1.72	1.29	2.29
Current Respiratory Irritant Exposure	0.64	0.51	0.80
Two Visits	ref		
Three Visits	0.76	0.60	0.98
Four Visits	1.13	0.86	1.47
Model AIC	2271		

**Table 19 Final fixed effects logistic regression, adjusted for lung function, using all data**

	Estimate	SE	
Intercept	-0.01	0.56	
	OR	95% Wald CL	
FEV1, percent predicted	0.97	0.96	0.98
Age	1.03	1.02	1.04
Female	1.63	1.16	2.28
Childhood asthma	2.16	1.50	3.11
Never Smoker	ref		
Former Smoker	1.01	0.78	1.32
Current Smoker	1.34	1.00	1.81
Current Respiratory Irritant Exposure	0.64	0.51	0.81
Two Visits	ref		
Three Visits	0.71	0.56	0.92
Four Visits	1.06	0.81	1.39
Model AIC	2427		

Again, the model was adjusted for FEV1 percent predicted (Table 19). In the adjusted model all included covariates remained significant. Lower FEV1 percent predicted was again associated with reporting dyspnea.

### 4.5.3 Proc Glimmix

The first mixed effect model using Proc Glimmix was constructed to describe the predictors of dyspnea. Subject was included as a random effect as was a random intercept variable. Variables offered into the model resulting in estimated coefficients that were significant at  $p < 0.10$  level were left in the model. The initial and final models are shown in Table 20 and Table 21. Older age, being female, smoking and reporting childhood asthma were risk factors for reporting dyspnea. Current exposure to respiratory irritants is protective for reporting dyspnea and there seems to be a trend towards less dyspnea in subjects who have a three complete visits. Recall that the random intercept variance (shown in all Proc Glimmix results) is a measure of variance in the random intercept between subjects in the model.

**Table 20 Initial mixed model using all data (n=2472) to determine predictors of reporting dyspnea at any point in time**

	Estimate	SE	
Intercept	-4.11	0.40	
Random Intercept Variance	1.37	0.18	
	OR	95% Wald CL	
Age	1.05	1.03	1.07
Female	1.70	0.97	2.98
Childhood asthma	2.99	1.71	5.23
Never Smoker	ref		
Former Smoker	1.20	0.80	1.81
Current Smoker	1.75	1.12	2.74
High Historical Asbestos Exposure	1.08	0.74	1.59
Current Respiratory Irritant Exposure	0.66	0.46	0.94
Two Visits	ref		
Three Visits	0.73	0.49	1.07
Four Visits	1.13	0.72	1.78
Model Pseudo-AIC	11792		

**Table 21 Final mixed model using all data (n=2472) to determine predictors of reporting dyspnea at any point in time**

	Estimate	SE	
Intercept	-4.09	0.40	
Random Intercept Variance	1.37	0.18	
	OR	95% Wald CL	
Age	1.05	1.03	1.07
Female	1.65	0.96	2.82
Childhood asthma	3.01	1.72	5.25
Never Smoker	ref		
Former Smoker	1.21	0.81	1.81
Current Smoker	1.76	1.12	2.74
Current Respiratory Irritant Exposure	0.67	0.47	0.95
Two Visits	ref		
Three Visits	0.73	0.50	1.07
Four Visits	1.15	0.74	1.80
Model Pseudo-AIC	11789		

**Table 22 Final mixed model describing predictors of dyspnea, adjusted for lung function.**

	Estimate	SE	
Intercept	-0.03	0.71	
Random Intercept Variance	1.32	0.18	
	OR	95% Wald CL	
FEV1, percent predicted	0.97	0.96	0.98
Age	1.04	1.02	1.05
Female	1.74	1.12	2.72
Childhood asthma	2.38	1.49	3.79
Never Smoker	ref		
Former Smoker	1.03	0.74	1.45
Current Smoker	1.32	0.91	1.93
Current Respiratory Irritant Exposure	0.67	0.50	0.90
Two Visits	ref		
Three Visits	0.67	0.49	0.93
Four Visits	0.99	0.68	1.42
Model Pseudo-AIC	11925		

After adjusting for lung function (Table 22) (using percent predicted FEV1) older age, being female and history of childhood asthma were risk factors for reporting dyspnea, but smoking was not a significant predictor of dyspnea. Again results indicated that working in a job with current exposure to respiratory irritants was associated with less dyspnea and subjects with three complete visits were also less likely to report dyspnea (as compared with having two visits).

The next Proc Glimmix models were constructed to determine whether the probability of reporting dyspnea was changing over time.

First, a model including visit year as the only covariate was used to estimate the correlation between repeated measures of dyspnea. The model did not converge when all subjects were included, likely due to the substantial amount of missing data. The model did converge when the data was limited to men with complete date (n=148). The estimated correlation matrix for the limited dataset is shown in Figure 3. All between-visit year correlation estimates are less than 0.5, implying that none of the dyspnea

responses were highly correlated. The variation in the correlation estimates indicates that in consecutive visits years, responses to the dyspnea question are more correlated than visits further apart ( $r=0.34$  between visit1 and visit2, but  $r=0.28$  between visit1 and visit4). The correlation estimates also suggest that the correlation between later consecutive visits is greater than between earlier consecutive visits ( $r=0.34$  between visit1 and visit2,  $r=0.47$  between visit3 and visit4). This may also be interpreted as an age effect, suggesting that as subjects aged their responses were more likely to be correlated at consecutive visits.

**Figure 3 Estimated correlation matrix (correlation between repeated measures of dyspnea) for men with complete data (n=148)**

	88	91	94	99
88	1.00	0.34	0.20	0.28
91		1.00	0.43	0.38
94			1.00	0.47
99				1.00

Next, a mixed effect model with a random intercept and dummy variables for each visit year was constructed. The estimated regression coefficients indicate that for both men and women (Table 24 and Table 25 respectively) the probability of reporting dyspnea changed over time.

In the last study year (1999), the probability of reporting dyspnea was greater than at baseline. This difference was only statistically significant in men. The estimated variance in the random intercept was significantly different from zero in both men and women indicating the subjects had different intercept values, but the variance appeared to be larger in men than in women.

**Table 23 Model for the effect of visit year on reporting dyspnea including random intercept term, all subjects (n=2472)**

	Estimate	SE	
Intercept	-1.76	0.11	
Random Intercept Variance	1.50	0.18	
	OR	95% Wald CL	
1988	ref		
1991	0.98	0.74	1.30
1994	1.24	0.90	1.71
1999	2.23	1.63	3.04

**Table 24 Model for the effect of visit year on reporting dyspnea including random intercept term, Men (n=2221)**

	Estimate	SE	
Intercept	-1.86	0.12	
Random Intercept Variance	1.54	0.19	
	OR	95% Wald CL	
1988	ref		
1991	0.96	0.70	1.30
1994	1.29	0.91	1.83
1999	2.30	1.65	3.21

**Table 25 Model for the effect of visit year on reporting dyspnea including random intercept term, Women (n=251)**

	Estimate	SE	
Intercept	-1.04	0.27	
Random Intercept Variance	1.06	0.44	
	OR	95% Wald CL	
1988	ref		
1991	1.09	0.54	2.19
1994	0.98	0.41	2.37
1999	2.09	0.82	5.32

This model can also be run as a saturated model with no intercept. When the intercept is excluded the coefficients for consecutive visit years can be easily compared. Results from the saturated model (Table 26) show that in 1988, 1991 and 1994 the odds ratio for reporting dyspnea are similar, but in 1999 the odds ratio is significantly greater as was shown in the random intercept model above (Table 23).

**Table 26 Saturated model demonstrating the effect of each visit year on reporting dyspnea.**

	Estimate	SE	
Residual Variance	1.00	0.028	
	OR	95% Wald CL	
1988	0.20	0.17	0.24
1991	0.20	0.16	0.24
1994	0.23	0.19	0.30
1999	0.43	0.35	0.53

A final model including the risk factors for dyspnea, an adjustment for FEV1 percent predicted as well as the visit year covariates (time trends) was constructed to determine whether the effect of visit year was actually an age effect. The results from this final model (Table 27) indicate that both visit year and the number of complete visits are significantly associated with dyspnea even after adjusting for identified risk factors (including age) and FEV1 percent predicted.

**Table 27 Final mixed model using all data (n=2473), risk factors, time covariates and an adjustment for lung function (% predicted FEV1)**

	Estimate	SE	
Intercept	0.88	0.70	
Random Intercept Variance	1.32	0.18	
	OR	95% Wald CL	
Age	1.03	1.01	1.04
Female	1.84	1.18	2.87
FEV1, percent predicted	0.96	0.95	0.97
Childhood asthma	2.19	1.38	3.50
Current Respiratory Irritant Exposure	0.68	0.51	0.91
Two Visits	ref		
Three Visits	0.63	0.46	0.87
Four Visits	0.86	0.59	1.25
1988 Visit	ref		
1991 Visit	0.82	0.61	1.10
1994 Visit	1.01	0.72	1.43
1999 Visit	1.51	1.05	2.17
Model Pseudo-AIC	11918		



It is also possible to use time, or visit year, as a continuous variable to determine whether the probability of reporting dyspnea is changing over time. A simple analysis of the population time trend is shown in Table 28. For each increase of one calendar year, the probability of dyspnea increases by seven percent.

**Table 28 Model of the effect of calendar year (continuous) on reporting dyspnea, all subjects (n=2472)**

	Estimate	SE		
Intercept	-2.04	0.12		
Random Intercept Variance	1.48	0.18		
	OR	95% Wald CL		
Time, in years	1.07	1.04	1.10	
Model Pseudo-AIC	11574			

#### 4.5.4 Proc Traj

Results from the Proc Traj model are shown here, beginning with the model fitting process. Table 29 outlines the process of selecting the “best” number of groups in the mixture model. Five models were fit to the data: a one group model up to a five group model. In each model all included groups were assigned a quadratic (second order) equation to describe their trajectory. Results are shown for the complete dataset (n=925) as well as for a subset of men with complete data (n=148).

For each model a BIC value was obtained from the output. This BIC value was recorded and compared to the null model. For each new model the null model was the group with one less group. For example, the two group model was compared to the one group model (null model), and the three group model was compared to the two group model (null model).

Results in Table 29 show that Model 2 was an improvement over Model 1 due to the smaller BIC and the larger  $2\Delta\text{BIC}$  value. According to the criteria of Jones, Nagin and Roeder (5) there is strong evidence against Model 1 when comparing Models 1 and 2. But when Model 2 and Model 3 were compared, there was no evidence against Model 2 suggesting that Model 2 is the better model.

**Table 29 Model fit statistics for stepwise iterations of Proc Traj model to determine number of groups using all subjects' data (n=925 subjects)**

Model	No. of Groups	BIC n=2472	2*ΔBIC	Model Comparison	Evidence
1	one	-1211.02	-	-	-
2	two	-1114.73	192.58	Model 2 : Model 1	Very strong evidence against Model 1
3	three	-1129.33	-29.20	Model 3 : Model 2	No evidence against Model 2
4	four	-1145.09	-31.52	Model 4 : Model 3	No evidence against Model 3
5	five	D.N.C.*	-	Model 5 : Model 4	-

\*model did not converge

**Table 30 Model fit statistics for stepwise iterations of Proc Traj model to determine number of groups only male subjects with complete data (n=148 subjects)**

Model	No. of Groups	BIC n=592	2*ΔBIC	Model Comparison	Evidence
1	one	-300.17	-	-	-
2	two	-279.98	40.38	Model 2 : Model 1	Strong evidence against Model 1
3	three	-287.15	-14.34	Model 3 : Model 2	No evidence against Model 2
4	four	D.N.C.*	-	Model 4 : Model 3	-
5	five	D.N.C.*	-	Model 5 : Model 4	-

\*model did not converge

Next, the shape of each group's change over time was determined using the output BIC. A two group model was fit to the model and the order of each group's trajectory was manipulated to determine which had the best fit. One of the groups in the two group model was constrained to be a flat trajectory (zero order equation) because descriptive analysis of dyspnea indicated that the majority of the population never reported dyspnea (67% of men, 51% of women - Table 12)

Results of fitting the shape of each group's trajectory are shown in Table 31. Again, the BIC output was used, this time the difference in BIC between models was assessed using Jeffreys's scale of the evidence (47). Values of less than one indicate that the alternate (complex) model is favored; a value greater than one indicates that the null model is favored. Results from this process suggest that the two group model with a zero order and a first order equation is the best fit model. In other words, there are two trajectories of change in dyspnea over time, one group with a low (but non-zero)

probability of dyspnea and another group with a linearly increasing probability of dyspnea.

After deciding on the number of groups and the shape of each groups' trajectory, the final model was run. The estimated coefficients in Table 33 describe two regression equations, one for each group. The estimate for the linear effect of year (continuous) on dyspnea can be transformed into an odds ratio of 1.15 (95% CL 1.08-1.22).

**Table 31 Model fit statistics for stepwise iterations of Proc Traj model to determine trajectory shape over time using all subjects' data (n=925 subjects)**

Model	Order of equations	BIC n=2472	Bij	Model Comparison	Evidence for/against
1	0 0	-1122.32	-	-	-
2	0 1	-1111.30	0.00	Model 2 : Model 1	Strong evidence for Model 2
3	0 2	-1114.13	16.95	Model 3 : Model 2	Strong evidence for Model 2
4	0 3	-1118.06	50.91	Model 4 : Model 3	Strong evidence for Model 3
5	0 4	-1121.58	33.78	Model 5 : Model 4	Strong evidence for Model 4

**Table 32 Model fit statistics for stepwise iterations of Proc Traj model to determine trajectory shape over time using subjects with complete data (n=148 subjects)**

Model	Order of equations	BIC n=592	Bij	Model Comparison	Evidence for/against
1	0 0	-272.88	-	-	-
2	0 1	-269.80	0.05	Model 2 : Model 1	Strong evidence for Model 2
3	0 2	-272.71	18.36	Model 3 : Model 2	Strong evidence for Model 2
4	0 3	-275.91	24.53	Model 4 : Model 3	Strong evidence for Model 3
5	0 4	-278.91	20.09	Model 5 : Model 4	Strong evidence for Model 4

**Table 33 Final two-group model describing the change in the probability of dyspnea over time, using all subjects' data (n=925)**

Group	Parameter	Estimate	SE	p	Membership Probability
1	Intercept	-2.92	0.25	<0.001	72.8%
2	Intercept	-12.96	2.78	<0.001	27.2%
	Linear (Year)	0.14	0.03	<0.001	

**Table 34 Final two-group model describing the change in the probability of dyspnea over time, using subjects with complete data (n=148)**

Group	Parameter	Estimate	SE	p	Membership Probability
1	Intercept	-2.59	0.27	<0.001	76.4%
2	Intercept	-17.42	5.78	0.003	23.6%
	Linear (Year)	0.19	0.06	0.003	

The final Proc Traj model for the entire population can be summarized in a system of two equations, one for each identified group, where:

$Y_{it}$  = the dyspnea for subject  $i$  at time,  $t$  and,

$P^1(Y_{it} = 1)$  = the probability of  $Y_{it}$  given membership in group one.

Resulting in two equations:

$$P^1(Y_{it} = 1) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$P^2(Y_{it} = 1) = \frac{e^{\beta_0 + \beta_1(\text{Year}_{it})}}{1 + e^{\beta_0 + \beta_1(\text{Year}_{it})}}$$

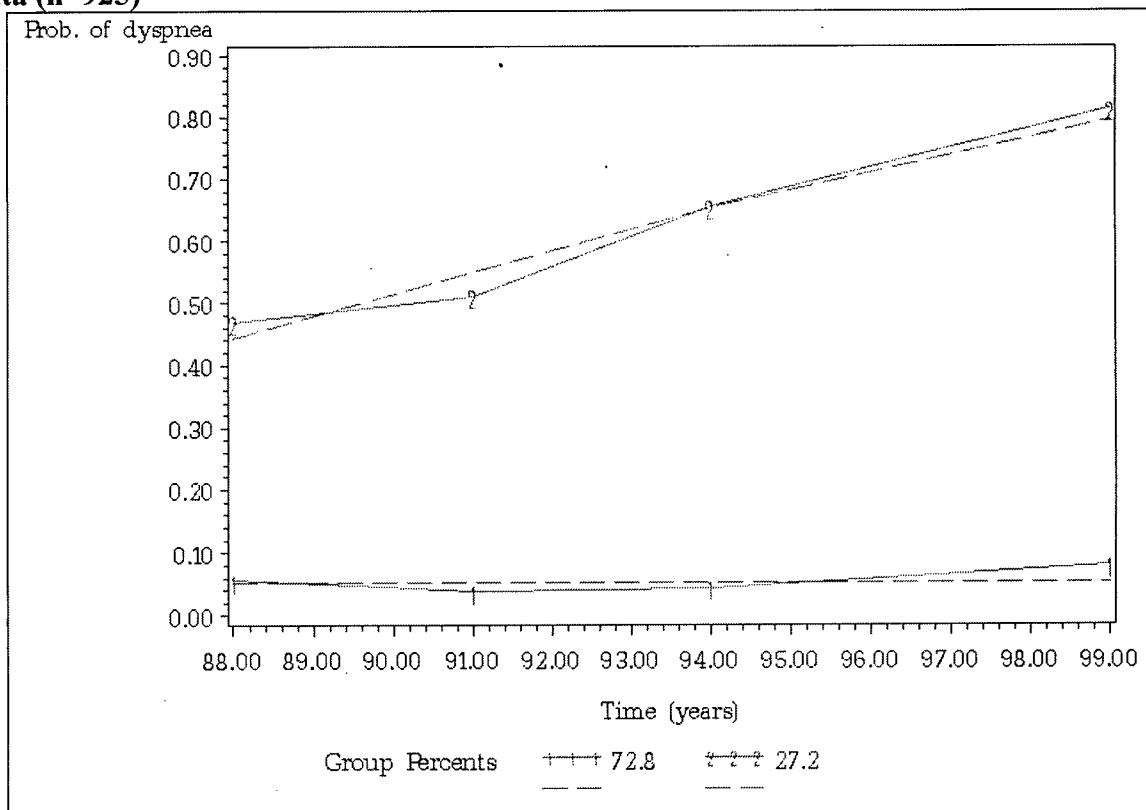
Inputting the parameter estimates from the model gives:

$$P^1(Y_{it} = 1) = \frac{e^{-2.92}}{1 + e^{-2.92}}$$

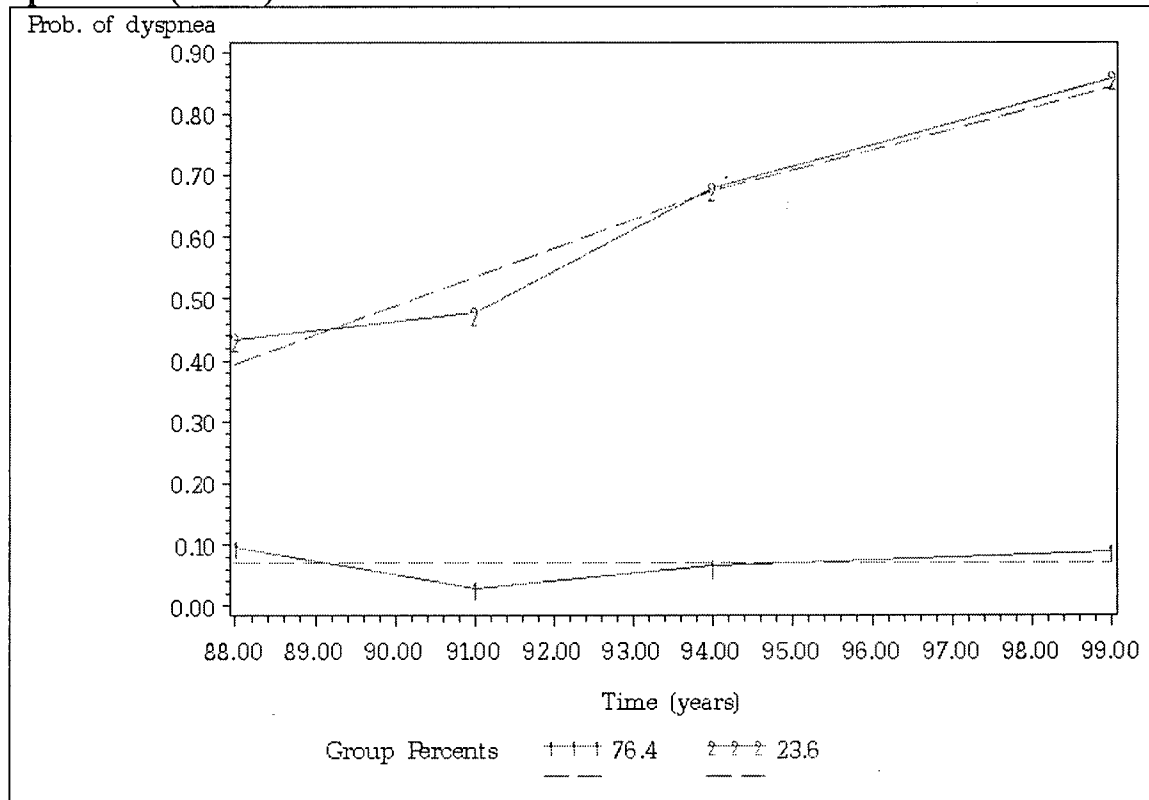
$$P^2(Y_{it} = 1) = \frac{e^{-12.96 + 0.14(\text{Year}_{it})}}{1 + e^{-12.96 + 0.14(\text{Year}_{it})}}$$

A zero order equation, or a horizontal straight line approximates group one. Subjects in this group have a stable risk (or probability) of reporting dyspnea. A straight line, or a linear first order equation approximates group two. Subjects in group two have an increasing risk of reporting dyspnea as follow-up progresses. These patterns of change over time can be observed in Figure 4 and Figure 5. In these figures, there are two lines visible for each group; the dashed line indicates the predicted trajectory described by the estimated regression coefficients and the solid line describes the average probability of reporting dyspnea at each measured time point for each group.

**Figure 4 Graphical output from Proc Traj showing final two-group model with group membership probability and shape of change over time using all subjects' data (n=925)**



**Figure 5 Graphical output from Proc Traj showing final two-group model with group membership probability and shape of change over time using men with complete data (n=148)**



Using the posterior group assignments from the output data set descriptive analysis can be completed to identify any risk factors associated with group assignment. Risk factors from the Fixed Effects and Proc Glimmix models were investigated by group assignment and results are shown in Table 35. Factors that may be related to membership in Group 2 include being female, older age, lower FEV1 percent predicted. Working in a job with exposure to respiratory irritants seemed to increase the probability of belonging to Group 1 (less dyspnea). These significant predictors are similar to the factors identified in the both fixed and mixed effects models.

**Table 35 Description of subgroups using posterior group assignments from Proc Traj output dataset (entire dataset, n=925)**

	Group 1				Group 2			
	n		%		n		%	
Overall	680		73%		245		27%	
	Men				Women			
	Group 1		Group 2		Group 1		Group 2	
	n	%	n	%	n	%	n	%
Frequency	624	76%	200	24%	56	55%	45	45%
Never Smoker	200	32%	48	24%	15	27%	12	27%
Former Smoker	258	41%	82	41%	15	27%	17	38%
Current Smoker	166	27%	70	35%	26	46%	16	36%
Childhood Asthma, no	600	96%	185	92%	53	95%	43	96%
Childhood Asthma, yes	24	4%	15	8%	3	5%	2	4%
Atopy, no	387	62%	125	63%	38	68%	36	80%
Atopy, yes	213	34%	64	32%	16	29%	9	20%
2 visits	290	46%	109	55%	32	57%	28	62%
3 visits	218	35%	59	30%	20	36%	13	29%
4 visits	116	19%	32	16%	4	7%	4	9%
Low Asbestos Exposure	207	33%	65	32%	50	89%	40	89%
High Asbestos Exposure	417	67%	135	68%	6	11%	5	11%
No Irritant Exposure	163	26%	79	40%	49	88%	43	96%
Irritant Exposure	461	74%	121	61%	7	12%	2	4%
Age (mean, SD)	44.8	9.8	49	9.1	44.4	10.4	49.8	10.9
FEV1 (mean, SD)	3.91	0.76	3.41	0.75	2.79	0.49	2.56	0.52
FEV1 percent predicted (mean, SD)	99.0	12.6	89.6	15.0	97.8	13.6	94.3	15.1

In addition, the categories of dyspnea change over time are shown stratified by group assignment and sex (Table 36). As expected, subjects never reporting dyspnea are members of Group 1, while subjects persistently reporting dyspnea are Group 2 members. The majority of Group 2 members (men and women) are subjects who developed dyspnea. Interestingly, the majority of subjects who resolved dyspnea or reported variable dyspnea were also assigned to Group 2, the group with increasing dyspnea over time.

**Table 36 Category of dyspnea change over time stratified by sex and group membership**

	Men				Women			
	Group 1		Group 2		Group 1		Group 2	
	n	%	n	%	n	%	n	%
Never Report Dyspnea	564	91%	0	0%	52	93%	0	0%
Always Report Dyspnea	0	0%	56	28%	0	0%	12	27%
Resolved Dyspnea	24	4%	29	14%	2	4%	12	27%
Developed New Symptom	13	2%	91	45%	1	2%	19	42%
Variable Dyspnea	20	3%	26	13%	1	2%	2	4%

## 4.6 Discussion

The use of categories to describe a pattern of change over time is a logical approach when there are only two time points because all possible patterns of change can be effectively captured. But when more time points are being considered, this crude categorization can potentially lead to biased results.

The results from the descriptive analysis of marine worker population highlight the finding that the more visits a subject completes, the more likely they are to report a symptom change during follow-up. This may be due to the fact that more visits means that the subject has a longer follow-up and that the probability of reporting a change in symptoms increases over time. Of more concern would be if the subject was responding to the research study itself and the experience of participating was increasing the probability of reporting a change in symptoms.

It is essential to note that when using more than two time points with the classification strategy, the fifth category (which was labeled 'Variable' in this analysis) is, by definition, limited to subjects who have greater than two visits. A subject with only two visits will slot into one of the four categories previously reported in the literature.

The frequency of symptom change categories in a population could also be affected by the baseline symptom prevalence rates. If a population reports very few symptoms at baseline there are few subjects "available" to resolve a symptom and thus a lower probability of subjects ending up in the 'Resolution' category. Vice versa, if the



population has very high symptom prevalence at baseline, there will be few subjects that can logically develop a new symptom during follow-up and the frequency of the 'New' category will be low. For these reasons, the baseline symptom prevalence should be considered when analyzing and comparing studies using a crude categorization method.

When there are more than two time points not only does the number of possible patterns increase, the timing of these changes over time will likely vary between individuals (i.e. some individuals may develop a new symptom early in follow-up while another may not develop a new symptom till the end of follow-up). As the length of follow-up increases the temporal location of respiratory symptom change will become variable and perhaps more critical to the research questions; this cannot be captured by a crude categorization strategy demonstrated here.

#### **4.6.1 Traditional Fixed Effects- Proc Logistic**

Results from the fixed effects models using Proc Logistic indicated that when using only baseline data (each subjects' first visit date) older age, smoking and current exposure to respiratory irritants are significant predictors of dyspnea. After adjusting for FEV1 percent predicted all predictors remained significant.

However, when a 'faulty' model is constructed using all subjects' data from all visits (ignoring autocorrelation) more significant predictors of dyspnea are identified. Older age, being female, history of childhood asthma smoking current exposure to respiratory irritants and the number of complete visits are all significant predictors. When FEV1 percent predicted is accounted for, only smoking ceases to remain a significant predictor.

#### **4.6.2 Proc Glimmix**

Results from two Proc Glimmix models suggested that the probability of reporting dyspnea did change over time. First, the marginal model provided correlation coefficients for each pair of visit dates and results showed correlation coefficients of less than 0.5 for all comparisons suggesting that the responses were not highly correlated from year to year. Second, mixed models were used to demonstrate the change over time in dyspnea reporting. The first mixed model included visit date as a covariate and a random intercept resulted in different odds ratios for reporting dyspnea at each visit date

and the odds of reporting dyspnea in 1999 (the last visit year) were significantly higher than reporting dyspnea in 1988 (the first visit year). The mixed models also showed the same results using a saturated model where each visit year was input as a dummy variable and no intercept was modeled. This model resulted in similar odds ratios for the first three visit years (1988, 1991, 1994) and a significantly larger odds ratio in 1999. The saturated model better facilitated comparison between visit years and also demonstrated that the probability of reporting dyspnea was low at all visits. The final mixed model included time as a continuous variable (calendar year) and a random intercept. Again, results indicated that probability of reporting dyspnea increased over time, approximately 7% per increase of one calendar year.

Proc Glimmix also modeled the predictors of reporting dyspnea at any point in time including a random intercept term (Table 20). Results from this model showed similar results as the fixed effects models that were completed. Older age, being female, childhood asthma, smoking and current exposure to respiratory irritants were significant predictors of dyspnea. As in the fixed models, after adjusting for FEV1 percent predicted, smoking was no longer a significant predictor of dyspnea but all other covariates remained significant. These results indicate that smoking is a risk factor only because it is related to having lower than expected lung function, whereas other risk factors, such as being female, are independent risk factors for reporting dyspnea irrespective of lung function.

The estimated variance component of the random intercept included in all the Proc Glimmix models was consistent in all models (range 1.33-1.37) and in all cases was statistically different from zero. This indicates that there was significant amount of variance in the intercept that was not accounted for by the fixed effects parameters.

Proc Glimmix models failed to converge when modeling the correlation structure between repeated measures of dyspnea using all of data (n=925). Despite the ability of generalized linear mixed models to handle missing data, the lack of convergence is likely due to the missing data in the population. When the dataset was limited to male subjects with complete data (n=148) the model converged and provided an estimated correlation matrix.

With the Proc Glimmix models it is difficult to infer anything about the shape or pattern of change over time in the outcome. It is a powerful tool for determining the predictors of an outcome by accounting for the autocorrelation in the repeated measures, but does not model change in the outcome with ease.

#### **4.6.3 Proc Traj**

Results from Proc Traj models indicated that there were two distinct patterns of change over time in the probability of reporting dyspnea in the study population. One group had a constant low level probability of reporting dyspnea and included 73% of the population. The second group had a linearly increasing probability of reporting dyspnea and represented 27% of the population.

Using the posterior group assignments, descriptive analysis showed that being female, older and lower FEV1 were associated with membership in Group 2. Current work in an area with exposure to respiratory irritants was associated with membership in Group 1, in other words it was associated with less dyspnea. The predictors of group membership identified from Proc Traj models were are very similar to the significant predictors identified using Proc Glimmix.

It is important to remember that the groups identified in the Proc Traj models are not real groups, they are approximations of patterns of change over time. The post-hoc analysis using the posterior group assignments (Table 35) would seem to go against this assertion, but as long as there is awareness of this limitation, potential relationships between group membership and risk factors can be explored using the group assignments. It is also possible to include risk factors in the model fitting stage so that the risk factors themselves impact the model selection process.

The strength of Proc Traj is that it uses all of the data collected on the outcome and risk factors (if included). Proc Traj models would allow for the inclusion of time varying or time stable covariates.

Unlike in the Proc Glimmix models, there were no convergence problems when using Proc Traj with the complete data set (n=925) that included missing data. However when the data set was limited to men with complete data (n=148) the models with larger

numbers of groups and higher order equations would not converge. This is most likely due to a lack of power, because as the number of groups increased and the order of the trajectory shape increased the number of parameters estimated increases and the sample size for these models was relatively small (n=148).

The results from Proc Traj suggested only a two group model of change in dyspnea over time. These results are quite different than the four groups identified in previous studies using the categorization method. This may be due in part to a smaller than necessary sample size, the proportion of missing data in the population or the small number of repeated measures. However, it may also be due to that fact that the simple categorization of subjects based on two time points ignores aspects of the data. Proc Traj included all subjects' visits and placed each visit at the appropriate visit year in the model; the categorization method does neither of these. The inclusion of time into the model may have complicated the modeling process because subjects were not all starting and ending at the same point in time.

Following the initial analysis, Proc Traj models were run to test whether the number of visits or the timing of visits affected the model outcome. When only male subjects with 4 visits (complete data) were included the same two group model was supported. A model excluding subjects with no dyspnea ever supported a one group model, where the trajectory was linearly increasing. This one group model was strikingly similar to the Group2 trajectory in the reported in the Proc Traj results (Figure 4 and Figure 5). Additionally, a Proc Traj model was run using visit1, visit2, visit3 and visit4 as the time variable (as opposed to calendar year) to determine whether the timing of subjects' visits was affecting the model. Again, the model supported a two group model similar to the results reported. Together, these additional models support the robustness of the reported Proc Traj model.

One drawback of Proc Traj is its inability to include random effects in the models. The result is that each subject is assumed to follow the estimated trajectory for the group they are assigned to. This is clear from the posterior group assignments that slot each subject into one of the estimated groups. A random effect would allow for estimation of variation around the estimated trajectories so that individuals were not expected to

exactly follow the estimated trajectory, but rather to follow a variation on the estimated trajectory.

Despite this drawback, the ability of Proc Traj to model multiple patterns of change and to model different shapes of change over time within the same model is beneficial. The result allows for the identification of discrete group and also the description of the unique patterns of change between these groups. This is in stark contrast to a normal regression or growth model where only one population mean is estimated and no inference about distinct subgroups can be made.

#### **4.6.4 Relevance to Previous Literature**

When using the categorization method, the frequencies of patterns of dyspnea change over time in the marine workers cohort were similar to previous studies. The never reporting dyspnea group was the largest, and was larger for men compared to women.

The 1993 paper by Sherrill et al (27) is a good example of how a mixed effects model can be applied to pulmonary function data. In the case of Sherrill (27), the outcome was a continuous lung function variable; no previous literature has applied these models to a binary outcome such as dyspnea.

The papers by Jedrychowski, Jaakkola and Brodtkin each used a fixed effects linear regression to model the change in lung function parameter (FEV1 or FVC) over time (19-21). Each of these studies is an example of where a mixed effects regression could have been applied. The use of mixed effects regression would have removed the requirement of running an individual linear regression for each subject. Instead, all the repeated lung function measures could have been used as the outcome variable. Then the outcome variable would be the exact measures, rather than an estimated coefficient with an ignored error. This approach would better account for the correlation between the repeated measures than the individual linear regression.

Studies using the pattern of symptom change over time (18-22) as a predictor, or those interested in symptom change as an outcome (7, 11, 28-30) may have benefited from the use of Proc Traj to describe the different patterns of change over time in their

populations. In addition, the group membership information could have been used as a predictor of lung health outcomes.

## **4.7 Conclusions**

In contrast with previous published analyses where data was excluded and subjects were categorized based on their symptom responses, this case study used dyspnea as an outcome in analyses using Proc Traj and Proc Glimmix. Proc Traj and Proc Glimmix remove this need to categorize subjects and at the same time use all of the collected data. These models also allow for the inclusion of time varying covariates, again removing the need for categorization or exclusion of covariate data.

When interpreting these results it is important to remember that the models are investigating a respiratory symptom that was self-reported by subjects. Compared to other binary data, respiratory symptoms may be more subjective and have more random variability. This may result in more random error and less systematic change over time.

Both Proc Traj and Proc Glimmix were able to provide information on whether dyspnea was changing over time in the study population as well as the predictors of dyspnea. However, Proc Traj is best suited for describing multiple patterns of change over time while Proc Glimmix is best suited for answering questions about the predictors of the outcome at a population level. These techniques provide valuable information about respiratory symptoms and how they change over time.

## 5 Perspectives

The work presented in this thesis has demonstrated that previous literature did not take full advantage of collected data, particularly in reference to respiratory symptom data resulting from the ATS questionnaire. Few studies have considered symptoms as an outcome in analyses. Some have incorporated symptoms as a predictor of lung function using a categorical variable based on two time points, ignoring significant amounts of data.

Two alternative methods for exploring respiratory symptoms as an outcome and using all of the data collected in longitudinal studies were reviewed in Chapter 3: Proc Traj and Proc Glimmix, also called discrete mixture models and a generalized linear mixed models. These models both handle longitudinal repeated measures data and permit the inclusion of time varying covariates so that the repeated nature of all of the data is considered. Proc Traj and Proc Glimmix both construct regression models to describe the data, however they permit inferences about different research questions.

Proc Traj is a special case of a mixture model. Proc Traj identifies multiple distinct subgroups within the population and models the change over time (in the outcome) within each subgroup. The result is information about multiple groups within the same population. Proc Traj provides information on the number of groups, the shape of their change over time (with respect to the outcome variable), what risk factors are associated with membership in different groups and the role of covariates in the pattern of change over time.

Proc Glimmix is a mixed regression model procedure that fits a single regression equation to the data. Proc Glimmix allows for the inclusion of random effects. The results from a mixed model (which include random effects by definition) allow inferences about the population trends with respect to change over time, and inferences about the predictors of the outcome variable. In addition, the deviation around the group mean is estimated for each included random variable. These results describe subject level variation in the effect of the random variable on the outcome.

Results from the case study of marine transportation workers indicated that the prevalence of dyspnea change over time categories was similar to those in previously published literature. But further analysis of the relationships between these categories and covariates suggested that the categories might be introducing bias into the results; within dyspnea change over time categories, subjects were not distributed equally based on of their number of visits. Also, the bias resulting from using only two data points was highlighted. When subjects' data from intermediate visits is ignored, observed symptom changes might be missed resulting in misclassification.

Results from both Proc Traj and Proc Glimmix models indicated that dyspnea changed over time. Proc Traj identified two patterns of dyspnea change over time; one group with a consistently low risk of reporting of dyspnea; and another group with a linearly increasing risk of reporting dyspnea. Proc Glimmix confirmed that dyspnea was changing over time in four separate models. The first model estimated the correlation between visits year and results suggested that the responses at each visit were not highly correlated. The second model a produced odds ratio estimates that indicated the probability of reporting dyspnea increased over time, and was significantly higher at the final visit compared with the first visit. The third model compared odds ratios between visit years (rather than to a reference group) and again showed that the odds of reporting dyspnea in 1999 were greater than in all other years. The final model used year as a continuous variable and indicated that for each increase of one calendar year, the probability of reporting dyspnea increased by seven percent.

Proc Glimmix models also provided evidence of several personal risk factors for reporting dyspnea: older age, being female and a history of childhood asthma were all associated with reporting dyspnea. Working in a current job where there was exposure to respiratory irritants was associated with less dyspnea, perhaps suggesting a healthy worker effect. The most notable result was that being female was an independent risk factor for dyspnea irrespective of lung function. Exploration of potential risk factors using the posterior group assignments from the final Proc Traj model identified the same risk factors as the Proc Glimmix model, but without any measure of statistical significance.



Proc Glimmix is ideally used when the researcher is interested in the group mean and the subject level variation around the single group mean. There are many situations where this may be appropriate, such as studies of childhood growth where individuals are all expected to increase in height over time. But, when thinking of respiratory symptoms, it is hard to assume that there is a single group mean that could adequately describe the trend in change over time. Proc Glimmix models are also useful when the researcher is interested in the predictors of a particular outcome in a population, for example, the predictors of reporting dyspnea.

When knowledge indicates that there are multiple patterns of change over time in the outcome, Proc Traj is a technique worth investigating. Proc Traj identifies and models multiple distinct pattern of change over time in the population. In the case of respiratory symptoms Proc Traj was a logical choice because previous literature had shown that individuals experience changes in symptom in different directions (gaining a symptom, or resolving a symptom). Proc Traj also incorporates covariates into the model and the impact of each covariate on the pattern of change over time can be determined.

Both Proc Traj and Proc Glimmix are suitable for research questions relating to longitudinal respiratory symptom data. Proc Glimmix is best for modeling the predictors of reporting a symptom at the population level, while Proc Traj is suited for modeling multiple subgroups in the population and their patterns of change over time. Proc Glimmix models an overall population mean, but the inclusion of random effects permits further inference about individual level differences. Proc Traj is limited to modeling the mean of each identified subgroups, with no subject level inference possible. A simple guide for applying Proc Traj and Proc Glimmix in studies of respiratory symptoms has been compiled and is located in Appendix A.

## **5.1 Strengths**

One of the main strengths of this thesis is that it highlighted issues with how respiratory epidemiologists currently handle symptom data. This thesis outlined the problems with categorizing subjects based on responses and the problems with ignoring collected data. Going further, this thesis reviewed and evaluated two potential methods for improving the way we handle longitudinal respiratory symptom data. Neither of these

two methods appears to have been used to analyze respiratory symptom data as an outcome in the peer-reviewed literature.

The exploratory results from these two methods, Proc Traj and Proc Glimmix, provide a beginning for future research. The results emphasize the strengths and limitations of each technique and, in combination with Appendix A, will help to guide other researchers in future studies of respiratory symptom data.

The analysis presented in this thesis was strengthened by the data available for exploring the research questions of interest. The size of the study population, the presence of multiple repeated measures of respiratory symptoms and the detailed information on personal risk factors and lung health measures were crucial for the case study analysis.

Overall, the results from this thesis contributed to the hypothesis that patterns of respiratory symptom change over time are important. Results demonstrated that there were risk factors for reporting dyspnea (independent of lung function) and that there were two distinct patterns of dyspnea change over time in the study population.

## **5.2 Limitations**

One limitation of this thesis is that it was not a theory based statistical evaluation. The statistical concepts and theories behind each of the methods investigated were not reviewed. Had a statistician undertaken this thesis, the evaluation of the methods may have been more thorough with regard to statistical theory, but the relevance to respiratory symptom data may have suffered. It is hoped that the additional relevance to our research area will make it easier for further research to flow from this work.

Missing data in the case study data may have resulted in some bias in the results. The data set employed for the exploratory analysis was from a large longitudinal occupational surveillance study. This provided a rich data source, but there was a significant amount of missing data resulting from subjects missing visit dates.

There are also limitations inherent to each of the statistical techniques evaluated. Proc Glimmix may produce biased coefficient estimates under some circumstances. This bias

can be determined with the assistance of a statistician, but this does limit the applicability of the procedure for respiratory epidemiologists.

Proc Traj provides estimates of the group mean for multiple groups that are identified within the population but these groups can be misleading. The estimated groups are not actual entities and should not be treated as absolute; they are estimations of multiple patterns of change. The fact that Proc Traj does not include random effects limits the ability of the procedure to make inferences about individual level change over time.

### **5.3 Future Research**

Future research should begin by constructing more complex models using Proc Traj and Proc Glimmix to further explore the patterns of change over time and the relevant personal risk factors. Researchers interested in this area should begin to formulate research questions specific to longitudinal respiratory symptoms and to design studies with these questions in mind. Appendix A should serve as a valuable reference to researchers for this purpose.

Now that two methods have been identified as useful tools for studying respiratory symptoms, perhaps other existing (maybe more complex) statistical methods can be identified to better study patterns of change over time. Respiratory epidemiologists should continue to work with statisticians to explore better approaches to analyzing complex data.

As with any research findings, it would be ideal to repeat the models presented in this thesis on another data set to determine whether the findings regarding dyspnea are consistent between populations. Other data sets may also have more complete data or contain information on risk factors that were not included in the marine workers study population (i.e. exposure measurements, physician visit data). The use of a data set with fewer missing data points and more information on potential risk factors will only further our understanding of respiratory symptoms, how they change over time and their relationship with lung health outcomes.

## References

- (1) Ferris BG. Epidemiology Standardization Project (American Thoracic Society). *Am Rev Respir Dis* 1978;118(6 Pt 2):1-120.
- (2) Malo J-L, Chan-Yeung M, Kennedy S. Occupational Asthma. In: Barnes P, Drazen J, Rennard S, Thompson N, editors. *Asthma and COPD: Basic Mechanisms and Clinical Management*. Amsterdam: Academic Press; 2002.
- (3) Chan-Yeung M, Malo JL. Occupational asthma. *N Engl J Med* 1995;333(2):107-12.
- (4) SAS Institute. *The GLIMMIX Procedure*: SAS Publishing; 2006.
- (5) Jones BL, Nagin DS, Roeder K. A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research* 2001;29(3):374-393.
- (6) Brinkman GL, Block DL, Cress C. Effects of bronchitis and occupation on pulmonary ventilation over an 11-year period. *J Occup Med* 1972;14(8):615-20.
- (7) Kongerud J, Samuelsen SO. A longitudinal study of respiratory symptoms in aluminum potroom workers. *Am Rev Respir Dis* 1991;144(1):10-6.
- (8) Schwartz DA, Davis CS, Merchant JA, Bunn WB, Galvin JR, Van Fossen DS, et al. Longitudinal changes in lung function among asbestos-exposed workers. *Am J Respir Crit Care Med* 1994;150(5 Pt 1):1243-9.
- (9) Beeckman LAF, Wang ML, Petsonk EL, Wagner GR. Rapid declines in FEV1 and subsequent respiratory symptoms, illnesses, and mortality in coal miners in the United States. *American Journal Of Respiratory And Critical Care Medicine* 2001;163(3):633-639.
- (10) Wolf C, Pirich C, Valic E, Waldhoer T. Pulmonary function and symptoms of welders. *Int Arch Occup Environ Health* 1997;69(5):350-3.
- (11) Carta P, Aru G, Barbieri MT, Avataneo G, Casula D. Dust exposure, respiratory symptoms, and longitudinal decline of lung function in young coal miners. *Occup Environ Med* 1996;53(5):312-9.
- (12) Preller L, Heederik D, Boleij JS, Vogelzang PF, Tielen MJ. Lung function and chronic respiratory symptoms of pig farmers: focus on exposure to endotoxins and ammonia and use of disinfectants. *Occup Environ Med* 1995;52(10):654-60.
- (13) Boezen HM, Schouten JP, Postma DS, Rijcken B. Relation between respiratory symptoms, pulmonary function and peak flow variability in adults. *Thorax* 1995;50(2):121-6.

- (14) James AL, Cookson WO, Buters G, Lewis S, Ryan G, Hockey R, et al. Symptoms and longitudinal changes in lung function in young seasonal grain handlers. *Br J Ind Med* 1986;43(9):587-91.
- (15) Larsson ML, Loit HM, Meren M, Polluste J, Magnusson A, Larsson K, et al. Passive smoking and respiratory symptoms in the FinEsS Study. *Eur Respir J* 2003;21(4):672-6.
- (16) Fletcher CM, Elmes PC, Fairbairn AS, Wood CH. The significance of respiratory symptoms and the diagnosis of chronic bronchitis in a working population. *Br Med J* 1959;5147:257-66.
- (17) Jakeways N, McKeever T, Lewis SA, Weiss ST, Britton J. Relationship between FEV1 reduction and respiratory symptoms in the general population. *Eur Respir J* 2003;21(4):658-63.
- (18) Sharp JT, Paul O, McKean H, Best WR. A longitudinal study of bronchitic symptoms and spirometry in a middle-aged, male, industrial population. *Am Rev Respir Dis* 1973;108(5):1066-87.
- (19) Jedrychowski W, Krzyzanowski M, Wysocki M. Are chronic wheezing and asthma-like attacks related to FEV1 decline? The Cracow Study. *Eur J Epidemiol* 1988;4(3):335-42.
- (20) Jaakkola MS, Jaakkola JJ, Ernst P, Becklake MR. Respiratory symptoms in young adults should not be overlooked. *Am Rev Respir Dis* 1993;147(2):359-66.
- (21) Brodtkin CA, Barnhart S, Checkoway H, Balmes J, Omenn GS, Rosenstock L. Longitudinal pattern of reported respiratory symptoms and accelerated ventilatory loss in asbestos-exposed workers. *Chest* 1996;109(1):120-6.
- (22) Krzyzanowski M, Camilli AE, Lebowitz MD. Relationships between pulmonary function and changes in chronic respiratory symptoms. Comparison of Tucson and Cracow longitudinal studies. *Chest* 1990;98(1):62-70.
- (23) Krzyzanowski M, Lebowitz MD. Changes in chronic respiratory symptoms in two populations of adults studied longitudinally over 13 years. *Eur Respir J* 1992;5(1):12-20.
- (24) Krzyzanowski M, Robbins DR, Lebowitz MD. Smoking cessation and changes in respiratory symptoms in two populations followed for 13 years. *Int J Epidemiol* 1993;22(4):666-73.
- (25) Christiani DC, Wang XR, Pan LD, Zhang HX, Sun BX, Dai H, et al. Longitudinal changes in pulmonary function and respiratory symptoms in cotton textile workers. A 15-yr follow-up study. *Am J Respir Crit Care Med* 2001;163(4):847-53.

- (26) Wang XR, Pan LD, Zhang HX, Sun BX, Dai HL, Christiani DC. Follow-up study of respiratory health of newly-hired female cotton textile workers. *Am J Ind Med* 2002;41(2):111-8.
- (27) Sherrill DL, Lebowitz MD, Knudson RJ, Burrows B. Longitudinal methods for describing the relationship between pulmonary function, respiratory symptoms and smoking in elderly subjects: the Tucson Study. *Eur Respir J* 1993;6(3):342-8.
- (28) Pahwa P, Senthilselvan A, McDuffie HH, Dosman JA. Predictors of onset of wheezing in grain elevator workers. *Can Respir J* 1998;5(3):200-5.
- (29) Boutet K, Malo JL, Ghezze H, Gautrin D. Airway hyperresponsiveness and risk of chest symptoms in an occupational model. *Thorax* 2006.
- (30) Gunnbjornsdottir MI, Franklin KA, Norback D, Bjornsson E, Gislason D, Lindberg E, et al. Prevalence and incidence of respiratory symptoms in relation to indoor dampness: the RHINE study. *Thorax* 2006;61(3):221-5.
- (31) Mahler DA, Tomlinson D, Olmstead EM, Tosteson AN, O'Connor GT. Changes in dyspnea, health status, and lung function in chronic airway disease. *Am J Respir Crit Care Med* 1995;151(1):61-5.
- (32) Lareau SC, Meek PM, Press D, Anholm JD, Roos PJ. Dyspnea in patients with chronic obstructive pulmonary disease: does dyspnea worsen longitudinally in the presence of declining lung function? *Heart Lung* 1999;28(1):65-73.
- (33) Hodgev VA, Kostianev SS, Torosian AA, Yanev IB, Mandoulova PB. Long-term changes in dyspnea, lung function, and exercise capacity in COPD patients. *Folia Med (Plovdiv)* 2004;46(3):12-7.
- (34) Wu J, Kreis IA, Griffiths D, Darling C. Respiratory symptoms and lung function of coke oven workers: a lung function surveillance system from 1990-2000. *J Occup Environ Med* 2004;46(9):906-15.
- (35) Jeansonne A. Loglinear Models. September 26, 2002 Retrieved on October 14, 2006 from <http://userwww.sfsu.edu/~efc/classes/biol710/loglinear/Log%20Linear%20Models.htm>.
- (36) Garson D. Univariate GLM, ANOVA, and ANCOVA. Retrieved on September 24, 2006 from <http://www2.chass.ncsu.edu/garson/PA765/anova.htm>.
- (37) Zeger SL, Liang KY, Albert PS. Models For Longitudinal Data - A Generalized Estimating Equation Approach. *Biometrics* 1988;44(4):1049-1060.
- (38) Mustillo S, Worthman C, Erkanli A, Keeler G, Angold A, Costello EJ. Obesity and psychiatric disorder: Developmental trajectories. *Pediatrics* 2003;111(4):851-859.

- (39) Schabenberger O. SUGI Paper 196-30 Introducing the GLIMMIX Procedure for Genrealized Linear Mixed Models. Statistics and Data Analysis: SAS User Group.
- (40) Redpath SM, Leckie FM, Arroyo B, Amar A, Thirgood SJ. Compensating for the costs of polygyny in hen harriers *Circus cyaneus*. Behavioral Ecology And Sociobiology 2006;60(3):386-391.
- (41) Mendell MJ, Fisk WJ, Deddens JA, Seavey WG, Smith AH, Smith DF, et al. Elevated symptom prevalence associated with ventilation type in office buildings. Epidemiology 1996;7(6):583-9.
- (42) Littell R, Milliken G, Stroup W, Wolfinger R, Schabenberger O. SAS for Mixed Models. 2nd ed. Cary NC: SAS Institute Inc.; 2006.
- (43) Bauer D. Advanced Topics in Fitting Mixed Models for Hierarchical Data Structures. In: Summer Programme in Data Analysis (SPIDA) Lecture Notes; 2005.
- (44) MacNab Y. Personal Communication with Arrandale VH. November 8, 2006.
- (45) Jones BL, Nagin DS. Advances in Group-based Trajectory Modeling and a SAS Procedure for Estimating Them. submitted 2005.
- (46) Nagin DS. Analyzing developmental trajectories: A semiparametric, group-based approach. Psychological Methods 1999;4(2):139-157.
- (47) Nagin DS. Group-based Modeling of Development. Cambridge, Massachusetts: Harvard University Press; 2005.
- (48) Nishimura K, Izumi T, Tsukino M, Oga T. Dyspnea Is a Better Predictor of 5-Year Survival Than Airway Obstruction in Patients With COPD\*. Chest %R 10.1378/chest.121.5.1434 2002;121(5):1434-1440.
- (49) Frostad A, Soyseth V, Andersen A, Gulsvik A. Respiratory symptoms as predictors of all-cause mortality in an urban community: a 30-year follow-up. J Intern Med 2006;259(5):520-9.
- (50) Krzyzanowski M, Kauffmann F. The relation of respiratory symptoms and ventilatory function to moderate occupational exposure in a general population. Results from the French PAARC study of 16,000 adults. Int J Epidemiol 1988;17(2):397-406.
- (51) Fell AK, Thomassen TR, Kristensen P, Egeland T, Kongerud J. Respiratory symptoms and ventilatory function in workers exposed to portland cement dust. J Occup Environ Med 2003;45(9):1008-14.

- (52) Greaves IA, Eisen EA, Smith TJ, Pothier LJ, Kriebel D, Woskie SR, et al. Respiratory health of automobile workers exposed to metal-working fluid aerosols: respiratory symptoms. *Am J Ind Med* 1997;32(5):450-9.
- (53) American Thoracic Society. Standardization of spirometry: 1987 Update. *American Review of Respiratory Disease* 1987;136:1285-1298.
- (54) Crapo RO, Morris AH, Gardner RM. Reference spirometric values using techniques and equipment that meet ATS recommendations. *Am Rev Respir Dis* 1981;123(6):659-64.
- (55) SAS Institute. SAS/STAT 9.1 User's Guide: SAS Publishing; 2004.
- (56) Archives of SAS-L@LISTSERV.UGA.EDU. Retrieved on multiple occasions, from <http://listserv.uga.edu/archives/sas-l.html>.



## **Appendix A.     How to use SAS® Proc Traj and SAS® Proc Glimmix in Respiratory Epidemiology**

### **A.1 Introduction**

This document outlines the use of two procedures capable of modeling repeated respiratory symptom data in the software package SAS®: Proc Traj and Proc Glimmix.

SAS® Proc Traj is a discrete mixture model which models the patterns of change over time in multiple subgroups within the population. SAS® Proc Glimmix is a procedure that fits a generalized linear model to non-linear outcome data either with or without random effects.

### **A.2 Goal**

The goal of this document is to provide a concise user's guide for applying discrete mixture models (Proc Traj) and generalized linear mixed models (Proc Glimmix) in the analysis of longitudinal respiratory symptom data using SAS® software. This document does not attempt to describe the statistical theory behind either of these techniques.

### **A.3 How to use this document**

This document presents an outline for setting up models in both Proc Traj and Proc Glimmix for analyzing repeated respiratory symptom outcomes. Data organization is explained, the modeling procedure is outlined, the basic syntax (appropriate for binary respiratory symptom outcomes) is described and the relevant modeling possibilities are discussed.

This document should be a starting point for modeling using Proc Traj and Proc Glimmix. Readers are advised to refer to the SAS® documentation as well as the noted reference texts for further explanations and for confirmation that the models are appropriate for the data in use.

---

## **A.4 SAS® Trajectory Procedure**

### **A.4.1 Overview**

The SAS® Trajectory Procedure (Proc Traj) is a user-friendly finite mixture model procedure designed to run easily on the SAS® platform. Proc Traj is capable of fitting a discrete mixture model to the data so that multiple distinct subgroups within the population can be identified.

The focus of the Proc Traj model is on group membership and identifying distinct subgroups within the population. Proc Traj does not provide any individual level information on the pattern of change over time; subjects are grouped and it is assumed that every subject in the group follows the same trajectory. There is no random effect capability within the Proc Traj model.

The documentation for SAS® Proc Traj is a peer-reviewed publication by Jones, Nagin and Roeder (5) and is available only on B. Jones' website<sup>5</sup>. A follow-up article to the documentation has been submitted for publication and is also available in draft format on B. Jones' website. A recent text authored by D. Nagin (47) is a valuable reference for users of Proc Traj and should be reviewed by those interested in the statistical theory behind Proc Traj.

### **A.4.2 Requirements**

To apply Proc Traj to your data, you need (at a minimum) multiple measures of the outcome of interest and information on the timing of the repeated measures. It would also be helpful to have multiple measures on a number of covariates you are also interested in.

You must also download the Proc Traj application from B. Jones' website<sup>5</sup> and have copied the files to the folders as directed on the website.

---

<sup>5</sup> <http://www.andrew.cmu.edu/user/bjones/>

### A.4.3 Data Organization

In order to use Proc Traj you must organize your data in a multivariate, or “wide” format, where there is only one row of data for each subject and multiple observations included in one line of data. An example of data ready for use with Proc Traj is shown in Table 37, a description of each variable is provided in Table 38. You can see in Table 37 that the outcome variable “wheeze” is denoted by the variables Wez1, Wez2 and Wez3. These three variables correspond to three repeated measurements taken at three different times. The time at which each of these measurements was collected is represented by the variables Year1, Year2 and Year3. If a subject did not complete a visits, all variables corresponding to that visits are blank, in this case a “.” is used to indicate missing data.

**Table 37 Mock data set up for analysis with Proc Traj**

ID	Sex	Byr	Csmk01	Csmk02	Csmk03	Wez01	Wez02	Wez03	Yr01	Yr02	Yr03
001	0	1947	0	0	0	0	0	0	1992	1994	1999
002	1	1953	1	.	0	0	.	1	1992	1994	1999
003	0	1951	0	1	1	1	0	1	1992	1994	1999
004	0	1946	0	0	.	1	1	.	1992	1994	1999
005	1	1950	1	0	1	1	1	1	1992	1994	1999

**Table 38 Description of variables in mock data (Table 37)**

Variable Name	Description	Values
ID	Subject ID	as assigned
Sex	Sex of subject	0= male 1= female
Byr	Year of birth	continuous, in years
Csmk01	Current smoker at visit 1	0= never or former smoker
Csmk02	Current smoker at visit 2	1= current smoker
Csmk03	Current smoker at visit 3	
Wez01	Response to wheeze question at visit 1	0= no wheeze
Wez02	Response to wheeze question at visit 2	1= wheeze
Wez03	Response to wheeze question at visit 3	
Yr01	Date of visit 1	values corresponding to data
Yr02	Date of visit 2	
Yr03	Date of visit 3	

The variables that describe repeated measures of the same outcome must be numbered consecutively (i.e. csmk1, csmk2, csmk3 etc.) before Proc Traj will accept them; this will

usually require some recoding. SAS® will not accept the data if the variables are labeled alternatively (i.e. smk1992, smk1994, smk1999) even if this is logical given your data set. By identifying the variables that contain information on the date of each repeat measure (i.e. Yr1, Yr2, Yr3) you are specifying the space between repeated measures. Time varying covariates (i.e. Csmk1, Csmk2, Csmk3 for smoking information at each visit) must also be named with consecutive numbers corresponding to the visit.

#### **A.4.4 Dummy Variables**

It is advisable to create dummy variables for each of your covariates that you plan to input into a Proc Traj model, as was done for Current Smoking in Table 37. Covariates can be input in a binary (dummy) form or a continuous form but Proc Traj does not handle categorical covariates.

#### **A.4.5 Missing Data**

Proc Traj is able to handle data that is missing completely at random (MCAR), but is unable to handle data that is missing for more complex reasons (47). Missing data can be entered in the dataset as shown in Table 37.

#### **A.4.6 Types of Research Questions**

In terms of respiratory symptom data, Proc Traj should be used when your research question is similar to one of the following:

- Are there multiple patterns of change in the outcome?
- How many patterns of change are there in the outcome?
- What is the shape of the change over time?
- What predicts membership in each of these groups?
- What are the characteristics that differ (or are similar) between the different groups?

#### **A.4.7 Syntax**

The entire Proc Traj syntax is outlined on B. Jones' website<sup>5</sup> and should be referenced for any further questions regarding syntax.

A simple Proc Traj syntax for a two group model of the respiratory symptom wheeze is presented here:

```
proc traj data=a.mockdata out=out outstat=os outplot=op;
var wez01-wez03;
indep year01-year03;
model logit;
ngroups 2;
order 0 1;
id ID;
run;
%trajplot (OP, OS, "Title of graph", "Subtitle", "Y-axis
label", "X-axis label");
```

As in all SAS® procedures, the Proc Traj statement outlines that data set to be used and in this case also defines the output from the procedure.

The ‘var’ statement defines the binary symptom outcome of interest. ‘Indep’ defines the time variables that you are modeling the outcome over. The ‘model’ statement identifies the outcome as binary and the ‘ngroups’ states how many groups you want to model. ‘Order’ assigns the order of each equation that will describe the change over time in each group. ‘ID’ identifies the subjects in your population and also denotes which variable you want to use to uniquely assign subjects to a specific group in the output data set (in this case, “out”).

The ‘%trajplot’ is a macro statement that results in the graphical output from Proc Traj. This macro includes references to the outplot and outstat statements in the ‘Proc Traj’ statement. If you make any changes in the ‘Proc Traj’ statement be sure to adjust the trajplot macro accordingly.

When including time independent covariates into a Proc Traj model the ‘risk’ or ‘tcov’ statements will also be added to the syntax. For example, a time stable covariate for sex could be added:

```
proc traj data=a.mockdata out=out outstat=os outplot=op;
var wez01-wez03;
indep year01-year03;
```

```

model logit;
ngroups 2;
order 0 1;
risk female;
id ID;
run;
%trajplot (OP, OS, "Title of graph", "Subtitle", "Y-axis
label", "X-axis label");

```

Or, a time varying covariate for current smoking could be added:

```

proc traj data=a.mockdata out=out outstat=os outplot=op;
var wez01-wez03;
indep year01-year03;
model logit;
ngroups 2;
order 0 1;
tcov csmk01-csmk03;
id ID;
run;
%trajplot (OP, OS, "Title of graph", "Subtitle", "Y-axis
label", "X-axis label");

```

#### **A.4.8 Selecting the Best Model**

The model fitting procedure with Proc Traj is iterative and requires a priori decisions based on substantive knowledge. In the most basic process, the following steps should be followed:

1. Decide on the maximum number of groups using a priori knowledge
2. Fit number of groups to data (start by fitting a one group model, and then fit up to the maximum logical number of groups in a step wise manner)
3. Fit the shape of the trajectory for each group
4. Perform further modeling if required (addition of covariates, inclusion of second outcome etc.)

To decide on the optimum number of groups for your data you must begin by fitting a basic one group model with all groups set to a second order (quadratic) equation. Then fit a two group, then three group model etc. until you have fit the maximum number of groups based on your a priori decision. Nagin suggests setting all group orders to second order during this process (47).

For each model you fit in this first step you will be given two Bayesian Information Criterion (BIC) values in the output, one relates to the overall sample size (total number of observations) and the other relates to the subject sample size (number of subjects). The true BIC for the model lies between these values (47). The BIC is the log-likelihood adjusted for the number of parameters and the sample size (5). In the Proc Traj procedure the BIC values given in the output are negative; the best fit model is the one with the smallest negative number.

Model selection in Proc Traj uses the BIC to select the best fitting model via two different methods. The first, described by Jones, Roeder and Nagin (5) uses the change in the BIC between two models to measure the weight of evidence against the null model. For each increasingly complex model that is tested, the BIC of the more complex (larger number of groups, or higher order equation) less the BIC of the less complex model is used to select the model that better fits the data.

$$\Delta BIC = BIC_{(complex)} - BIC_{(null)}$$

The difference in BIC between the two models is a measure of the evidence against the null model. Jones, Nagin and Roeder (5) suggest criteria for strength of evidence against the null model (Table 39). Using the difference in the logged Bayes factor between successive models, the difference between the alternate and the null model can be qualified. The null model is always the simpler model (i.e. less groups, or lower order equations). The interpretation of the logged Bayes factor ( $2\Delta BIC$ ) in terms of model preference is shown in Table 39.

**Table 39 Interpretation of logged Bayes factor ( $2*\Delta BIC$ ) for model selection (Adapted from Table 2 in (5))**

$2*\Delta BIC$	Evidence against $H_0$
0 to 2	Not worth mentioning
2 to 6	Positive
6 to 10	Strong
> 10	Very Strong

The second method is called Jeffreys's scale of the evidence and is described by Nagin (47). Jeffreys's scale of the evidence uses the exponentiated difference between the BIC values of models,  $i$  and  $j$ :

$$\text{Bayes Factor} \approx e^{BIC_i - BIC_j}$$

In this case it does not matter which model is the null model; only that the researcher remembers which model is which. The interpretation of Jeffreys's scale of the evidence is outlined in Table 40. Further description and explanation can be found in Chapter 4 of Nagin (2005) (47).

**Table 40 Interpretation of Bayes Factor ( $e^{BIC_i - BIC_j}$ ) for model selection (Adapted from Table 4.2 in (47))**

Bayes Factor ( $B_{ij}$ )	Interpretation
$B_{ij} < 1/10$	Strong evidence for model $j$
$1/10 < B_{ij} < 1/3$	Moderate evidence for model $j$
$1/3 < B_{ij} < 1$	Weak evidence for model $j$
$1 < B_{ij} < 3$	Weak evidence for model $i$
$3 < B_{ij} < 10$	Moderate evidence for model $i$
$B_{ij} > 10$	Strong evidence for model $i$

When selecting the 'best' model it is important to base decisions on substantive knowledge about the research area, and remember the rule of parsimony to select the simplest model that best describes the data.

Again, in reference to the example with respiratory symptoms, if we tested five models (one group up to five groups) we would have five BIC values to review. The comparisons are completed in a step-wise manner so that the two-group model is



compared to the one-group model, and the three-group model to the two-group model and so on. In each case, the model with the smaller number of groups is the null model.

The next step in fitting a model using Proc Traj is selecting the shape of each group's trajectory over time. Proc Traj can model up to a fourth order polynomial and can model both linear and non-linear trajectories within the same model. This can be done using substantive knowledge (i.e. we expect one group to never report symptoms so this group's trajectory will be a zero-order equation, or a straight line) or it can be done using the  $\Delta BIC$ . It seems ideal to use a combination of substantive knowledge and statistical inference to make the decision regarding the shape of each group's trajectory.

#### **A.4.9 Output**

The output from Proc Traj includes the parameter estimates for each group (with standard errors), group membership probabilities (population level) and model fit statistics. The output data set (out= in 'Proc Traj' statement) includes all the variables included in the analysis (not all the variables in the original dataset), the variable identified in the 'id' statement, posterior subject specific group membership probabilities and a group assignment for each individual.

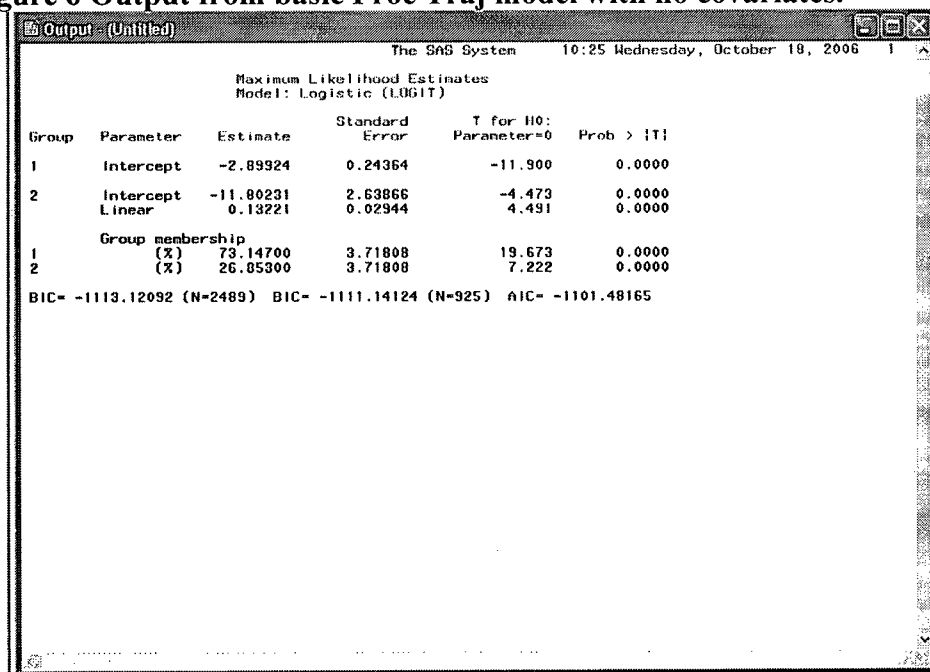
The parameter estimates can be used to construct regression equations for each group and a system of equations to describe the population. The relative differences between the estimates for the same covariate between groups can be used to make inferences about differences between the groups.

The posterior group membership probabilities and the group assignment variables in the output data set can be used to explore between group differences in covariates not included in the model and potentially as predictor variables in separate analyses. The posterior group probabilities are calculated for each individual based on the estimated parameters, and the individual is assigned to a group based on their highest posterior group probability (47).

The output from a basic model (no covariates) is shown in Figure 6. The intercept parameters represent the estimated intercept for each group. For Group 2 the linear parameter represents the estimated coefficients for the linear time component of the

regression equation. The group membership probabilities indicate what proportion of the population is estimated to belong to each group. And, the BIC values are the final portion of the output. Note the BIC values are shown for two sample sizes; first for all the data points and second for the number of subjects.

**Figure 6 Output from basic Proc Traj model with no covariates.**



Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
1	Intercept	-2.89924	0.24364	-11.900	0.0000
2	Intercept	-11.80231	2.63866	-4.473	0.0000
	Linear	0.13221	0.02944	4.491	0.0000
	Group membership				
1	(%)	79.14700	3.71808	19.673	0.0000
2	(%)	20.85300	3.71808	7.222	0.0000
BIC= -1113.12092 (N=2489) BIC= -1111.14124 (N=925) AIC= -1101.48165					

#### **A.4.10 User Information**

Base SAS® is required to run Proc Traj. You can download the procedure from the B. Jones' Proc Traj Home Page and copy the downloaded files into the appropriate folders on your hard drive (instructions provided on web page). Proc Traj will then be installed and functional in the SAS® platform.

Because Proc Traj is an add-on to SAS®, there is no formal SAS® documentation in the traditional SAS® format. Users are advised to thoroughly review the reference texts listed below.

#### **A.4.11 Cautions**

Researchers using Proc Traj are advised to remember that the multiple groups estimated are not reified groups. The identified groups are estimations of multiple

patterns of change within the population, and we must be careful not to think of group membership and trajectory shape as absolute certainties.

#### **A.4.12 Reference Texts**

Nagin, Daniel S. Group-based Modeling of Development. Harvard University Press: Massachusetts (2005).

Jones, B., Nagin, D., & Roeder, K. A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories. Sociological Methods & Research (2001) 29: 374-393.

SAS® Proc Traj Home. <http://www.andrew.cmu.edu/user/bjones>.

## **A.5 SAS® Glimmix Procedure**

### **A.5.1 Overview**

Generalized linear mixed models are a way to apply mixed models procedures to non-linear outcomes (i.e. binary, Poisson). Mixed models are models that incorporate both fixed effects and random effects in the model. Fixed effects are effects that are expected to have the same direction and magnitude of effect on each subject in a repeated measures study. Random variables are variables where the effect on the outcome is allowed to vary between subjects in the data set.

Mixed models are particularly useful in the modeling of longitudinal data because repeated measurements are collected over time on subjects and are inevitably correlated. A fixed effects model requires all of the measurements to be independent; in a longitudinal repeated measures data set this assumption is violated.

Within the SAS® program there are several procedure for constructing mixed models. The most common procedure is Proc Mixed, which models continuous outcomes. When the outcome is binary or count data, Proc Glimmix (general linear mixed models) should be employed. Proc Nlmixed (non-linear mixed models) can also be used to model binary, count or continuous outcomes but is primarily for use with advanced non-linear modeling and requires more programming. Both Proc Mixed and Proc Glimmix can be run using SAS® syntax.

This guide is focused on modeling binary respiratory symptom data using Proc Glimmix. Further information on Proc Glimmix (as well as Proc Mixed) can be found in the reference texts.

### **A.5.2 When to Use Mixed Effects**

In reference to longitudinal study designs, random effects should be introduced into a regression model when there are correlated outcome measures (repeated measures on individuals) and when you want to allow the effect of a particular covariate on the outcome to vary randomly among your subjects.

### A.5.3 Requirements

For a mixed model you should have outcome measures that you expect are correlated; this occurs when you have collected repeated measures on individuals. To use Proc Glimmix, you should also have a non-linear outcome variable, in this case binary symptom data. If you are using a continuous measure of a respiratory symptom, you should consult SAS® Proc Mixed.

### A.5.4 Data Organization

For Proc Glimmix models, data must be organized in a univariate, or “long”, format where there is one observation per line of data and multiple lines of data per subject. An example of data organized this way is shown in Table 41.

**Table 41 Mock data set up for analysis with Proc Glimmix**

ID	Age	Vyr	Vis2	Vis3	Sex	Fsmk	Csmk	Wez
001	53	1992	0	0	0	0	0	0
001	55	1994	1	0	0	0	0	0
001	60	1999	0	1	0	0	0	0
002	49	1992	0	0	1	0	1	0
002	56	1999	1	0	1	1	0	1
003	40	1992	0	0	0	1	0	1
003	42	1994	1	0	0	0	1	0
003	47	1999	0	1	0	0	1	1
004	60	1992	0	0	0	0	0	1
004	62	1994	1	0	0	0	0	1
005	36	1992	0	0	1	0	1	1
005	38	1994	1	0	1	1	0	1
005	43	1999	0	1	1	0	1	1

**Table 42 Description of variables in mock data (Table 41)**

Variable Name	Description	Values
ID	Subject ID	as assigned
Age	Subject's Age	age in years
Vyr	Year of Visit	date in years
Vis2	Complete Visit 2	yes/no
Vis3	Complete Visit 3	yes/no
Sex	Sex of subject	0= male 1= female
Fsmk	Former Smoker	yes/no
Csmk	Current Smoker	yes/no
Wez	Response to wheeze question	0= no wheeze 1= wheeze

The data setup is quite straightforward, but note that the data includes dummy variables for otherwise categorical variables (smoking, visit number). It is easier to deal with dummy variables, rather than categorical variables, in Proc Glimmix.

### **A.5.5 Dummy Variables**

Proc Glimmix does have a 'class' statement in the syntax, and therefore theoretically you can input categorical variables without any recoding. However, it is not easy to adjust the reference groups using the class statement. Instead, researchers are advised to create dummy variables for each categorical variable.

### **A.5.6 Missing Data**

Proc Glimmix does handle missing data. Observations are not excluded if variable values are missing within the observation. However, if the amount of missing data is substantial the specified models may not converge. In this case, you can limit your dataset to subjects with less missing data in an attempt to run the models successfully, but this will result in a smaller sample size and a loss of power.

### **A.5.7 Types of Research Questions**

In terms of longitudinal respiratory symptom data, Proc Glimmix should be used when your research question is similar to one of the following:

- Considering the repeated measures on individuals, what are the risk factors that predict the outcome?
- How much variation exists between individuals for a given main effect?
- How are the repeated measures on individuals correlated?
- Does the probability of the outcome change over time?

### A.5.8 Syntax

The syntax for a basic Proc Glimmix model is outlined here. For further discussion of the Proc Glimmix syntax, including the specification of a marginal model for estimating correlation structures, readers should refer to the official SAS® documentation (4).

The first model presented is a mixed model estimating the risk factors for wheeze:

```
proc glimmix data=a.mockdata ;
model wez (event='1')= age sex vis2 vis3 fsmk csmk /
s dist=binary link=logit or ;
random intercept / subject=case ;
ods output oddsratios=a.oddratio ;
run;
```

Again, the procedure statement specifies the dataset to be used. The model statement indicates that the outcome is 'wez' and that Proc Glimmix is modeling the probability of 'wez=1'. Beyond that, the model statement lists the covariates to include in the model (in this case they are all dummy variables except age) and the model options. The included model options in this example are 's' (can also be written as 'solution') to provide the fixed effects parameter estimates, 'dist' to specify the distribution of the outcome, 'link' to specify the link function and 'or' to provide the odds ratios for the fixed effects. An explanation of the 'dist' and 'link' options as well as a table of possible values is provided in the Proc Glimmix documentation (4). When the outcome is a binary respiratory symptom, the 'dist' option will be binary and the 'link' will always be logit.

The random statement specifies the random variables. In this case only a random intercept was specified, but any other random variables would be listed before the forward slash. The random statement options used here are 'subject', which identifies the

variable for which there are repeated measurements. The entry here will always be the subject or case identification variable when the repeated measures are on individuals.

The odds ratio option in the model statement gives a very long output table that is difficult to interpret from the SAS® window. For this reason it is advisable to use the 'ods output' statement to specify that the odds ratio table be output as a dataset. Once the odds ratio table is seen as a dataset file it is much easier to interpret. For more information on ODS output and how to limit the output to specific portions (using the 'ods select' statement) or output specific tables to a new dataset, refer to the SAS/STAT® documentation (55).

### **A.5.9      Selecting the Best Model**

Unfortunately there is no easy way to select the best fitting model using Proc Glimmix. Proc Glimmix does not provide a likelihood value for the estimated models, instead pseudo-likelihood is calculated and this value cannot be used in a likelihood ratio test.

Instead, users are advised to construct their model in a stepwise manner using substantive knowledge. A priori hypotheses should drive decision making while constructing the model. Once the model is assembled, the significance of individual estimates and prior knowledge should guide what remains in the model.

Additional fit statistics can be requested in the Proc Glimmix statement by including the following the command:

IC = PQ

When this command is included, pseudo-AIC and pseudo-BIC values will be included in the output fit statistic table. In the case of both pseudo-AIC and pseudo-BIC values, a smaller value indicates a better model fit.

More information on the complexities of fitting models in Proc Glimmix can be found in the documentation (4). There is also an on-going discussion of this and other pertinent SAS® issues on the SAS® user's list serve (56).



### **A.5.10 Output**

Proc Glimmix provides extensive text output in SAS®. A sample of Proc Glimmix output, limited to the key pieces of output, is shown in Figure 7. The fit statistics are shown, including the pseudo-likelihood mentioned previously. Covariance parameter estimates are the estimates of variance in each of the specified random effects, in this case only a random intercept was included in the random statement. The covariance parameter estimates provide a measure of the between subject variability in the random variable.

The next table shown is the estimates of the fixed effects included in the model statement. These are the regression coefficients describing the effect of each independent variable on the probability of reporting the symptom outcome.

If odds ratios had been requested in the output they would follow after the fixed effects parameter estimates.

The complete Proc Glimmix output is extensive, including information on the model optimization, iterative process of model fitting and the convergence criteria. Specific portions of the default output can be selected for viewing in the output window using the 'ods select' statement (55) as was done in Figure 7.

### **A.5.11 User Information**

Proc Glimmix does not ship with SAS®, instead the procedure and documentation can be downloaded from the SAS® Support website. The files are self-extracting and will copy all necessary files to the correct location (unlike Proc Traj, where you have to manually move the downloaded files into the correct folders).

The Glimmix procedure is supported by SAS® and has traditional SAS® documentation (4). In addition, the book SAS® for Mixed Models contains an intensive chapter (with examples) on generalized linear mixed models that should be reviewed.

**Figure 7 Sample output from mixed model using Proc Glimmix.**

Output - (Untitled)

The SAS System 10:25 Wednesday, October 18, 2006 18

The GLIMMIX Procedure

Number of Observations Read	2472
Number of Observations Used	2472

Fit Statistics

-2 Res Log Pseudo-Likelihood	11820.21
Generalized Chi-Square	1483.84
Gener. Chi-Square / DF	0.60

Covariance Parameter Estimates

Cov Para	Subject	Estimate	Standard Error
Intercept	CASE	1.3567	0.1763

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept	-4.3395	0.4093	921	-10.60	<.0001
AGE2	0.05232	0.006807	1539	7.69	<.0001
FEMALE	0.4411	0.2342	1539	1.88	0.0593
threevis	-0.3112	0.1622	1539	-1.92	0.0552
fourvis	0.1189	0.1903	1539	0.62	0.5321
FSMOKE	0.1215	0.1709	1539	0.71	0.4773
CSMOKE	0.5036	0.1877	1539	2.68	0.0074
ukarterm	0.7909	0.2056	1539	3.85	0.0001
ukardeas	-0.2562	0.2365	1539	-1.08	0.2789
asmachck	1.1135	0.2320	1539	4.80	<.0001
wkexsome	-0.1777	0.1890	1539	-0.94	0.3473
wkexoft	-0.3564	0.2051	1539	-1.74	0.0825

### A.5.12 Cautions

Researchers using Proc Glimmix should be aware that there are acknowledged issues with the estimation technique used in Proc Glimmix and that the procedure may result in biased coefficient estimates. A statistician can assess the magnitude of this problem using simulation techniques. Researchers should consult a statistician to ensure that their results are not biased.

### A.5.13 Reference Texts

Diggle PJ HP, Liang K, Zeger SL. Analysis of Longitudinal Data. 2nd ed. New York: Oxford University Press; 2003.

Fitzmaurice GM LN, Ware JH. Applied Longitudinal Analysis. New Jersey: John Wiley & Sons; 2004.

Littell R, Milliken G, Stroup W, Wolfinger R, Schabenberger O. SAS for Mixed Models. 2nd ed. Cary NC: SAS Institute Inc.; 2006.

SAS Institute. SAS/STAT 9.1 User's Guide: SAS Publishing; 2004.

SAS Institute. The Glimmix Procedure: SAS Publishing; 2006.

**Appendix B. Cross-tabulations of Personal Risk  
Factors for Dyspnea**

**Table 43 Cross-tabulations and Chi-square p-values for personal risk factors in Men (frequencies in percentages)**

		Atopy		Child Asthma		Age				Race		Number of Visits			Asbestos		Resp. Irritants	
		No	Yes	No	Yes	<40	40-50	50-60	>60	White	Non	2	3	4	No	Yes	No	Yes
Childhood Asthma	No	65	30															
	Yes	1	3															
	p<0.0001																	
Age	<40	19	13	31	1													
	40-50	22	9	30	2													
	50-60	20	9	28	1													
	>60	5	3	7	1													
	p=0.03		p=0.09															
Race	White	58	8	82	4	29	27	23	7									
	Non-white	28	6	13	1	3	4	6	1									
	p=0.07		p=0.8		p=0.0005													
Number of Visits	Two	31	17	46	2	14	14	15	6	41	8							
	Three	23	11	33	1	11	11	10	1	29	5							
	Four																	
		p=0.7		p=0.05		p<0.0001				p=0.2								
History of Asbestos Exposure	No	23	10	32	1	10	11	10	2	26	7	19	11	3				
	Yes	43	24	64	3	22	21	19	6	60	7	29	23	15				
	p=0.2		p=0.8		p=0.9				p<0.0001		p<0.0001							
Curr. Exposure Respiratory Irritants	No	21	9	28	2	7	9	10	3	26	3	16	9	4	16	13		
	Yes	46	25	68	3	25	22	19	5	60	11	32	25	14	17	54		
	p=0.1		p=0.4		p=0.006				p=0.2		p=0.02			p<0.0001				
Smoking Status	Never	19	11	29	2	13	9	7	1	24	7	12	11	7	10	20	8	22
	Former	28	14	39	2	10	13	13	5	38	3	21	12	8	12	29	13	28
	Current	20	9	28	1	9	9	8	2	25	4	15	10	3	11	18	8	20
	p=0.5		p=0.5		p<0.0001				p<0.001		p=0.002			p=0.1		p=0.4		

**Table 44 Cross-tabulations and Chi-square p-values for personal risk factors in Women (frequencies in percentages)**

		Atopy		Childhood Asthma		Age				Race		Number of Visits			Asbestos		Resp. Irritants	
		No	Yes	No	Yes	<40	40-50	50-60	>60	White	Non	2	3	4	No	Yes	No	Yes
Childhood Asthma	No	71	24															
	Yes	4	1															
	p=0.8																	
Age	<40	25	15	38	2													
	40-50	13	4	15	2													
	50-60	26	3	28	1													
	>60	12	3	15	0													
	p=0.07		p=0.4															
Race	White	73	25	93	5	40	17	27	15									
	Non-white	2	0	2	0	0	0	2	0									
	p=0.4		p=0.7		p=0.2													
Number of Visits	Two	48	12	59	0	20	11	19	10	58	1							
	Three	22	11	29	4	15	5	9	4	32	1							
	Four	6	2	7	1	5	1	1	1	8	0							
	p=0.4		p=0.02		p=0.8				p=0.09									
History of Asbestos Exposure	No	68	21	86	3	34	15	28	13	87	2	55	30	4				
	Yes	7	4	9	2	6	2	1	2	11	0	4	3	4				
	p=0.3		p=0.03		p=0.5				p=0.6		p=0.001							
Curr. Exposure Respiratory Irritants	No	70	21	87	4	33	16	29	14	89	2	55	31	5	84	8		
	Yes	5	4	8	1	7	1	0	1	9	0	4	2	3	6	3		
	p=0.2		p=0.4		p=0.08				p=0.7		p=0.01			p=0.02				
Smoking Status	Never	22	5	25	2	11	3	10	3	25	2	13	12	2	25	2	24	3
	Former	24	8	32	0	7	6	8	11	35	0	20	10	2	29	3	29	3
	Current	30	12	39	3	22	8	11	1	42	0	27	11	4	36	6	39	3
	p=0.6		p=0.3		p=0.005				p=0.06		p=0.6			p=0.6		p=0.9		