ANALYSES OF IDIOPATHIC PULMONARY FIBROSIS (IPF) INPATIENTS IN THE UNITED STATES

By

Christopher Policelli, MS

A Dissertation Submitted to

the Rutgers University – School of Health Professions

in partial fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics

Spring 2016

© Christopher Policelli, MS

All Rights Reserved

Final Dissertation Approval Form

ANALYSES OF IDIOPATHIC PULMONARY FIBROSIS (IPF) INPATIENTS IN THE UNITED STATES

By:

Christopher Policelli, MS

Dissertation Committee:

Shankar Srinivasan, PhD, Committee Chair
Frederick Coffman, PhD, Committee Member
Richard Haddad MD, Committee Member

Approved by the Dissertation Committee:

_____ Date: _____

_____ Date: _____

_____ Date: _____

ABSTRACT

Idiopathic pulmonary fibrosis (IPF) is a rapidly progressive immune mediated lung disorder that leads to death in the majority of patients, regardless of treatment, within 3-5 years of diagnosis. 1-3 Pulmonary fibrosis is often the final common pathway of many known causes of interstitial lung disease, such as sarcoidosis, silicosis, drug reactions, infections and collagen vascular diseases. The overall goal of the project is to identify the factors and costs associated with IPF patients in terms of mortality, length of stay and costs in different types of clinical settings across the United States.

We used data rom the Nationwide Inpatient Sample (NIS) for calendar years 2007-2012 to perform a series of analyses to identify any factors and costs associated with IPF patients, particularly those pertaining to mortality, length of stay (LOS), and costs across the various clinical settings in the US. A series of parametric and non-parametric methods were used. Parametric methods included linear regression models, correlation analysis using Pearson correlation coefficients, paired and unpaired t-tests, and one-way ANOVA. A series of non-parametric methods were also used including: Wilcoxon rank sum tests, Mann-Whitney tests, Spearman correlations, and Kruskal-Wallis tests. Descriptive analyses were employed to compute sample mean, median, standard deviation, and interquartile ranges for study variables. Finally, binary outcomes and categorical variables were examined using the following: logistic regression models, chi-square tests, contingency coefficients, and McNemar's test.

We found a mean expenditure per IPF patient of $68,925 (standard deviation = $16,4049), with a median of $28,257. We found the distribution of total charges to be significantly different from the normal distribution ($p<0.001$). Further investigation revealed a maximum charge of $4,162,849, which is approximately six times as high as the 99th percentile total charge of and 147 times as high as the median. These results indicate that the distribution of total costs are skewed heavily by a relative few patients with extremely high bills.

LOS also varied wildly with the mean stay being 7.84 days with a standard deviation of 13.60 days. LOS ranged from zero days to 1158 days, which corroborates our conclusion about the distribution of total costs. The patient who stayed 1158 days was hospitalized for 858 days more than any other patient. The next highest LOS values were 300 days, 250 days, 196 days, and 160 days, which further plays to our notion of the extremely expensive few. LOS varied in a statistically significant way between white and black patients ($p<0.05$).

Only 11.2% of all patients in the data set died during the study period (11.2% of whites, 9.5% of blacks, 11.2% Hispanics, 12.2% of Asian/Pacific Islander, 13.0% of Native Americans). Finally, only 12% of all IPF patients passed away, while slightly more (13.8%) of all IIP patients died during the study period.

Keywords: idiopathic pulmonary fibrosis, Nationwide Inpatient Sample, inpatient care, pulmonary rehabilitation

ACKNOWLEDGEMENTS

As this dissertation represents the end of a long scholastic journey that has included many "twists and turns "over the past 18 years in collegiate institutions, I feel compelled to add some acknowledgments. Dr. Shankar Srinivasan has been the perfect advisor for someone in my circumstances and has been extremely helpful, patient, and understanding. His counsel, guidance, and expertise have been instrumental in making this dissertation useful and something I am proud of. Dr. Srinivasan was always available for me when I needed him for resources, tools, or quick answers. I would also like acknowledge the graduate faculty Biomedical Informatics of the School of Health Related Professions at Rutgers University: Dr. Syed S. Haque, Department Chair and Program Director, who always inspired me to continue to challenge myself and learn. Drs. Dinesh Mital and Masayuki Shibata have through their hard work on colloquiums and their contributions and suggestions to the content, presentation, and overall organization of this dissertation study; the entire faculties dedication to education and the field of Biomedical Informatics is very inspiring. I would also like to acknowledge Ms. Yvonne Rolley, the Administrative Coordinator for our program for her hard work and dedication; she was very supportive throughout. Also, I would like to acknowledge my fellow students whose questions, presentations, dedication, and interactions always helped to come up with new ideas. Finally, I would like to acknowledge my company, Celgene for their financial support and flexibility throughout this process.

DEDICATION

I would like to dedicate this dissertation to my family: My parents Mark and Judy Policelli worked very hard to always provide opportunity for me to succeed and pushed me in just the right way through understanding and love. My wife Nancy Policelli, who is an amazing wife, mother and person for living through my professional and academic career as we built our family. And my children Luke and Mark Policelli who from the outside may have made this more difficult to get through but in reality, my love and dedication to them drove me to always strive for more. I love you all more than anything in the world.

Table of Contents

Chapter One - Introduction 1

1.1 Background of the Disease: 1

1.2 Goals and Objectives: 2

1.3 Data & Methods: 3

Chapter Two - Literature Review 5

2.1. Introduction: 5

2.2 Signs and Symptoms 8

2.2.1 Diagnostic findings 9

2.2.2 Physiological changes 11

2.3 Natural History and Prognosis 12

2.4. Treatment: 13

2.5. Risk Factors for IPF: 15

2.5.1. Personal demographics: 15

2.5.2. Cigarette smoking: 16

2.5.3. Diabetes: 17

2.5.4. Diet/Gastroesophagael Reflux: 18

2.5.5. Genetic Influences: 19

2.5.6 Environmental Factors 19

2.5.7. Length of Hospital Stay & Cost: 21

Chapter Three - Methodology 22

3.1. Nationwide Inpatient Sample Data: 22

3.2. Goals and Objectives: 25

3.3. Research Design & Methods: 25

3.4. Statistical Methodology: 26

3.5. Statistical Analyses: 27

3.6. Modeling Techniques Overview: 29

3.7. Cox Proportional Hazards Regression: 29

3.8. C-Statistic: 30

3.9. Linear Regression:.....31
3.10. Kolmogoro Smirnov Test: 31
3.11. Logistic Regression: 32
Chapter Four - RESULTS.....34
4.1 IPF Descriptive Analysis for 2008-2012 34
Post-proposal analysis 74
Post Defense Analysis.....84
References 99

List of Figures

Figure 1: Normal airways and breathing…….……6

Figure 2: IPF incidence studies……..……...7

Figure 3: IPF survival by country……..……..13

Figure 4: Prominent clinical studies evaluating gastroesophageal reflux in IPF……18

Figure 5. Environmental exposure pathways and IPF…….21

Table 1. Methods for diagnosing IPF……..……..…..10
Table 2. NIS data observations by year…….……..23
Table 3. IPF diagnosis by year……..24
Table 4. Data Variables Used for Analysis ……...……... …….24
Table 5. The UNIVARIATE Procedure Total Charge - IPF…..…….35
Table 6. The UNIVARIATE Procedure Total Charge (Basic Statistical Measures) – IPF…….....35
Table 7. The UNIVARIATE Procedure Total Charge (Basic Confidence Limits Assuming Normality) -IPF……..……35
Table 8. The UNIVARIATE Total Charge (Tests for Location) – IPF..……36
Table 9. The UNIVARIATE Procedure Total Charge (Test for Normality) – IPF.....…..37
Table 10. The UNIVARIATE Procedure Total Charge Quantiles - IPF……..38
Table 11. The UNIVARIATE Procedure Total Charge (Extreme Observations) – IPF…….....38
Table 12. The UNIVARIATE Procedure Total Charge (Missing Values) - IPF……...39
Table 13. The UNIVARIATE Procedure Total Charge – NIS……...39
Table 14. The UNIVARIATE Procedure Total Charge (Basic Statistical Measures) – NIS…….40
Table 15. The UNIVARIATE Procedure Total Charge Basic Confidence Limits Assuming Normality – NIS……40
Table 16. The UNIVARIATE Procedure Total Charge (Tests of Location) – NIS…….41
Table 17. The UNIVARIATE Procedure Total Charge (Tests for Normality) – NIS……...41
Table 18. The UNIVARIATE Procedure Total Charge Quantiles – NIS…….42
Table 19. The UNIVARIATE Procedure Total Charge (Extreme Observations) – NIS…42
Table 20. The UNIVARIATE Procedure Total Charge (Missing Values) – NIS…..……43
Table 21. The UNIVARIATE Procedure of (Length of Stay) - IPF…..……...44
Table 22. The UNIVARIATE Procedure of LOS (Basic Statistical Measures) – IPF…….....45
Table 23. The UNIVARIATE Procedure of LOS (Basic Confidence Limits Assuming Normality) - IPF…..……..…45
Table 24. The UNIVARIATE Procedure of LOS (Test for Location) – IPF…..……46
Table 25. The UNIVARIATE Procedure of LOS (Tests for Normality) - IPF…....46
Table 26. The UNIVARIATE Procedure of LOS Quantiles ……..……..47

Table 27. The UNIVARIATE Procedure of LOS (Extreme Observations) – IPF……...48
Table 28. The UNIVARIATE Procedure of LOS (Missing Values) – IPF……..48
Table 29. The UNIVARIATE Procedure of (Length of Stay) – NIS…..……49
Table 30. The UNIVARIATE Procedure of LOS (Basic Statistical Measures) – NIS……..…….....49
Table 31. The UNIVARIATE Procedure of LOS (Basic Confidence Limits Assuming Normality) – NIS..……..49
Table 32. The UNIVARIATE Procedure of LOS (Test for Location) – NIS……...50
Table 33. The UNIVARIATE Procedure of LOS (Tests for Normality) - NIS….....50
Table 34. The UNIVARIATE Procedure of LOS Quantiles - NIS ……..…….51
Table 35. The UNIVARIATE Procedure of LOS (Extreme Observations) - NIS……...51
Table 36. The UNIVARIATE Procedure of LOS (Missing Values) – NIS.……..51
Table 37. The FREQ Procedure (AGE by DIED)……..…….....53
Table 38. The FREQ Procedure – Primary payor/insurance (PAY1) – IPF……...54
Table 39. The FREQ Procedure – Pay1 NIS……..……54
Table 40. The FREQ Procedure – Pay1 by RACE – IPF……..55
Table 41. The FREQ Procedure – Pay1 by RACE – NIS…….56
Table 42. The FREQ Procedure (Pay1 by DIED) – IPF..…….....57
Table 43. The FREQ Procedure (Pay1 by DIED) – NIS……...57
Table 44. The FREQ Procedure Statistics for RACE by DIED……..……..58
Table 45. Total charges, length of stay, and number of diagnoses by Payor – IPF-only……60
Table 46. Total charges, length of stay, and number of diagnoses by payor – NIS…….60
Table 47. Total charge, LOS, number of diagnoses, severity, risk of mortality by death status – IPF……..62
Table 48. Total charge, LOS, number of diagnoses, severity, risk of mortality by death – NIS…….62
Table 49. The FREQ Procedure of DX1 by DIED……..63
Table 50. The FREQ Procedure AGE by DX1……..…….....64
Table 51. Mean total charge, length of stay, and number of diagnoses among top 10 diagnoses within IPF patients……68
Table 52. Total charges, length of stay, and number of diagnoses (per person) by top diagnoses – NIS/HCUP data set……...69
Table 53. The GLM Procedure Class Level Information……..……....70

Table 54. The GLM Procedure Number of Observations Read and Used…..……..70
Table 55. The GLM Procedure (LOS) – Model Summary – Length of Stay = Race…….....71
Table 56. Predictive Power of Race for Length of Stay.……..…….…….…....71
Table 57. The GLM Procedure Type I & III of SS…….……72
Table 58. The GLM Procedure by HSD test for LOS……..…..……72
Table 59. The GLM Procedure (LOS) Comparisons significant at the 0.05 level. ……..……..73
Table 60. Primary diagnosis by race – IPF…….76
Table 61. Top primary diagnoses by race – NIS/HCUP data set……77
Table 62. Cost comparison by Diagnosis-Related Group…….…….....78
Table 63. Most common procedures among IPF patients…….…….. …81
Table 64. AHRQ comorbidities with IPF…….…….82
Table 65. Reported severity and risk of mortality among IPF patient……..85
Table 66. Total charges and lengths of stay by severity…….....87
Table 67. Mean age – all NIS/HCUP and IPF only…….....88
Table 68. Average age by race…….90
Table 69. Pearson corerlations age and race (NIS)……..…...90
Table 70. Age and race correlations among IPF patients…….91
Table 71. Average age by Payor/Insurance…….……...91
Table 72. Insurance/Payor by data grouping……..92
Table 73. Pearson Correlations for Age and Payor (NIS)……...……..92
Table 74. Pearson Correlations for Age and Payor (IPF).……92
Table 75. Mean ages among primary diagnoses.……...95
Table 76. Age distribution among the top 10 primary diagnoses (IPF)……..95
Table 77. Pearson coefficients for age and top diagnoses (NIS).…….....96
Table 78. Pearson coefficients for age and top diagnoses (IPF).……97
Table 79. Remaining correlations for variables analyzed within the IPF data set……98
Table 80. Pearson Coefficients for severity, risk of mortality, length of stay, total charge, sex, and race variables in the NIS data set…….……98

Chapter One - Introduction

1.1 Background of the Disease:

Idiopathic pulmonary fibrosis (IPF) is a rapidly progressive immune mediated lung disorder that leads to death in the majority of patients, regardless of treatment, within 3-5 years of diagnosis. 1-3 Pulmonary fibrosis is often the final common pathway of many known causes of interstitial lung disease, such as sarcoidosis, silicosis, drug reactions, infections and collagen vascular diseases, the etiology for the fibrosis often cannot be determined, and the disease is classified as an idiopathic interstitial pneumonia (IIP). IPF, which is defined histologically as usual interstitial pneumonitis (UIP), is the most common subcategory of IIP, accounting for greater than 60% of cases. 4,5 Unfortunately, although IPF is the most common IIP, it is also the least treatable and has the worst prognosis of all of the IIPs. 3,6

IPF predominantly afflicts older individuals with approximately 2/3 of the patients over the age of 60 years at presentation. 4,7 The prevalence of this disease has been estimated to be 13.2-20.2 per 100,000 with an annual incidence of 7.4-10.7 per 100,000 new cases a year 8. The initial signs and symptoms of IPF are often subtle and insidious in onset with progressive dyspnea on exertion and dry cough2,4,9,10. Unfortunately, the cough, which occurs in 73-86% of patients, is often disabling and resistant to traditional therapy. IPF almost universally progresses regardless of current treatments 11,12.

Although many factors have been shown to predict disease progression such as mortality, hospitalization, need for supplemental oxygen, acute exacerbations, dyspnea, decline in Forced Vital Capacity (FVC) and Carbon Monoxide Diffusing Capacity (DLCo), 6- minute walk, honeycombing on High Resolution Computed Tomography (HRCT), and declining quality of life scores, none of these measures have proven robust as short term outcomes in clinical trials.

1.2 Goals and Objectives:

The overall goal of the project is to identify the factors and costs associated with IPF patients in terms of mortality, length of stay and costs in different types of clinical settings across the United States. Specifically the objectives are to determine:

1. what clinical factors (such as number and types of comorbidities and procedures) influence the mortality, costs and length of stay
2. whether mortality, costs and length of stay differ with race, age, or socio-economic status
3. whether there are differences in the mortality, costs and length of stay across the various regions of the US
4. Whether there are differences in the mortality, costs and length of stay amongst the different types of hospital settings – rural/urban/hospital with and without teaching.

1.3 Data & Methods:

In this project we plan to utilize the datasets obtained from the Nationwide Inpatient Sample (NIS) database patients. The NIS is the largest all-payer inpatient care database in the United States containing data from 1998 to 2012. It contains data from approximately 8 million hospital stays each year accruing from all discharge data from 1,050 hospitals located in 44 States, approximating a 20-percent stratified sample of U.S. community hospitals. The 2012 NIS is a sample of hospitals that comprises approximately 95 percent of all hospital discharges in the United States. The NIS includes more than 100 clinical and nonclinical data elements for each hospital stay. These include:

- Primary and secondary diagnoses
- Primary and secondary procedures
- Admission and discharge status
- Patient demographics (e.g., gender, age, race, median income for ZIP Code)
- Expected payment source
- Total charges
- Length of stay
- Hospital characteristics (e.g., ownership, size, teaching status).

Furthermore, the NIS is the only national hospital database containing charge information on all patients, regardless of payer, including persons covered by Medicare, Medicaid, private insurance, and the uninsured.

The outcomes of interest as indicated in the goals and hypotheses above are the mortality, the length of stay and the costs involved. Using the datasets obtained from the NIS database appropriate descriptive and inferential statistics will be effected. To relate the factors associated with the research outcome, the length of stay and the costs a multiple regression model will be setup and validated. Predictive models such as logistic regression will be employed to determine the risks and ratios for the various factors influencing mortality such as race, age groups, number and types of procedures and comorbidities. Details as to the state of art knowledge and research into Idiopathic Pulmonary Fibrosis and its management are provided in the next chapter.

Chapter Two - Literature Review

2.1. Introduction:

Idiopathic pulmonary fibrosis (formerly known as cryptogenic fibrosing alveolitis).”is a serious and lethal disease where the alveoli and lung tissue next to the alveoli become damaged and scarred. As scar tissue accumulates, lung tissue thickens, which decreases the lungs’ capacity to properly move oxygen into the bloodstream. As a result the brain and other organs do not receive the amount of oxygen required to perform their work. The etiology of IPF is consistent, hence the name “idiopathic”, meaning not following a convention or not conforming to a set of rules. In most case, the physician is unable to determine the exact cause of IPF. Figure 1 (found below) depicts a normal airway, and zooms to the alveoli to illustrate the formation of IPF.

IPF is a rare disease that affects only a small portion of the population. The incidence and prevalence of IPF are difficult to determine because of uniform diagnostic criteria have only been defined since the mid 2000’s. Historical information relating to vital statistics relied on population studies which utilized diagnostic coding data and death certificates to identify cases. The accuracy of this information can be questioned, especially when studies performed in the era of undefined diagnostic criteria.

The best available data suggested that incidence of IPF is approximately 10.7 per 100,000 persons for men; and 7.4 per 100,000 persons for women. The prevalence of IPF is slightly greater than 20.2 men per 100,000 and 13.2 women per 100,000. Data from

around the world suggest that IPF favors no particular race, ethnic group, or social environment. It is estimated that IPF affects at least 5 million persons worldwide. It also appears that incidence of IPF is on the rise, due in large degree to an increased diagnostic ability and increased attention given to the disease (Coultas, 1994; Hansell 1999).

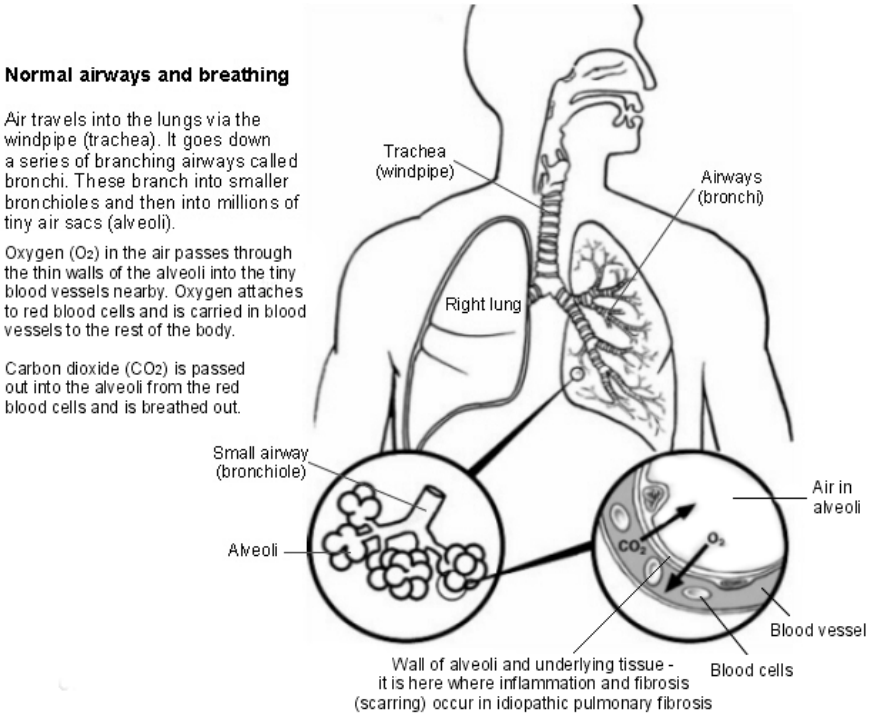

Figure 1: Normal airways and breathing

Source: http://medical.cdn.patient.co.uk/images/212.gif

There are no large-scale data sets on the incidence or prevalence of IPF. However, several smaller studies have suggested that new (incident) cases of IPF develop in 5-10 people per 100,000 per year. Prevalence of IPF is estimated at between 14-43 persons per 100,000 population. (American Thoracic Society, 2000). The Coalition For Pulmonary

Fibrosis estimates that 128,000 people in the United States suffer from IPF, with 15,000 new cases being diagnosed annually. However, the true number of patients who suffer from IPF is through to be higher as the disease is often underdiagnosed or misdiagnosed (Gribbin, 2006; Navaratnam, 2001; Raghu, 2006; Fernandez, 2010; Coultas, 2010; Meltzer, 2008). The chart below outlines the incidence and prevalence of IPF as described by five studies in the United States and the United Kingdom.

Study	Study period	Patient age (years)	Location	Incidence rate per 100,000 person-years	Prevalence per 100,000
Gribbin[3]	1991 – 2003	> 40 at diagnosis	UK	4.6	Not measured
Navaratnam[4]	2000 – 2008	> 40 at diagnosis	UK	7.4	Not measured
Raghu[5]	1996 – 2000	≥ 18	USA	6.8[a] – 16.3[b]	14.0[a] – 42.7[b]
Fernandez Perez[6]	1997 – 2005	≥ 50	USA	8.8[a] – 17.4[b]	27.9[a] – 63.0[b]
Coultas[7]	1988 – 1990	≥ 18	USA	7.4[c] – 10.7[d]	13.2[c] – 20.2[d]

a Narrow case definition; b Broad case definition; c Females; d Males

Figure 2: IPF Incidence studies

Source: http://www.inipf.com/about-ipf/disease-information/what_is_the_incidence_and_prevalence_of_ipf/_jcr_content/par/media/image.-1064343630.image.png

IPF has no known cure. Many IPF patients live only 3 to 5 years after diagnosis, with many dying of respiratory failure. Other causes of death related to IPF include pulmonary hypertension, heart failure, pulmonary embolism, pneumonia, and lung cancer. In a well-known study of 596 potential IPF patients by Fernandez, et, al., median survival for narrow-criteria and board criteria incidence cases was 3.5 and 4.4 years respectively (Fernandez, 2010).

2.2 Signs and Symptoms

IPF is gradual in onset, initially presenting as dyspnea (a.k.a. shortness of breath) upon exertion and dry cough. Early and non-specific symptoms are often mistaken for the natural process of aging, cardiac disease, emphysema, bronchitis, asthma, or COPD. Misdiagnosis, or the ignoring of early symptoms entirely, may explain why IPF is rarely diagnosis immediately A correct diagnosis may take several months or even years. Respective analysis of IPF patients suggests that symptoms precede diagnosis by a period of 6 months to 2 years. (ATS/ERS, 2000; Collard, 2007; Meltzer, 2008; Borchers, 2011; Cottin, 2012; Swigris, 2005; Raghu, 2011).

A common early sign of IPF is bibasilary inspiratory ‘Velcro crackles’ on lunch auscultation. Abnormal crackles can be observed (heard) in greater than 80% of patients. These can be detected when a physician listens to patient’s lung sounds with a stethoscope during inspiration. The crackles initially appear in the basal areas of the lung where the disease initiates. As the disease progresses, crackles spread to the upper zones of the lungs (Cottin, 2012).

Finger clubbing is another symptom of IPF. It is characterized by the spreading out and rounding of finger tips. Finger clubbing occurs in 25-50% of all patients. (ATS/ERS, 2000; Borchers, 2011). Other extrapulmonary symptoms are rate; however, they include weight loss malaise, and fatigue, all of which could be mistaken as symptoms of a litany of other disease, or as part of a number of normal processes, thus further contributing to misdiagnosis.

Shortness of breath and coughing (especially dry coughing) can be especially impairing to the quality of life for IPF patients. Symptoms progress and worsen over time, as patients in the later stages of IPF often experience curtailing of physical activity. Cough has been described by patients as being “dry and non-productive” or “hacking” and “occurring when talking for long period” with many patients experiencing an intense nagging to cause constantly or a lack of relief once coughing occurs. Fatigue is another symptom of IPF that limits a patient’s physical activity and reduces quality of life.

The examination of patients with IPF should attempt to identify those signs suggesting an alternative diagnosis such as systemic sclerosis or polymyositis that can be associated with secondary pulmonary fibrosis. To this end, the examiner should look for sclerodactily, scleroderma, proximal muscle weakness and telangiectasias. The history should exclude Raynaud’s phenomenon.

2.2.1 Diagnostic findings

Because patients with Pulmonary Fibrosis experience similar symptoms and scarring patterns to those with other lung disorders, pulmonary fibrosis can be difficult to diagnose. An estimated 50% of cases are initially misdiagnosed as another form of

respiratory disease. In fact, until 2000, pulmonary fibrosis was classified as a distinct clinical disorder by the American Thoracic Society (ATS) and European Respiratory Society (ERS). Until recently, the medical community has not agreed upon a standard of diagnosis for IPF. As a consequence, other related lung diseases were often incorrectly classified as Pulmonary Fibrosis. With new diagnostic standards now in place, the recognition and management of IPF should be substantially improved. The table below (from the Coalition of Pulmonary Fibrosis) outlines a series of tools that are commonly used to diagnose IPF.

Table 1: Methods for diagnosing IPF

DIAGNOSTIC	DESCRIPTION	PURPOSE
Chest imaging	Use of radiologic machines to take pictures of your lungs, such as x-ray or High Resolution Computer Tomography (HRCT)	To view lung structures look for scar tissue and assess patterns of scarring
Pulmonary function test	A test using a device with a mouthpiece to measure a patient's breathing capacity	To measure the degree of impairment in lung function
Arterial blood gas test	A measurement of oxygen and carbon dioxide levels in blood taken from an artery in the wrist	To determine how well the lungs are performing vital gas exchange
Exercise Test(or desaturation study)	A test in which the patient is monitored while using a treadmill or stationary bicycle	To measure how well the lungs and heart respond to physical activity and evaluate oxygen levels with exertion
Six Minute Walk Test (SMWT)	A test where a patient walks on a flat surface as far as possible in six minutes	To measure the distance you are able to walk as well as lung function during the walk.
Bronchoalveolar lavage (BAL)	A "lung-washing" procedure conducted through a flexible tube (bronchoscope) inserted into the airways through the nose or mouth; fluid (salt water) is injected into	To examine cells and fluid to look for signs of inflammation in the lungs, or markers of

	the lungs and then removed for inspection	disease activity
Lung biopsy	A procedure in which a tissue sample is obtained through a bronchoscope (see BAL, above) or by means of a small surgical incision (VATS- video-assisted thoracic surgery) between the ribs (open-lung biopsy)	To obtain a sample of lung tissue for direct examination

2.2.2 Physiological changes

Routine spirometry reveals decreased measures of forced vital capacity (FVC) and forced expiratory volume in one second (FEV_1) in most IPF patients. The ratio of FEV_1/FVC remains normal (or increased) in IPF, consistent with restrictive physiology. Lung volume measurements typically confirm restrictive physiology, which is usually manifested by a reduction in total lung capacity (TLC). Restrictive physiology is the consequence of reduced pulmonary compliance. Changes in compliance can be attributed to the accumulation of parenchymal scar tissue, which leads to the distortion of normal lung architecture.

Gas exchange, or the ability of the lungs to deliver oxygen to the bloodstream, and eliminate carbon dioxide from the bloodstream to the lungs, is impaired in IPF, which can be demonstrated by measurement of diffusion capacity. Declining diffusion capacity can sometimes precede changes in lung volume. Isolated impairment of diffusion capacity can be found during the early stages of IPF.

Resting arterial blood gas is usually normal. Mild hypoxemia and mild respiratory alkalosis can occur in end-stage disease. Although resting arterial oxygen saturation remains normal, oxygen desaturation is commonly found during exercise. The main

cause for exercise-induced hypoxemia is ventilation-perfusion (V/Q) mismatching, as opposed to anatomic shunting or reduced diffusion capacity (Augusti, 1991).

2.3 Natural History and Prognosis

IPF's natural history is incompletely known. IPF usually assumes a course of relentless physiological deteriorations. However, some patients remain stable for extended periods of time, and individual outcomes can be highly variable (Kim, 2006). Nonetheless, longer-term survival with biopsy proven IPF is not expected. New insight into the natural history of IPF has been gleaned from the results of secondary analysis of placebo groups assembled for recently conducted multi-center clinical trials (Azuma, 2005; Raghu, 2004; Demedts, 2005).

Overall prognosis for IPF patients is poor, with a median survival of 2-5 years from the time of diagnosis (Raghu, 2004; Frankel, 2009). The survival rate is even lower than that of several common cancers. Estimated mortality rates are 64.3 deaths per million in men and 58.4 deaths per million in women. The graph below is adapted from Vencheri et.. 2010. Only lung cancer and Pancreatic cancer have lower 5-year survival percentages than IPF. The Coalition For Pulmonary Fibrosis estimates a 5-year survival of between 30% to 50% (COPF, 2015, Olson, 2007).

Estimates are that 60% of patients with IPF die from IPF, as opposed to with IPF. Of those patients who die with IPF, most commonly it is after an acute exacerbation of the disease. When an acute exacerbation of IPF is not the cause of death, an increased cardiovascular risk and an increased venous thromoboembolic disease risk contribute to the cause of death. The most common causes of death in patients with IPF are acute

exacerbation of IPF, acute coronary syndromes, congestive heart failure, lung cancer, infectious disease due to a immunosuppression, and venous thromboembolic disease.

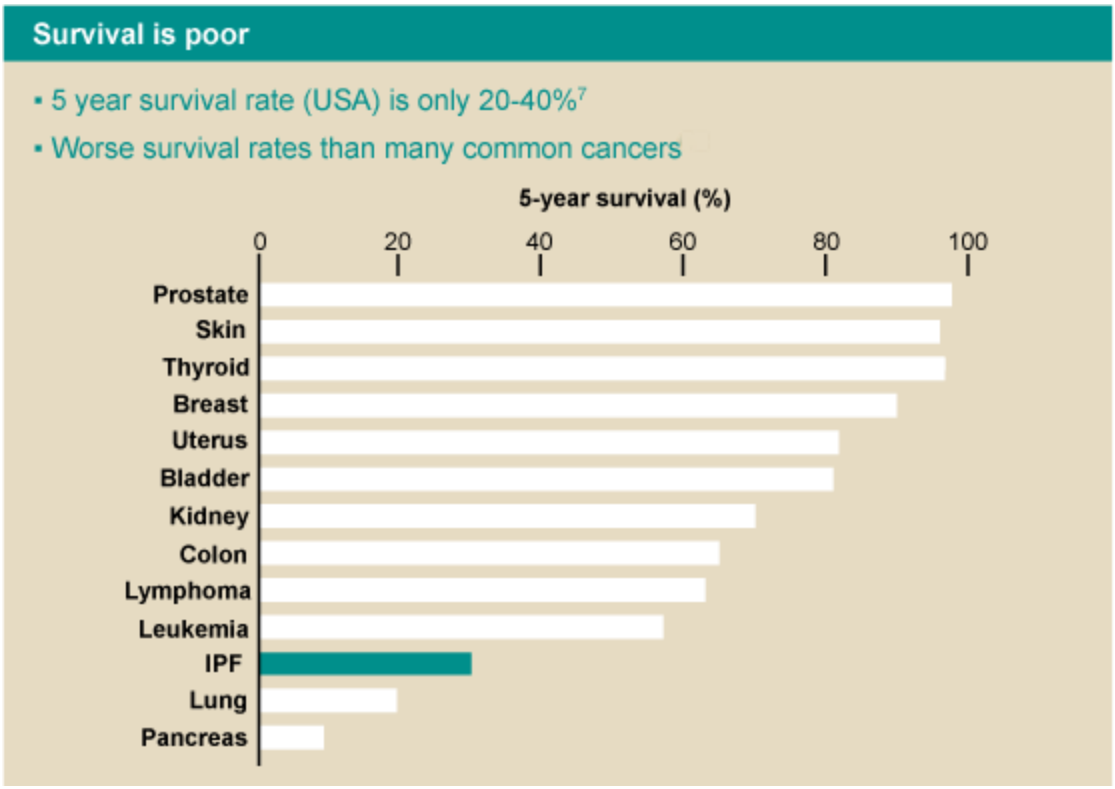

Figure 3: IPF Survival by country

2.4. Treatment:

The most common treatments for IPF include steroids, antioxidents, oxygen therapy, pulmonary rehabilitation, and lung transplant. Currently, no medicines have proven to slow the progression of IPF. Three prescriptions, however, have traditionally been used for the treatment of symptoms, including predinisone (anti-inflammatory), azathrioprine (immune-suppressant), and N-acetylcysteine (antioxidant used to prevent lung damage). As of 2014, nintedanib and pirfenidon have been shown to slow disease

progression in separate Phase III clinical trials and, for the first time, two treatment drug-based alternatives might become available for IPF patients (King, 2014; Richeldi,2014).

Because of the lack of success of traditional prescription-based treatments, other medicines and treatment options are currently being explored. Oxygen therapy is employed in many situations to help reduce shortness of breath and to allow IPF patients to be more active. Oxygen is usually give through nasal prongs or a mask. Initially, oxygen is administered only during exercise or sleep; however, as the disease worsens, it may be needed more frequently. Eventually, oxygen is needed on a constant basis.

Lung transplantation is another commonly used treatment for IPF, and is so far, the only treatment with proven benefit, conferring a better survival for some carefully selected patients. However, the number of lung transplantations performed is limited primarily by the supply of donor organs, and survival is poor for IPF patients relative to most other disease categories (OPTN, International Society For Heart & Lung Transplantation). Both single and bilateral transplantations are performed in patients with IPF, and debate remains as to whether single lung transplantations (SLT) or bilateral transplantation (BLT) is the better choice. Among IPF patients who received a lung transplant in the US from 1987 to 2009, the leading cause of death was infection (24% of deaths) (Thabut, 2009). Articles resulting from single-center studies and one dual center study also reported that infection (or sepsis) was a leading cause of death, along with bronchiolitis, obliterans syndrome, chronic rejection, or unspecified graft failure (Erasmus, 2008; Grossman, 1990; Mason, 2007; Neurohr, 2010; Meyers, 2000; Rusanov, 2012; Saggar, 2010; Schachna, 2006; Thabut, 2003; Willie, 2008).

Pulmonary rehabilitation (PR) is the current standard treatment (as of 2015) for chronic lung disease patients. PR encompasses a broad program whose goal is to improve the well-being of individuals with breathing problems. The program typically involves treatment administered by a team of specialists within a treatment facility. The goal of PR is to teach individuals how to manage breathing conditions (including IPF), and equip them with knowledge they will need to function at their best.

PR is not meant to replace medical therapy, rather, it is used in conjunction with ongoing IPF treatment. PR treatment activities include:

- Exercise training
- Nutritional counseling
- Education on lung disease and its management
- Energy-conserving techniques
- Breathing strategies
- Psychological counselling and/or group support

2.5. Risk Factors for IPF:

2.5.1. Personal demographics:

IPF affects men more than women. The incidence of IPF increases with age. IPF most commonly appears between the fifth and seventh decades, with two-thirds of lal cases arising in patients over 60 years of, with the mean age at presentation being 66 years old (American Thoracic Society, 2000). IPF occurs infrequently in those younger than 40 and rarely affects children, if at all. One large U.S. population-based study noted

a significant difference by age (Coultas, 1994). The study found that the prevalence of IPF was only 2.7 cases per 100,000 amongst those aged 34 to 44 years old; meanwhile 175 cases per 100,000 were found among persons over the age of 75 years. Worldwide, the incidence of IPF is estimated to be 10.7 cases per 100,000 person-years for males and 7.4 cases per 100,000 for females. The prevalence of IPF is estimated to be 20 cases per 100,000 persons for males and 13 cases per 100,000 persons for females (Collard, 2006). Other estimates, derived using data obtained from a large US healthcare claims database estimate the incidence of IPF from 0.4 – 1.2 cases per 100,000 person-years for person aged 18-34 years. The same estimates incidence in persons age 75 years and older at between 27.1 and 76.4 cases per 100,000 person-years (Raghu, 2006).

2.5.2. Cigarette smoking:

The prevalence of tobacco use in IPF ranges from 41% to 83% (Antinou, 2008; Oh, 2008), depending on the case definition used in studies. Current or former smokers have consistently been overrepresented (Carrington, 1978; Johnston, 1997; King, 2000; Ryu, 2001; Schwartz, 1994; Turner-Warwick, 1980; Watters, 1987). As with other pulmonary morbidities, cigarette smoking is a suspected risk factor for IPF, with recent work suggesting that smoking may have a detrimental effect on survival of patients with IPF. In theory, increased oxidative stress in current and former smokers may promote disease progression. However, formal research studies have disagreed as to the role that smoking plays in the development and progression of IPF. Previous research had counter-intuitively suggested that current cigarette smokers with IPF tend to experience longer survival that ex-smokers (Antoniou, 2008). The mechanism by which smoking may contribute to the pathogenesis of IPF is unknown despite studies suggesting a strong

association between the development of IPF and cigarette smoking. Because IPF is recognized as a disease of aging, it has been speculated that smoking may contribute to the development of interstitial lung disease in an age-dependent manner.

Antiniou, et al., studied the medical records of 249 IPF patients and examined the extent and severity of their disease, smoking history, and survival. Their initial findings were, which were unadjusted for disease severity, indicated that smokers had longer survival times than ex-smokers. Once adjustments for IPF severity were made, the results revealed the following (as dictated by Antoniou):

> “"We established that current smokers live longer, but this is mostly because they have much milder disease. Clearly, many patients stop smoking precisely because their disease is getting worse. This is the 'healthy smoker' effect: that current smoking is a marker for milder disease because advancing disease causes smoking cessation," said Dr. Wells. "Symptomatic patients with more severe disease may be more likely to stop smoking for perceived health reasons. It can, therefore, be argued that current smoking might be a marker of less severe disease, associated with better survival."

Cigarette smoking is strongly associated with IPF. One study reported a correlation between smoking history (20-40 pack years) and risk for IPF, with an odds ratio of 2.3 (95% confidence interval: 1.3 – 3.8) for smokers (Baumgartner, 1997).

2.5.3. Diabetes:

Among lifestyle-related diseases, Diabetes Mellitus is a frequent complication of patients with IPF; however, the prevalence is unknown. Several small-scale studies have cited Type 2 diabetes as a risk factor for IPF. One such study examined 657 patients, and determined that patients with Type 2 diabetes experienced a 4.3 times as likely to develop IPF (95% confidence Interval: 1.9 – 9.8) (Perez-Padilla, 2009). Another such study

estimated an unadjusted odds ratio of 3.88 (95% confidence interval: 1.85 – 8.12). After making adjustments for obesity, and smoking, the odds ratio moved only slightly to 4.08 (1.80 – 9.15), thus cementing diabetes as a risk factor for IPF.

2.5.4. Diet/Gastroesophagael Reflux:

The notion of recurrent microaspiration aas a potential cause of pulmonary fibrosis is not a new one, with reported case series dating back to the 1960's. Several prominent observational clinical studies have noticed an association between gastroesophagael reflux disease (GERD) and IPF. There was a limited correlation between typical oesophageal reflux symptoms (e.g., heartburn) and objective reflux events. A number of studies have subsequently attempted to evaluate the exact prevalence of reflux in IPF (as illustrated in Table 1 from Fahim, 2010).

Table 1: Prominent clinical studies evaluating gastroesophageal reflux in IPF.

Study	Methodology	Number of subjects	Prevalence of GERD	Other outcomes
Tobin et al. 1998 [35]	Prospective with non-IPF ILD control	17 IPF 8 controls	94% IPF 50% controls	25% of IPF patients had typical reflux symptoms
Raghu et al. 2006 [36]	Prospective, control group without ILD	65 IPF 133 asthmatics	87% IPF 68% Asthma	47% of IPF patients had heartburn and regurgitation. No significant difference in proximal reflux in IPF and asthma, 63% versus 61%, respectively
Raghu et al. 2006 [19]	Retrospective case review	4 IPF	100% as one of the inclusion criteria	2–6 year follow up with stable FVC and TLCO with proton pump inhibitors
Salvioli et al. 2006 [37]	Prospective	18 IPF 10 secondary pulmonary fibrosis	67% of IPF patients had abnormal distal reflux	57% of total patients had heartburn and regurgitation
Bandiera et al. 2009 [38]	Prospective	28 IPF	35.7%	Participants divided into $GRED^{+}$ and $GERD^{-}$ groups

Figure 4: Prominent clinical studies evaluating gastroesophageal reflux in IPF

2.5.5. Genetic Influences:

Changes in several genes have been suggested as risk factors for IPF. Mutations in genes known as TERC and TERT have been found in about 15 percent of all caes of familial pulmonary fibrosis and a smaller percentage of cases of sporadic idiopathic pulmonary fibrosis. The TERC and TERT genes provide instructions for making components of telomerase, which maintains structures at the ends of chromosomes known as telomeres. It is not well understood how defects in telomerase are associated with IPF.

Most cases of IPF are sporadic, meaning that they occur in people with no history of the disorder in the family. However, familial pulmonary fibrosis appears to have a pattern of autosomal dominant inheritance, meaning that one copy of an altered gene in each cell s sufficient to cause the disorder. Some individuals who inherit the altered gene never develop features of familial pulmonary fibrosis (known as reduced penetrance). It is unclear why some individuals with a mutated gene develop the disease while others with the mutated gene do not (Genetics Home Reference, 2015).

2.5.6 Environmental Factors

Several sources of evidence, including investigations of pathogenesis and observational studies, support the hypothesis that environmental agents may have an etiologic role in idiopathic pulmonary fibrosis (IPF). Since 1990, six case-control studies have been conducted in three countries and have consistently demonstrated increased risk

of IPF with exposures to a number of environmental and occupational agents. In a meta-analysis of these studies, six exposures were significantly associated with IPF (summary odds ratios [95% confidence intervals]), including ever smoking (1.58 [1.27–1.97]), agriculture/farming (1.65 [1.20–2.26]), livestock (2.17 [1.28–3.68]), wood dust (1.94 [1.34–2.81]), metal dust (2.44 [1.74– 3.40]), and stone/sand (1.97 [1.09–3.55]). Although there are a number of limitations of the case-control design and these results alone do not establish a causal link, an assessment of all of the available evidence strongly suggests that IPF may be a heterogeneous disorder caused by a number of environmental and occupational exposures.

The association observed from available studies provide support for the hypothesis that IPF may be a heterogeneous disorder caused by a number of environmental and occupational exposures. Although causation of IPF may never be directly observable, we can conclude that four proposed mechanisms and potential variations in lung responses have been proposed to contribute to IPF. They include: (1) delivery and persistence of agent, (2) biochemical response, (3) immunological response, and (4) fibrotic response (Nemery, 2001). Figure 1 from the 2006 Proceedings of The American Thoracic Society (found below) explains a prevailing paradigm for the development of IPF due to environmental factors.

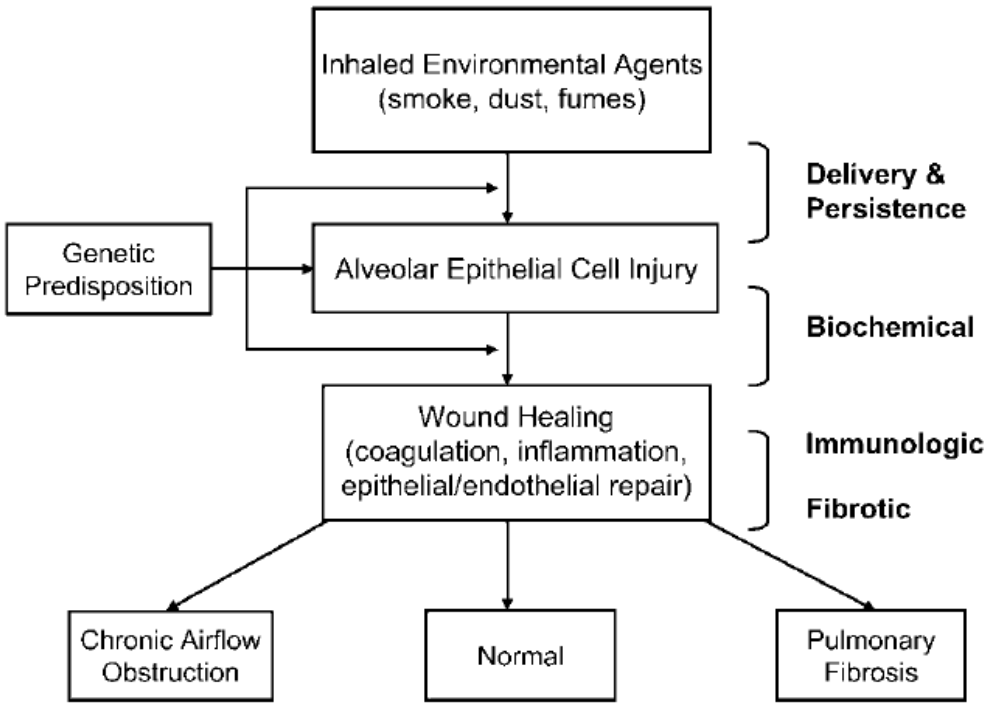

Figure 5: Environmental exposure pathways and IPF

2.5.7. Length of Hospital Stay & Cost:

A 2008 analysis of the HCUP database (the same one we are using) found a steady increase in the number of hospital discharges with IPF. IPF discharges have increased since 1993, while LOS and in-hospital mortality decreased. This is most likely due to increased ability and knowledge to treat IPF across the U.S. Hospital charges increased dramatically during this time, rising from just over $15,000 in 1993 to over $40,000 in 2008. Some of this increase is due to inflation, as $15,000 in 1993 dollars is equal to $22,349 in 2008 dollars. In essence, the cost of treatment for IPF has almost doubled when adjusted for inflation. Hospitalization costs increased 3.8 fold from 1993-2008 and doubled between 2007 and 2008 to $81,000 (Fioret, 2011).

Chapter Three - Methodology

3.1. Nationwide Inpatient Sample Data:

The Nationwide Inpatient Sample (NIS) is the largest all-payer care database in the United States. It contains a sample of discharge records from all HCUP-participating hospitals, and uses the definitions of hospitals and discharges supplied by the statewide data organizations that contribute to HCUP, rather than the definitions used by the AHA Annual Survey. This is done to provide conformity within the data set. NIS data cover all patients, including individuals covered by Medicare, Medicaid, or private insurance, as well as those who are uninsured, so as to comprise an accurate sample of all healthcare patients within the United States. Its large sample size is ideal for developing national and regional estimates and enables analyses of rare conditions, uncommon treatments, and special populations. The data we are using to address the research question ranges 2007 – 2012. NIS data contain information on over 47,133,557 hospital stays across the data range.

Table 2 illustrates the total number of stays by year over the observation period.

Table 2: NIS data observations by year

Year	Study population (stays)
2007	8,043,415
2008	8,158,381
2009	7,810,762
2010	7,800,441
2011	8,023,590
2012	7,296,968
Total	47,133,557

In order to isolate IPF patients, we combined the yearly HCUP data sets and identified any IPF diagnosis via ICD-9 codes. The HCUP data for 2007 and 2008 contained 15 diagnosis codes, while data for 2009 – 2012 contained 25 diagnosis code variables, Within our data set, the particular codes that applied to IPF were as follows: 5163, 51630, and 51631. These represent ICD9 codes of 516.3, 516.30 and 516.31. Codes 516.30 and 516.31 were only used for the calendar year 2011 and 2012. Any patient who received a diagnosis of IPF in any of the available diagnosis variables (whether dx1-dx15 or dx1-dx25 were included in the analysis data set. Of the total HCUP observations, only 18,790 (0.0399% of the total) patients were diagnoses with IPF Table 3 illustrates the distribution of IPF patients by year.

Table of IPF by YEAR							
IPF	**YEAR(Calendar year)**						
	2007	**2008**	**2009**	**2010**	**2011**	**2012**	**Total**
1	2826	3063	2691	2571	3519	4120	18790
Total	2826	3063	2691	2571	3519	4120	18790

Table 3: IPF diagnosis by year

Table 4. Data Variables Used for Analysis:

Study Variables	Original Variable Name in the NIS Data Set	Variable Description
Age	AGE	Age in years, Numerical Variable
Mortality	DIED	Patient did not die during hospitalization (DIED=0); Patient died during hospitalization (DIED=1), Categorical (binary) Variable
GENDER	FEMALE	Gender of patient FEMALE = 1 is Male; FEMALE=0 is female, Categorical (binary) Variable
TOTAL CHARGE	TOTCHG	Total charges , Numerical Variable
RACE	RACE	1 = White, 2 = Black, 3 = Hispanic, 4 = Asian/Pacific, 5 = Native Am., 6 = Other, Categorical Variable
INSURANCE TYPE	PAY1	1=Medicare, 2=Medicaid, 3=Private insurance,4=Self-pay,5=No charge,6=Other, Categorical Variable
NUMBER OF PROCEDURES	NPR	The number of procedures performed while patient was hospitalized, Numerical Variable
SOCIO_ECONOMIC STATUS (SES)	ZIPINC	Median household income for patient's ZIP Code, 1=$1-24,999, 2=$25,000-34,999, 3=$35,000-44,999, 4=45,000 or more, Categorical Variable

COMORBIDITIES	CM_DRUG, CM_ALCOHOL, CM_OBESE, CM_ULCER, CM_DM, CM_HTN	Comorbidities (drug abuse, alcohol abuse, obesity, ulcer, diabetes, hypertension), Categorical (binary) Variables
LENGTH OF STAY	LOS	The number of days patient was hospitalized, Numerical Variable
NUMBER OF DIAGNOSES	NDX	The number of diagnoses on the patient record, Numerical Variable
REGION	REGION	Four regions are included Northeast = 1, Midwest =2, South = 3, west =4 , Categorical Variable

3.2. Goals and Objectives:

The goal of the initial analysis is to identify any factors and costs associated with IPF patients as it pertains to mortality, length of stay, and costs in various clinical settings across the US.

3.3. Research Design & Methods:

The following research questions were asked as part of this analysis:

- Are there statistically significant associations between the number and types of comorbidities and procedures and mortality, costs and length of stay of IPF patients?
- Are there statistically significant differences in mortality, costs and length of stay of IPF patients with race, age, or socio-economic status?
- Are there statistically significant differences in the mortality, costs and length of stay of IPF patients across the various regions of the US?

- Are there statistically significant differences in the mortality, costs and length of stay of IPF patients amongst the different types of hospital settings – rural/urban/hospital with and without teaching?

3.4. Statistical Methodology:

The following methods will be used to analyze the data as appropriate.

The following parametric methods will be used to analyze continuous data that are normally distributed within our data set:

- Linear regression models
- Correlation analysis: Pearson correlation
- Paired and unpaired t-test
- One-way ANOVA and
- Mean, SD for descriptive analyses

The following non-parametric methods will be used to analyze variables that are not normally distributed within, or those data that are ranks or scores.

- Wilcoxon Rank sum test and Mann-Whitney test
- Correlation analysis: Spearman correlation
- Kruskal-Wallis test and
- Median, interquartile range for descriptive analyses

For binary outcomes and categorical variables, we will use the following methods, as appropriate.

- Logistic regression models.
- Chi-square test or Fisher exact test.
- Contingency coefficients (Cochran-Mantel-Haenszel tests).

- McNemar's test
- Proportion for descriptive analyses.

3.5. Statistical Analyses:

We will combine all of the data points from the 2007 – 2012 through concatenation in SAS, and will examine key variables by each year to determine unrealistic or erroneous values. IPF cases will be selected using the 5163, 51630, and 51631 diagnosis codes within the DX1 – DX25 fields. The analysis data set will consist of only cases of IPF, isolated by the aforementioned codes. Data will be categorized as appropriate to investigate research questions. All computations will be performed with SAS® Release 9.3 running on the Windows 8 operating system. All invalid data will be reported and a reason given for why the data is considered invalid (example –missing value). Where outlying data are observed, analyses will be performed with and without the outlying data. Sound statistical evidence that the data are outlying (i.e. outlying data is more than 4 standard deviations beyond the mean of comparable data) will be documented. Outlying data can be removed from an analysis if it can be shown to improve the power of the statistical tests or if not removing it would skew the result.

We will then perform a descriptive statistical for the remaining variables Continuous variables will be assessed for normality using the UNIVARIATE procedure in SAS. PROC UNIVARIATE provides descriptive analysis and measures of central tendency (mean, median, mode, standard deviation, variance, range, interquartile range) for a single variable within a dataset. Through the same procedure, we will also test for

normality and record extreme values for each variable. Extreme values are those values that lie at both ends of the numerical range of a continuous variable.

If the data is normally distributed, parametric methods will be used to analyze data otherwise non-parametric methods will be used. Non-parametric methods will be used to analyze score data. We will group numerical variables into categories, according to their distribution (tertiles, quartiles, quintiles, categories of 5 and 10, etc). Categorical analyses with the appropriate methods will be used to compare categorical variables. Cochran-Mantel-Haenszel tests (for categorical variables) or linear models (for continuous variables) will be used to compare the baseline clinical characteristics. Relationships between outcome and clinical characteristics will be tested by using Pearson correlations.

If the data are not normally distributed, nonparametric tests such as Spearman correlation and the Wilcoxon rank sum test will be used where appropriate. We will compare categorical variablesthe chi-square test or Fisher exact test (where appropriate) to make comparisons between or among groups. A two-sample student's t test will be used to compare difference in scores between clinical groups. We will use the CORR procedure within SAS to employ Pearson correlation or Spearman rank correlation coefficients to test independence between variables. Spearman rank correlation coefficients will be used when making comparisons between categorical factors and the continuous research outcomes (length of stay, total charges and mortality). For the comparison of means, the Student t-test will be used, and where appropriate, a paired t-test will be performed.

The following SAS procedures will be used to perform statistical modeling, hypothesis testing, and comparisons among groups: The CORR Procedure, The CATMOD Procedure, The FREQ Procedure, The GLM Procedure, the LOGISTIC Procedure and The MEANS Procedure.

3.6. Modeling Techniques Overview:

The following sections constitute a brief summary of some major statistical modeling techniques such as Cox regression, c-statistic, linear regression, Kolmogoro Smirvov test and Logistic regression were frequently mentioned for predicting outcome models.

3.7. Cox Proportional Hazards Regression:

The Cox proportional hazards model describes the relationship between the time that passes before some event occurs to one or more covariates that may be associated with that quantity of time using the hazard function, $\lambda_0(t)$. The Cox proportional hazards model operates on three key assumptions, as described by Hosmer and Lemeshow, 1999. The linearity assumption states that the relationship between a predictor and the outcome takes a linear functional form. For variables that have a skewed distribution, as is commonly the case with values of laboratory tests or measures, the axis may be transformed by a square root or logarithmic function so as to force the transformed variable to adhere better to the linearity assumption.

The Cox proportional hazard model further assumes that the total effect of different predictors may be estimated simply by summing beta coefficients of the individual effects of each predictor. In cases where a more complex variable interaction is

believed to exist, the researcher or data modeler may create a composite variable by joining two individual variables together in a single term, thus creating a single beta coefficient. The proportional hazards assumption states that the impact of each predictor on survival does not change over time. Extensions to the Cox model exist, such as the use of time-dependent covariates, to allow for situations where this assumption does not hold (Hosmer & Lemeshow, 2001). Cox assumptions are not rigid, and some alternative methods exist to account for situations when they are violated. Said alternative methods require a priori preference for some functional form by the experimenter.

3.8. C-Statistic:

Concordance (C) statistic model translates into a graphical plot which demonstrates the performance of a two-classifications; true positive versus false positive or sensitivity versus specificity. It is also known as a "receiving operating characteristic" or simply ROC curve and used during WWII for radar signals analysis. US military utilized ROC to predict Japanese aircraft from their radar signals (Green, 1966; Uno, 2011). Currently, c-statistic analysis is commonly used in the evaluation of diagnostic tests.

For modern evidence-based medicine, a well thought-out risk scoring system for predicting the occurrence of a clinical event can play a critical role in selecting prevention and treatment strategies. Such an index system is often established based on the subject's "baseline" genetic or clinical markers via a working parametric or semi-parametric model. To evaluate the adequacy of such a system, C-statistics are routinely

used in the medical literature to quantify the capacity of the estimated risk score in discriminating among subjects with different event times. The C-statistic provides a global assessment of a fitted survival model for the continuous event time rather than focuses on the prediction of t-year survival for a fixed time.

3.9. Linear Regression:

Linear regression is modeling the relationship between two variables using the least squares approach. In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Beta coefficients outputted from a statistical software indicate the amount of statistical influence a given predictor or covariate has on the slope of the line between the exposure and the outcome in question. A fitted linear regression model can be used to identify the relationship between a single predictor x and the response variable y.

3.10. Kolmogoro Smirnov Test:

Kolmogoro Smirvov (K-S) test is a nonparametric test for the equality of continuous, one dimensional probability distribution that can be used to compare a sample with a reference distribution or to compare two samples. The K-S statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between empirical distribution functions of two samples. The K-S test seeks to determine if the distribution of the same variable(s) differs significantly between two data sets. The K-S test has the advantage of making no assumption regarding the distribution of the data.

3.11. Logistic Regression:

Logistic regression, or logit regression, or logit model is a direct probability that is used to predict a binary response based on one or more predictor variables. Logistic regression is typically used when the outcome in question is dichotomous (yes/no). Logistic regression can be binomial or multinomial, i.e. dead versus alive, success versus failure, yes versus no, better versus no change versus worse. Generally, the outcome is coded as "0" and "1". Logistics regression was developed by D.R. Cox in 1958 (Cox, 1958; Walker, 1967).A simple logistic function is defined as $P(t) = 1/(1+e^{-t})$, where variable P is considered a population, variable e is Euler's number and variable t is time. Logistic regression is used to predict the odds of being a case based on the values the values of the independent variables. The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non-case.

As with other forms of regression analysis, logistic regression makes use of one or more predictor variables that may be either continuous or categorical. Given that the outcome analyzed in logistic regression has only two levels, it is necessary that the regression take the natural logarithm of the odds of the odds of the dependent variable being a case (referred to as the logit or log-odds) to create a continuous criterion as a transformed version of the dependent variable.

The logit of success (being a case) is then fitted to the predictors using linear regression analysis. The predicted value of the logit is converted back into predicted odds via the inverse of the natural logarithm, which transformation is called the exponential function. Thus, although the observed dependent variable in logistic regression is dichotomous, the logistic regression estimates the odds as a continuous variable that

represents the probability of being a success (case). Logistic regression is typically performed in SAS using PROC LOGISTIC or PROC GENMOD with the logit link.

Chapter Four - RESULTS

4.1 IPF Descriptive Analysis for 200-2012

Total Charge

Tables 5 through 12 and Tables 21 through 28 represent output obtained through the use of the UNIVARIATE procedure to perform a descriptive analysis of key variables (total charge and length of stay) and comparing within a data set restricted to IPF patients only in SAS. Tables 13 through 20 and Tables 29 through 36 are displayed for the purpose of making comparisons between the IPF-only subgroup and the entire NIS/HCUP data set.

Table 5 describes the data set as a whole, providing the total number of observations with IPF (N = 14,395), mean expenditures per patient, the standard deviation for each patient (Std Deviation), total charges for IPF patients across all years measured (Sum Observations), and several other variables related to the overall distribution of the sample (variance, skewness, kurtosis, uncorrected sum of squares, corrected sum of squares, coeffient of variation, and standard error of the mean). Table 6 further provides the portion of SAS output called Basic Statistical Measures, which gives the median, mode, and range of total charges within the data set. Table 7 provides the 95% confidence limits for the mean, standard deviation, and variance of total charge.

Moments			
N	14395	**Sum Weights**	14395
Mean	68925.6892	**Sum Observations**	992185296
Std Deviation	164048.757	**Variance**	2.6912E10
Skewness	9.57275945	**Kurtosis**	134.420702
Uncorrected SS	4.55758E14	**Corrected SS**	3.87371E14
Coeff Variation	238.008148	**Std Error Mean**	1367.31037

Table 5: The UNIVARIATE Procedure Total Charge – IPF

Basic Statistical Measures			
Location		**Variability**	
Mean	68925.69	**Std Deviation**	164049
Median	28257.00	**Variance**	2.6912E10
Mode	8433.25	**Range**	4162790
		Interquartile Range	46014

Table 6: The UNIVARIATE Procedure Total Charge (Basic Statistical Measures) – IPF

Basic Confidence Limits Assuming Normality			
Parameter	**Estimate**	**95% Confidence Limits**	
Mean	68926	66246	71606
Std Deviation	164049	162176	165966
Variance	2.6912E10	2.63009E10	2.75447E10

Table 7: The UNIVARIATE Procedure Total Charge Basic Confidence Limits Assuming Normality - IPF

Table 8 provides three tests of location for the variable Total Charge, Student's t test, the sign test, and Wilcoxon signed rank rest. All three tests produce a test statistic for the null hypothesis that the mean or median is equal to a given population mean μ_0 against the alternative that the mean or median is not equal the value μ_0. Student's t test is appropriate when the data are from an approximately normal population; otherwise nonparametric tests such as the sign test or the signed rank test should be used. The results of these three tests seem to indicate that the measures of central tendency computed in the tables above are in agreement with the population mean.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	50.40969	**Pr > |t|**	<.0001
Sign	**M**	7197.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	51807605	**Pr >= |S|**	<.0001

Table 8: The UNIVARIATE Procedure Total Charge (Test of Location) - IPF

Table 9 provides the tests of normality as performed by PROC UNIVARIATE. Smaller p-values in these three tests indicate a severe lack of normality in the distribution. The three tests outlined in these tables all produced statistically significant p-values, meaning that the distribution of total charges for IPF was significant from the normal (bell curve) distribution.

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.337884	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	554.653	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2752.925	**Pr > A-Sq**	<0.0050

Table 9: The UNIVARIATE Procedure Total Charge (Test for Normality) - IPF

Table 10 provides information pertaining to data quartiles and other percentiles for the total charge variable. The minimum charge for a patient with an IPF diagnosis was $58.49. The maximum charge for someone with an IPF diagnosis was $4,162,849, a markedly higher cost. The table outlines the interquartile as the distance between quartile 1, and quartile 3 (Q1 and Q3, respectively), which range is $46,014. The top 1% of total charges among IPF patients differs from the 99th percentile by $3,463,991, while the 99th percentile is $457,221 more expensive than the 95th percentile. This phenomenon could be caused by a lower quantity of individuals who were diagnosed with IPF as a primary diagnosis who might have survived longer than others or undergone very expensive surgery to attempt to slow the progression of the disease. Table X7 provides the values for the top (highest) and bottom 5 (lowest) IPF-related charges within the data set and provides the number of each of these observations.

Finally, Table X8 gives the number of observations in the total charge variable that had missing values (4,395). It is important to note that 23.39% of all observations were missing total charge information in this combined data set.

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	4162849.00
99%	698858.00
95%	241637.00
90%	140266.00
75% Q3	60821.00
50% Median	28257.00
25% Q1	14807.00
10%	8556.02
5%	6341.00
1%	3235.73
0% Min	59.49

Table 10: The UNIVARIATE Procedure Total Charge Quantiles - IPF

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
59.49	1091	2866021	13719
63.54	1026	3117820	5304
63.55	1089	3240276	13778
74.79	1086	3417425	5275
86.63	1095	4162849	3324

Table 11: The UNIVARIATE Procedure Total Charge (Extreme Observations) - IPF

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	4395	23.39	100.00

Table 12: The UNIVARIATE Procedure Total Charge (Missing Values) – IPF

We compared the univariate distribution of total charge among IPF patients to the entire NIS/HCUP population by running the UNIVARIATE Procedure on the entire NIS/HCUP data set. The results are found in Tables 13-20 (below). We immediately see that the average cost per hospital admission among IPF patients was more than two times (2.17 exactly) that of all patients in the NIS/HCUP (Tables 13 and 14).

Moments			
N	46221782	**Sum Weights**	46221782
Mean	31742.4125	**Sum Observations**	1.46719E12
Std Deviation	57598.4095	**Variance**	3317576774
Skewness	11.6250662	**Kurtosis**	334.689735
Uncorrected SS	1.99916E17	**Corrected SS**	1.53344E17
Coeff Variation	181.455677	**Std Error Mean**	8.47202344

Table 13: The UNIVARIATE Procedure Total Charge – NIS

Basic Statistical Measures			
Location		**Variability**	
Mean	31742.41	**Std Deviation**	57598
Median	16758.00	**Variance**	3317576774
Mode	7170.00	**Range**	4993914
		Interquartile Range	26746

Table 14: The UNIVARIATE Procedure Total Charge (Basic Statistical Measures) – NIS

Table 15 provides three tests of location for the variable Total Charge among the NIS population, Student's t test, the sign test, and Wilcoxon signed rank rest. All three tests produce a test statistic for the null hypothesis that the mean or median is equal to a given population mean μ_0 against the alternative that the mean or median is not equal the value μ_0. The results of these three tests seem to indicate that the measures of central tendency computed in the tables above are in agreement with the population mean.

Basic Confidence Limits Assuming Normality			
Parameter	**Estimate**	**95% Confidence Limits**	
Mean	31742	31726	31759
Std Deviation	57598	57587	.
Variance	3317576774	3316224616	.

Table 15: The UNIVARIATE Procedure Total Charge Basic Confidence Limits Assuming Normality – NIS

Table 16 provides the tests of normality as performed by PROC UNIVARIATE. As before, smaller p-values in these three tests indicate a severe lack of normality in the

distribution. The three tests outlined in these tables all produced statistically significant p-values, meaning that the distribution of total charges for NIS/HCUP was significant from the normal (bell curve) distribution.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	3746.733	**Pr > |t|**	<.0001
Sign	**M**	23110891	**Pr >= |M|**	<.0001
Signed Rank	**S**	5.341E14	**Pr >= |S|**	<.0001

Table 16: The UNIVARIATE Procedure Total Charges (Test of Location) – NIS

As with the IPF population, the highest percentiles of total charges among NIS/HCUP patients was highly skewed. The highest charge in the data set was nearly $5 million, while many in the lower percentile had charges that were recorded only as $100. It would appear that, when a charge was recorded, the default selection for any charge whatsoever was $100. There are no observations between $0 and $100. It is of great alarm also that only 1.93% of the total charge observations were missing in this, the much larger data set, as opposed to the IPF case-restricted data which was missing 23.39% of its total charge data points.

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.293368	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	1243813	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	6448157	**Pr > A-Sq**	<0.0050

Table 17: The UNIVARIATE Procedure Total Charge (Tests for Normality) – NIS

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	4994014
99%	244690
95%	103747
90%	67942
75% Q3	34811
50% Median	16758
25% Q1	8065
10%	3668
5%	2254
1%	1174
0% Min	100

Table 18: The UNIVARIATE Procedure Total Charge Quantiles – NIS

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
100	4.69E7	4954306	2.07E7
100	4.63E7	4956982	2.96E7
100	4.63E7	4958781	3.13E7
100	4.52E7	4972660	1.74E7
100	4.51E7	4994014	4.69E7

Table 19: The UNIVARIATE Procedure Total Charge (Extreme Observations) – NIS

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	911775	1.93	100.00

Table 20: The UNIVARIATE Procedure Total Charge (Missing Values) – NIS

Length of Stay

The next set of tables provides the SAS output for the UNIVARIATE procedure pertaining to the length of stay variable. All tables from this output mirror those of the output from the Total Charge variable, thus explanations are limited. Table 21 provides a high-level overview of the distribution of the Length of Stay variable. As can be seen, the Length of Stay variable experienced a standard deviation that was greater than the mean length of stay. A Skewness of 44.11 indicates that the mean value for Length of Stay was greater than the median and the mode. Such a high value for skewness indicates that the right tail of the distribution is much, much longer than the left, meaning that the values for Length of Stay are very heavily skewed to the right.

A kurtosis value of 0 indicates that the distribution of the variable maintains a Gaussian (normal) distribution, while a negative value indicates a flatter distribution, and a positive value indicates a more peaked distribution. In this data set, as it pertains to Length of Stay, the distribution is very peaked, meaning that the vast majority of values fall within a very small range, despite the standard deviation seen above, which could have been caused by a few to several dozen very long hospital stays.

The UNIVARIATE Procedure
Variable: LOS_X (Length of stay (as received from source))

Moments			
N	14601	**Sum Weights**	14601
Mean	7.84432573	**Sum Observations**	114535
Std Deviation	13.6041824	**Variance**	185.073778
Skewness	44.1082144	**Kurtosis**	3530.79218
Uncorrected SS	3600527	**Corrected SS**	2702077.15
Coeff Variation	173.427046	**Std Error Mean**	0.11258516

Table 21: The UNIVARIATE Procedure of (Length of Stay) - IPF

Table 22 provides a further illustration of the location and variability of the data (the combination of which is known as the distribution). This representation adds the median, mode, range, and interquartile range. It can be seen that the interquartile range is 6.0 days, meaning that 75% of the observations experienced a length of stay of 6 days or less. The most common length of stay is 3 days. We can also see that the mean measurement is over 50% greater than the median, as was illustrated in the "Moments" table above. Table 23 provides the confidence limits for mean, standard deviation, and variance.

Basic Statistical Measures			
Location		**Variability**	
Mean	7.844326	**Std Deviation**	13.60418
Median	5.000000	**Variance**	185.07378
Mode	3.000000	**Range**	1158
		Interquartile Range	6.00000

Table 22: The UNIVARIATE Procedure of LOS (Basic Statistical Measures) - IPF

Basic Confidence Limits Assuming Normality			
Parameter	**Estimate**	**95% Confidence Limits**	
Mean	7.84433	7.62364	8.06501
Std Deviation	13.60418	13.44993	13.76204
Variance	185.07378	180.90056	189.39378

Table 23: The UNIVARIATE Procedure of LOS (Basic Confidence Limits Assuming Normality) - IPF

Tables 24 and 25 provide the tests of location and tests of normality for Length of Stay. The test of location are very similar to those of the Total Charge variable, meaning that the Length of Stay in this sample are in agreement with the population mean. The tests for normality all indicate that the distribution of Length of Stay departs severely from the normal distribution.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	69.6746	**Pr > |t|**	<.0001
Sign	**M**	7231.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	52298208	**Pr >= |S|**	<.0001

Table 24: The UNIVARIATE Procedure of LOS (Test for Location) - IPF

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.297994	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	388.0543	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2032.185	**Pr > A-Sq**	<0.0050

Table 25: The UNIVARIATE Procedure of LOS (Tests for Normality) - IPF

Table 26 provides a description of the percentiles of Length of Stay. This table corroborates the hypothesis that there was a very long (and possibly expensive) stay for at least one IPF patient, which stay lasted 1,158 days. The median stay is indicated by the "50% Median" row label. The interquartile range for Length of Stay is 6 days, also as indicated above. We can also see that stay lengths increase dramatically in the 99th percentile, as compared to the 95th. 99th percentile stays were 45 days, which was double the stay of a 95th percentile patient. This finding seems to indicate why the LOS and Total Charge variables were highly skewed.

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	1158
99%	45
95%	22
90%	16
75% Q3	9
50% Median	5
25% Q1	3
10%	2
5%	1
1%	1
0% Min	0

Table 26: The UNIVARIATE Procedure of LOS Quantiles - IPF

Tables 27 and 28 display extreme values (five lowest and five highest) and missing values for the Length of Stay variable. The lowest five stays were of zero days, which the highest observations varied greatly in length. This table also reveals that the stay of 1,158 days was (by far) the longest, in fact, by 858 days. Also, there were 4,189 missing observations, which comprised 22.29% of the sample.

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	14578	160	3156
0	14492	196	9352
0	14433	250	3324
0	14400	300	4693
0	14130	1158	8739

Table 27: The UNIVARIATE Procedure of LOS (Extreme Observations) - IPF

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	4189	22.29	100.00

Table 28: The UNIVARIATE Procedure of LOS (Missing Values) - IPF

As with total charge, we sought to compare the univariate distribution of the variable in question between the IPF-only and the NIS/HCUP data sets. Tables 29-35 provide the appropriate data to compare these two groups. As with the total charge variable, we can see that the IPF group stays in the hospital 71% longer than the rest of the NIS/HCUP population.

Moments			
N	47124400	**Sum Weights**	47124400
Mean	4.58166364	**Sum Observations**	215908150
Std Deviation	6.81089817	**Variance**	46.3883339
Skewness	11.2451743	**Kurtosis**	286.266472
Uncorrected SS	3175240874	**Corrected SS**	2186022354
Coeff Variation	148.655569	**Std Error Mean**	0.00099216

Table 29: The UNIVARIATE Procedure (Length of Stay) – NIS

Basic Statistical Measures			
Location		**Variability**	
Mean	4.581664	**Std Deviation**	6.81090
Median	3.000000	**Variance**	46.38833
Mode	2.000000	**Range**	365.00000
		Interquartile Range	3.00000

Table 30: The UNIVARIATE Procedure of LOS (Basic Statistical Measures) NIS

Basic Confidence Limits Assuming Normality			
Parameter	**Estimate**	**95% Confidence Limits**	
Mean	4.58166	4.57972	4.58361
Std Deviation	6.81090	6.80952	.
Variance	46.38833	46.36961	.

Table 31: The UNIVARIATE Procedure of LOS (Basic Confidence Limits Assuming Normality) – NIS

Tables 32 and 33 again provide the tests of location and tests of normality for Length of Stay. The test of location are very similar to those of the Total Charge variables and of the length of stay in the IPF group, meaning that the Length of Stay in the NIS/HCUP sample are in agreement with the population mean. The tests for normality all indicate that the distribution of Length of Stay departs severely from the normal distribution. The final striking finding in the comparison of LOS between the IPF and NIS/HCUP groups is the level of missing data. As with total charge, the IPF group was missing 22.29% of its total observations, while the NIS/HCUP is only missing 0.02% of its LOS observations. This indicates that total charge and length of stay are missing at an unusually high level among IPF patients.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	4617.87	**Pr > |t|**	<.0001
Sign	**M**	23103571	**Pr >= |M|**	<.0001
Signed Rank	**S**	5.338E14	**Pr >= |S|**	<.0001

Table 32: The UNIVARIATE Procedure of LOS (Tests for Location) – NIS

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.280024	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	1212049	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	6285496	**Pr > A-Sq**	<0.0050

Table 33: The UNIVARIATE Procedure of LOS (Tests for Normality) – NIS

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	365
99%	29
95%	14
90%	9
75% Q3	5
50% Median	3
25% Q1	2
10%	1
5%	1
1%	0
0% Min	0

Table 34: The UNIVARIATE Procedure of LOS Quantiles (Definition 5) – NIS

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0	4.71E7	365	4.51E7
0	4.71E7	365	4.51E7
0	4.71E7	365	4.51E7
0	4.71E7	365	4.51E7
0	4.71E7	365	4.7E7

Table 35: The UNIVARIATE PROCEDURE of LOS (Extreme Observations) - NIS

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	9157	0.02	100.00

Table 36: The UNIVARIATE Procedure of LOS (Missing Values) – NIS

Tables 37 outlines the distribution of age by death. We can see that the vast majority of deaths (2013 or 97.48%) occur after the age of 45 years, which is consistent with the pattern of death described in the IPF literature. 73.02% of all deaths occurred within the Over 65 years category. The percentage of people who die vs those who do not die increased in each category. The following percentages of patients died in each age category:

- Less than 16 years: 2.899%
- 16 – 25 years 2.985%
- 26 – 35 years 6.025%
- 46 – 55 years 9.732%
- 56 – 65 years 11.348%
- Over 65 years 11.395%

Table of agecat by DIED			
agecat(Age of Patient)	DIED(Died during hospitalization)		
	0	1	Total
Less than 16 years	67	2	69
16 – 25 years	65	2	67
26 – 35 years	195	13	208
36 – 45 years	497	35	532
46 – 55 years	1382	149	1531
56 – 65 years	2781	356	3137
Over 65 years	11726	1508	13234
Total	16713	2065	18778
Frequency Missing = 12			

Table 37: The FREQ Procedure (AGE by DIED) - IPF

Tables 38 and 39 display the primary expected payor (PAY1) for both the NIS/HCUP and IPF data configurations. It becomes very clear that the IPF data set is heavily skewed toward the Medicare eligible population. This is because IPF is a disease that primarily occurs in older individuals. The 5.65% on Medicaid are most likely disabled (most likely as the IPF population is older) or severely disadvantaged individuals. We also assume that the private insurance group are older, but not yet eligible for Medicare. The payor distirubiton is much more balanced among the NIS/HCUP database with nearly equal proportions receiving care through Medicare and Private Insurance, and 20.05% being covered by Medicaid. The distribution is most likely

affected by the presence of births in the general population, who would almost exclusively fall within the private insurance, and Medicaid brackets. It is highly unlikely that anyone on Medicare would give birth, as to be eligible to receive Medicare, you must be well beyond the child-bearing age.

Primary expected payer (uniform)				
PAY1	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Medicare	13613	72.54	13613	72.54
Medicaid	1061	5.65	14674	78.20
Private Insurance	3409	18.17	18083	96.37
Self-pay	296	1.58	18379	97.94
No charge	27	0.14	18406	98.09
Other	359	1.91	18765	100.00
Frequency Missing = 25				

Table 38: The FREQ Procedure – Primary payor/insurance (PAY1) - IPF

Primary expected payer (uniform)				
PAY1	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Medicare	17801096	37.85	17801096	37.85
Medicaid	9427013	20.05	27228109	57.9
Private Insurance	15522695	33.01	42750804	90.91
Self-pay	2450690	5.21	45201494	96.12
No charge	232915	0.5	45434409	96.61
Other	1592273	3.39	47026682	100
Total	**Frequency Missing = 106875**			

Table 39: The FREQ Procedure – Pay1 NIS

Tables 40 and 41 outline the distribution of primary insurance provider by race. It is striking that the vast majority of those on Medicare are white. Meanwhile, that percentages drops dramatically for Medicaid, where the balance is picked up by the Black and Hispanic populations. This phenomenon would indicate that the Black and Hispanic populations either: 1) do not seek care as frequently as whites, or 2) blacks and Hispanics are not insured as commonly as whites unless they are participating in income-based entitlement programs. The tables also indicate that blacks and Hispanics self-pay or are not charged at a higher rate than whites, which only corroborates our assumptions.

Table of PAY1 by RACE							
	RACE(Race (uniform))						
Pay1	White	Black	Hispanic	Asian PI	Native AM	Other	Total
Medicare	9261	803	815	180	82	272	11413
	81.14	7.04	7.14	1.58	0.72	2.38	
Medicaid	379	194	234	27	17	40	891
	42.54	21.77	26.26	3.03	1.91	4.49	
Private Insurance	2053	342	229	80	15	78	2797
	73.4	12.23	8.19	2.86	0.54	2.79	
Self-Pay	127	52	43	7	3	12	244
	52.05	21.31	17.62	2.87	1.23	4.92	
No charge	11	4	9	0	0	2	26
	42.31	15.38	34.62	0	0	7.69	
Other	237	39	25	1	6	10	318
	74.53	12.26	7.86	0.31	1.89	3.14	
Total	12068	1434	1355	295	123	414	15689
	76.92	9.14	8.64	1.88	0.78	2.64	100
Frequency Missing = 3101							

Table 40: The FREQ Procedure – Pay1 by RACE - IPF

Table of PAY1 by RACE							
	RACE(Race (uniform))						
Payor	White	Black	Hispanic	Asian PI	Native AM	Other	Total
Medicare	1.17E+07	1855835	971460	257541	86988	343535	1.52E+07
	76.93	12.18	6.38	1.69	0.57	2.25	
Medicaid	3337853	1883016	2120525	228498	89221	415962	8075075
	41.34	23.32	26.26	2.83	1.1	5.15	
Private Insurance	9213920	1399583	1227387	503493	76774	472287	1.29E+07
	71.46	10.85	9.52	3.91	0.6	3.66	
Self-Pay	1092284	423064	411893	55095	17156	121844	2121336
	51.49	19.94	19.42	2.6	0.81	5.74	
No charge	106195	42048	55235	2605	1700	9013	216796
	48.98	19.4	25.48	1.2	0.78	4.16	
Other	839167	196644	184656	29206	21420	49884	1320977
	63.53	14.89	13.98	2.21	1.62	3.78	
Total	2.63E+07	5800190	4971156	1076438	293259	1412525	3.99E+07
	66	14.55	12.47	2.7	0.74	3.54	100
Frequency Missing = 7271006							

Table 41: The FREQ Procedure – Pay1 by RACE - NIS

Table 42 outlines the distribution of deaths among payment status. A total of 1435 deaths (69.62%) occurred within the Medicare category. This is consistent with the trend in the age category. Medicare beneficiaries are over the age of 65, which provides almost an exact match as the number of deaths in the Over 65 years category above. 418 of the total deaths were paid via private insurance. Table 43 illustrates the distribution among the NIS database. In both the IPF and NIS/HCUP groups, more than half the people who die are on Medicare.

Table of PAY1 by DIED			
PAY1(Insurance Type)	DIED(Died during hospitalization)		
	0	1	Total
Medicare	12173	1435	13608
Medicaid	947	112	1059
Private Insurance	2987	418	3405
Self-pay	271	24	295
No charge	23	4	27
Other	291	68	359
Total	16692	2061	18753
Frequency Missing = 37			

Table 42: The FREQ Procedure (Pay1 by DIED) - IPF

Table of PAY1 by DIED			
Pay1	0	1	Total
Medicare	1.72E+07	598946	1.78E+07
Medicaid	9345747	76620	9422367
Private Insurance	1.54E+07	159344	1.55E+07
Self-pay	2416371	33030	2449401
No charge	230012	2597	232609
Other	1557442	31761	1589203
Total	4.61E+07	902298	4.70E+07
Frequency Missing = 139663			

Table 43: The FREQ Procedure (Pay1 by DIED) - NIS

Table 44 outlines the distribution of race by death category (died vs not). 77.5% of all those who died were white. The following percentages of patients died within each category:

- White 11.24%
- Black 9.47%
- Hispanic 11.23%
- Asian/Pacific 12.20%
- Native American 6.92%
- Other 13.04%

Table of RACE by DIED			
RACE(Patient Race/Ethnicity)	**DIED(Died during hospitalization)**		
	0	**1**	**Total**
White	10719	1357	12076
Black	1299	136	1435
Hispanic	1201	152	1353
Asian/Pacific	259	36	295
Native Am.	107	16	123
Other	360	54	414
Total	13945	1751	15696
Frequency Missing = 3094			

Table 44: The FREQ Procedure Statistics for RACE by DIED – IPF

Tables 45 and 46 display the total charge, length of stay, and number of diagnoses by payor for IPF and NIS/HCUP respectively. Even when stratifying by payor, the IPF group is more expensive, carries more diagnoses, and stays longer in the hospital. This evidence is further corroboration of the extreme financial burden that being diagnosed with IPF can have on an individual or family. It is also of note that the standard deviations of total charges among IPF patients are extremely wide and therefore largely inconclusive. The mean can be considered reliable as the average charge across the distribution; but we have no real understanding of the range of costs to diagnose and treat IPF. According to the data, we can only be confident that 95% of the distribution lies somewhere between 0 and $286,892.704 (the upper limit on the 95% Confidence Interval associated with the mean). Unfortunately, the only way to develop a true measurement of the total cost is to incur more IPF patients.

Table 45: Total charges, length of stay, and number of diagnoses by Payor - IPF-only

	Total Charges			Length of Stay			Number of Diagnoses		
Payor	N	Mean	Std Dev	N	Mean	Std Dev	N	Mean	Std Dev
Medicare	13378	$56,956.48	$117,314.01	13612	7.17	8.25	13613	13.71	5.52
Medicaid	1050	$74,146.35	$161,434.93	1061	8.97	14.07	1061	12.11	5.67
Private	3333	$100,138.66	$190,027.32	3408	8.86	12.26	3409	12.17	5.70
Self-pay	292	$49,289.70	$72,700.99	296	7.36	8.29	296	11.04	5.33
No charge	27	$76,086.70	$99,254.65	27	12.41	18.18	27	12.00	6.87
Other	359	$105,342.01	$261,594.84	359	9.63	13.39	359	13.31	6.06
MISSING	25	$35,364.84	$34,658.65	25	6.48	5.80	25	13.76	6.94

Table 46: Total charges, length of stay, and number of diagnoses by payor - NIS

	Total Charges			Length of Stay			Number of Diagnoses		
Payor	N	Mean	Std Dev	N	Mean	Std Dev	N	Mean	Std Dev
Medicare	17501333	$38,795.80	$57,239.82	17799059	5.48	6.50	17801096	10.98	5.19
Medicaid	9341000	$24,269.56	$62,109.75	9425394	4.37	8.59	9427013	5.73	4.48
Private	15030857	$28,989.67	$55,275.16	15518464	3.82	5.86	15522695	6.14	4.45
Self-pay	2425854	$25,875.29	$45,457.70	2449822	3.88	5.81	2450690	6.33	4.31
No charge	231956	$29,135.12	$47,597.36	232815	4.51	7.32	232915	6.75	4.42
Other	1585423	$33,624.68	$64,332.38	1591976	4.36	7.09	1592273	6.73	4.69
MISSING	105359	$27,848.20	$57,324.63	106870	4.42	6.62	106875	7.25	4.63

Tables 47 and 48 compare the distribution of total charge, length of stay, number of diagnoses (NDX), severity of disease mix (as per DRG), and the patient's risk of mortality by whether or not they died during hospitalization. It becomes apparent quickly that total charges among those who die are much higher (approximately twice among IPF and 2.3 times among the NIS/HCUP database). Length of stay is approximately twice as long when a patient dies in either group. Statistically speaking, the number of diagnoses is the same in the IPF population. However, in the NIS population, this is not the case. The number of diagnoses is nearly twice as much in those who die. Interestingly, severity was not dramatically higher among those experiencing death vs those who did not in the IPF group. This seems to indicate that patients with IPF present at the hospital already ill and either accept the treatment well or risk death. Risk of mortality in the NIS group is more than double among those who die compared to those who did not.

Table 47: Total charge, LOS, number of diagnoses, severity, and risk of mortality by death status - IPF

Variable	Died = 0			Died = 1		
	N	Mean	Std Dev	N	Mean	Std Dev
Total Charge	16450	$60,560.10	$133,539.84	2002	$115,901.85	$180,161.01
Length of Stay	16713	7.17	8.68	2064	11.44	14.95
Number of diagnoses	16713	13.03	5.52	2065	15.43	5.95
Severity	14018	2.94	0.74	1698	3.63	0.57
Risk of mortality	14018	2.71	0.76	1698	3.48	0.63

Table 48: Total charge, LOS, number of diagnoses, severity, and risk of mortality by death - NIS

Variable	Died = 0			Died = 1		
	N	Mean	Std Dev	N	Mean	Std Dev
Total Charge	45304152	$30,771.09	54030.66	885006	$81,620.75	145564.51
Length of Stay	46191406	4.51	6.58	904799	8.29	13.78
Number Diagnoses	46195325	7.81	5.24	905297	13.72	6.11
Severity	38207471	1.95	0.89	739768	3.47	0.76
Risk of mortality	38207471	1.58	0.85	739768	3.45	0.78

Table outlines the primary diagnosis by whether or not a person died. A total of 620 deaths occurred in patients with a diagnosis code consistent with IPF. These deaths include those who perished as a result of idiopathic interstitial pneumonia, idiopathic interstitial pneumonia, not otherwise specified, and idiopathic pulmonary fibrosis. More than 100 deaths occurred in four other diagnosis categories, namely Acute Respiratory Failure (254 deaths), Acute and chronic respiratory failure (174 deaths), "Unspecified" (151 deaths),and pneumonia (126 deaths).

Table of DX1 by DIED		
	DIED(Died during hospitalization)	
DX1(Diagnosis 1)	**0**	**1**
Unspecified	366	151
Subendocardial infacrtion, initial episode of care	199	23
Coronary atherosclerosis of native coronary artery	172	6
Pulmonary embolism and infarction	130	24
Other chronic pulmonary heart diseases	99	9
Atrial fibrillation	228	14
Congestive heart failure	374	28
Acute systolic heart failure	23	0
Acute on chronic systolic heart failure	125	8
Acute on chronic diastolic heart failure	202	8
Acute bronchitis	99	1
Pneumonia	1393	126
Obstructive chronic bronchitis without exacerbation	10	0
Obstructive chronic bronchitis with (acute) exacerbation	698	25
Obstructive chronic bronchitis with acute bronchitis	196	3
Chronic obstructive asthma with (acute) exacerbation"	121	7
Pneumonitis due to inhalation of food or vomitus	160	29
Post inflammatory pulmonary fibrosis	168	38
Idiopathic interstitial pneumonia	2802	450

Idiopathic interstitial pneumonia, not otherwise specific	36	5
Idiopathic pulmonary fibrosis	1196	164
Other specified alveolar and parietoalveolar pneumonopathies	100	17
Acute respiratory failure	487	254
Other pulmonary insufficiency, not elsewhere classified	12	2
Chronic respiratory failure	24	4
Acute and chronic respiratory failure	518	174
Acute kidney failure with lesion of tubular necrosis	14	1
Acute kidney failure, unspecified	170	19
Urinary tract infection, site not specified	152	4
Total from common causes	10274	1594
Total from all admissions	16708	2065
Frequency Missing = 17		

Table 49: The FREQ Procedure of DX1 by DIED

Table 50 displays the various common diagnoses by age among IPF patients. This table provide overwhelming evidence that IPF and many of its associated diseases, conditions, and complication are age-related. Twenty eight conditions are listed on Table 50. Of these twenty eight, at twenty-two of them have at least 100 cases. In all but three of these, no more than five cases appear before the age 36-45 bracket. We should probably assume that the majority of these cases occur in the later years of this age range for one reason. When the age range crosses into 46-55 years, the number of diagnoses at least double in almost every category. However, all diagnoses of all twenty eight of the listed diseases and conditions are only 14.4% (n = 1,466) of the total disease burden for all age groups. In other words, the 56-65 and over 65 categories contribute 85.6% (n = 10,196) of the total disease burden from common causes among IPF patients.

Table 50: Age category among common diagnoses - IPF								
	agecat(Age of Patient)							
DX1(Diagnosis 1)	>16	16-25	26-35	36-45	46-55	56-65	65+	Total
Unspecified	0	0	4	14	44	74	382	518
Subendocardial infacrtion, initial episode of care	0	0	0	0	5	25	192	222
Coronary atherosclerosis of native coronary artery	0	0	1	0	8	41	128	178
Pulmonary embolism and infarction	0	1	0	2	16	23	112	154
Paroxysmal ventricular tachycardia	0	0	0	0	0	3	19	22
Atrial fibrillation	0	0	1	2	6	19	214	242
Atrial flutter	0	0	0	0	0	3	31	34
Congestive heart failure	0	0	1	3	22	42	334	402
Acute systolic heart failure	0	0	0	2	2	3	16	23
Acute on chronic systolic heart failure	0	0	0	0	6	11	116	133
Acute on chronic diastolic heart failure	0	0	0	2	3	19	187	211
Acute bronchitis	0	1	1	2	10	11	75	100
Pneumonia	3	3	16	54	125	201	1117	1519
Obstructive chronic bronchitis with (acute) exacerbation	0	0	2	5	45	131	540	723
Obstructive chronic bronchitis with acute bronchitis”	0	0	1	3	12	30	153	199
Chronic obstructive asthma with (acute) exacerbation	0	0	3	9	20	16	80	128
Pneumonitis due to inhalation of food or vomitus	0	1	0	1	6	18	163	189
Post inflammatory pulmonary fibrosis	2	0	2	9	19	44	130	206
Idiopathic interstitial pneumonia	16	7	45	126	307	721	2034	3256
Idiopathic interstitial pneumonia, not otherwise specific	1	0	1	2	5	11	21	41
Idiopathic pulmonary fibrosis	10	3	11	50	105	299	883	1361
Other specified alveolar and parietoalveolar pneumonopathies	0	0	2	7	17	27	64	117
Acute respiratory failure	1	4	5	18	87	123	503	741
Other pulmonary insufficiency, not elsewhere classified	1	0	1	0	5	0	7	14
Chronic respiratory failure	0	0	0	1	5	6	16	28
Acute and chronic respiratory failure	1	1	4	15	71	138	463	693
Acute kidney failure with lesion of tubular necrosis	0	0	0	0	3	6	6	15
Acute kidney failure, unspecified	1	0	5	4	14	29	136	189
Urinary tract infection, site not specified	0	1	1	0	2	8	144	156
Total from common causes	36	22	107	331	970	2082	8266	11814
Total from all causes	70	67	208	532	1534	3137	13237	18785

Charge, LOS, and Number of Diagnoses among most common primary diagnoses

We also examined the distribution of total charge, LOS, and the number of diagnoses among the top diagnoses within both groups. Table 51 examines the distribution of these three variables among IPF patients. As previously mentioned, the IPF group was significantly older than the mean age of the entire NIS/HCUP data set and, as such, eliminated any of the birth diagnoses that were so prevalent within the NIS/HCUP data set. The most common diagnosis among the IPF-only group was Interstitial Pulmonary Fibrosis which, as we identified in the methodology section of this document, as considered to be an IPF-related diagnosis code. IPF was only diagnosed approximately 41% as much. It becomes apparent by the cost figures that IPF and its related diseases are much more expensive to treat and require longer hospital stays than many other diseases. The average interstitial pulmonary fibrosis patient stayed 8.84 days and accrued an average cost of $94,687, while the average IPF patient accrued $109,116.22. It is important to note that the standard deviations on these cost figures are extremely wide, being approximately two times the mean themselves. Interestingly, IPF patients did not vary significantly in the number of diagnoses.

Table 52 identifies the top 10 diagnoses among the NIS/HCUP data set and outlines the costs, length of stay, and number of diagnoses associated with each. The most common non-birth diagnosis in the full data set was pneumonia, with hearing loss registering a distant second. As can be seen on this table, average costs are significantly lower than the top IPF diagnoses, corroborating the fact that IPF and chronic lung diseases are dramatically more expensive to treat and require longer hospital stays than most other diagnoses.

As previously mentioned, birth-related diagnoses were far and away the most common within the NIS/HCUP data set. Vaginal birth itself was three times as common as the most common non-birth diagnosis, with a total of 3,149,069 occurrences. It is interesting, however, that birth by C-section is so commonplace, with nearly 1.5 occurrences. This finding serves to establish the pseudo "rate" for C-sections among hospital births who had not had a previous C-section at 32.2%. Upon further examination, we find that C-sections are more than twice as expensive, require nearly twice the length of stay, and contribute 0.6 more diagnoses per event than a vaginal birth. Births to mothers having had a prior C-section are 2.5 times as expensive as vaginal births and contributed nearly twice the number of diagnoses.

Table 51: Mean total charge, length of stay, and number of diagnoses among top 10 diagnoses within IPF patients

	Total Charge			Length of Stay			Number of Diagnoses		
Diagnosis	N	Mean	Std Dev	N	Mean	Std Dev	N	Mean	Std Dev
Hearing loss	497	$77,914.97	$116,053.42	518	8.42	7.92	518	17.24	5.64
Subendo infarction	215	$67,317.69	$79,241.73	222	6.65	5.86	222	15.26	5.64
Atrial fibrillation	239	$34,174.11	$54,150.28	242	4.47	3.70	242	12.90	5.25
CHF	399	$34,178.74	$45,249.04	402	5.63	4.46	402	12.74	5.05
Pneumonia	1498	$42,516.83	$83,076.35	1519	6.64	7.24	1519	12.80	5.38
Obstructie chronic bronchitis	717	$28,092.08	$31,729.50	723	5.36	4.31	723	12.50	5.22
Interstitial pulmonary fibrosis	3206	$94,687.53	$207,843.81	3255	8.84	12.91	3256	11.50	5.32
IPF	1339	$109,116.22	$249,185.33	1361	8.59	13.29	1361	13.43	5.95
Acute respiratory failure	729	$81,535.87	$117,594.88	741	9.53	9.90	741	13.65	5.37
Acute and chronic respiratory failure	672	$78,805.15	$158,746.00	693	9.80	14.02	693	14.62	5.70

Table 52: Total charge, Length of Stay, and Number of diagnoses (per person) by top diagnoses - NIS/HCUP data set

	Total Charges			Length of Stay			Number of diagnoses		
Diagnosis	N	Mean	Std Dev	N	Mean	Std Dev	N	Mean	Std Dev
Hearing loss	686253	$60,664.20	$93,727.16	707289	7.68	8.89	707306	14.75	5.73
Coronary artherosclerosis	750164	$60,935.95	$62,926.97	759008	3.70	4.51	759206	9.48	4.41
Atrial fibrillation	508806	$27,462.05	$37,499.77	515954	3.48	3.80	515993	9.38	4.68
Congestive Heart Failure	541290	$33,007.39	$57,330.95	548994	4.92	5.80	549044	11.07	4.61
Pneumonia	1099094	$25,980.76	$38,500.51	1116491	4.78	4.63	1116552	9.65	5.17
Obstructive chronic bronchitis	599742	$24,166.91	$33,543.51	606057	4.51	4.27	606089	10.04	4.61
Osteoarthritis	521503	$47,081.07	$26,865.74	530393	3.30	1.67	530395	7.12	3.80
Chest pain	533281	$17,942.60	$15,434.46	539806	1.89	1.87	539976	8.41	4.24
Vaginal birth	3149069	$6,064.32	$33,571.19	3261254	2.43	5.01	3261263	2.63	1.85
Birth - C-section	1496987	$14,190.59	$66,669.91	1543674	4.50	9.43	1543681	3.21	2.62
Vaginal birth after C-section	625114	$15,043.90	$9,772.75	636605	2.78	1.21	636607	4.77	2.51

Table 53 further expands the distribution of primary diagnoses by age category. We see from this table that 27 of the IPF codes were found in the 16 and under age group, 10 diagnoses in 16-25, 57 diagnoses in 26-35, 178 diagnoses in 36-45, 417 diagnoses in 46-55, 1,031 diagnoses in 56-65, and 2,034 diagnoses in patients over 65 years of age. The next most commonly diagnoses disease in IPF patients were pneumonia (1519), acute or chronic respiratory failure (1,434), Unspecified diagnosis (518), and congestive heart failure (402).

The following tables, labeled Tables 53 through 55 display the initial output of the GLM procedure in SAS with Length of Stay as the dependent variable and Race as the independent (predictor). As can be seen in Table 53, there were six "levels" of race. Table 54 gives the number of observations read into the model, and the observations used. The observations used in modeling is affected by a missing variable in any component of the model, hence the subtraction of observations used

Class Level Information		
Class	**Levels**	**Values**
RACE	6	1 2 3 4 5 6

Table 53: The GLM Procedure Class Level Information

Number of Observations Read	18790
Number of Observations Used	11787

Table 54: The GLM Procedure Number of Observations Read and Used

Table 55 displays the basic characteristics of the model. A statistical model typically has x-1 degrees of freedom. As there were six race categories, the model would have five degrees of freedom (6-1 = 5). Table 29 displays the R-square statistic for Race. A high R-square indicates that a predictor variable has strong predictive ability. In this case, the R-square for Race is 0.000885, which is almost 0. This indicates that, overall, race is a very poor predictor of Length of Stay. As Length of Stay is very strongly correlated with the Total cost of treatment and boarding (Total Charge), we can also infer that Race will be a poor predictor of Total Charge.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2252.527	450.505	2.09	0.0639
Error	11781	2543423.857	215.892		
Corrected Total	11786	2545676.384			

Table 55: The GML Procedure (LOS) Model Summary - Length of stay = Race

R-Square	Coeff Var	Root MSE	LOS_X Mean
0.000885	182.1590	14.69326	8.066175

Table 56: Predictive Power of Race for Length of Stay

Table 57 shows the sequential sumo of squares (Type I), and the partial sum of squares (Type III). The sum of squares is a measure of the total variability of a set of scores around a particular number, usually the mean of the set of scores, in this case, the Length of Stay for patients, based on their race category. A statistically significant score

p-value based on an F-statistic indicates that the dispersion of Length of Stay causes the Race variable not to a good fit for modeling. In this case, our p-value of 0.0639 indicates a near statistically significant lack of fit. We should strongly consider not using the variable in any subsequent statistical model. Table 58 simply outlines Alpha, Error Degrees of Freedom, and the critical value for significance using Student's T-test.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
RACE	5	2252.526987	450.505397	2.09	0.0639

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RACE	5	2252.526987	450.505397	2.09	0.0639

Table 57: The GLM Procedure Type I & III of SS

Alpha	0.05
Error Degrees of Freedom	11781
Error Mean Square	215.892
Critical Value of Studentized Range	4.03076

Table 58: The GLM Procedure by HSD test for LOS

Finally, Table 59 displays comparisons of Length of Stay among all race categories. As previously mentioned, race was assigned into six categories by the NIS data set as follows:

- White = 1
- Black = 2

- Hispanic = 3
- Asian/PI = 4
- Native Am = 5
- Other = 6

The table below indicates Length of stay differed in a significant way when comparing Whites vs Blacks only. There are two cells in the table below that are indicated as being significant at the 0.05 alpha level; however, upon further review, we discover that both of these are for the white vs black comparison. The first reads 2-1, while the other reads 1-2. The confidence intervals for these comparisons are simply switched.

Comparisons significant at the 0.05 level are indicated by *.**				
RACE Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
2 – 6	0.6070	-2.1365	3.3504	
2 – 4	0.9102	-2.1551	3.9755	
2 – 3	1.2119	-0.6019	3.0256	
2 – 1	1.3503	0.0102	2.6903	***
2 – 5	3.1462	-1.3997	7.6920	
6 – 2	-0.6070	-3.3504	2.1365	
6 – 4	0.3032	-3.4008	4.0072	
6 – 3	0.6049	-2.1543	3.3640	
6 – 1	0.7433	-1.7304	3.2170	
6 – 5	2.5392	-2.4596	7.5380	
4 – 2	-0.9102	-3.9755	2.1551	

Comparisons significant at the 0.05 level are indicated by ***.				
RACE Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
4 – 6	-0.3032	-4.0072	3.4008	
4 – 3	0.3017	-2.7777	3.3810	
4 – 1	0.4401	-2.3864	3.2665	
4 – 5	2.2360	-2.9465	7.4184	
3 – 2	-1.2119	-3.0256	0.6019	
3 – 6	-0.6049	-3.3640	2.1543	
3 – 4	-0.3017	-3.3810	2.7777	
3 – 1	0.1384	-1.2334	1.5103	
3 – 5	1.9343	-2.6210	6.4897	
1 – 2	-1.3503	-2.6903	-0.0102	***
1 – 6	-0.7433	-3.2170	1.7304	
1 – 4	-0.4401	-3.2665	2.3864	
1 – 3	-0.1384	-1.5103	1.2334	
1 – 5	1.7959	-2.5924	6.1842	
5 – 2	-3.1462	-7.6920	1.3997	
5 – 6	-2.5392	-7.5380	2.4596	
5 – 4	-2.2360	-7.4184	2.9465	
5 – 3	-1.9343	-6.4897	2.6210	
5 – 1	-1.7959	-6.1842	2.5924	

Table 59: The GLM Procedure (LOS) Comparisons significant at the 0.05 level.

Diagnoses by Race

Table 60: provides the top diagnoses among IPF patients by race. Most notable from this table is the vast majority of these hospital-related diagnoses occur within the white population. As all of these diagnoses are related to chronic conditions, it would appear that greater than 75% of all chronic disease diagnoses related to the human lung and IPF (with the exception of "chest pain", which can be interpreted in a variety of ways and attributed to an endless number of conditions) occur within the white population. The same type of distribution occurs within the payor and race cross tabulations.

The distribution of the top diagnoses within the NIS database (Table 61) is even more skewed. The top diseases/conditions are heavily white and almost exclusively chronic. This could be an indicator that the white population has one, two, or all of the following attributes. First, they may be more likely to seek care in the hospital setting. Second, and most likely, they do not care for themselves as well as the other race-based sub-populations in the database. Third, they may be most likely to have insurance, and therefore, most likely to actually seek care for their chronic conditions and medical emergencies. Another possibility would be that the NIS/HCUP database is not representative of the general population; however, due to its vastness, this scenario is highly unlikely, but worth stating.

Table 60: Primary diagnosis by race - IPF

	White		Black		Hispanic		Asian/PI		Native American		Other	
DX1(Diagnosis 1)	N	%	N	%	N	%	N	%	N	%	N	%
Idiopathic interstitial pneumonia	1998	74.39	227	8.45	309	11.5	53	1.97	20	0.74	79	2.94
Pneumonia	978	75.7	107	8.28	120	9.29	33	2.55	12	0.93	42	3.25
IPF	947	74.8	121	9.56	113	8.93	29	2.29	10	0.79	46	3.63
Obstructive chronic bronchitis with acute exacerbation	468	81.39	46	8	38	6.61	11	1.91	4	0.7	8	1.39
Acute respiratory failure	466	77.15	54	8.94	50	8.28	10	1.66	7	1.16	17	2.81
Acute and chronic respiratory failure	428	75.09	67	11.75	53	9.3	8	1.4	1	0.18	13	2.28
Hearing loss	348	76.15	35	7.66	43	9.41	13	2.84	6	1.31	12	2.63
Congestive Heart failure - unspecified	237	74.53	37	11.64	26	8.18	4	1.26	1	0.31	13	4.09
Atrial fibrillation	177	88.94	9	4.52	8	4.02	0	0	2	1.01	3	1.51
Subendo infarction - 1st	145	82.86	11	6.29	13	7.43	3	1.71	1	0.57	2	1.14

Table 61: Top primary diagnoses by race - NIS/HCUP data set

	White		Black		Hispanic		Asian/PI		Native American		Other	
Diagnosis	n	%	n	%	n	%	n	%	n	%	n	%
Pneumonia	687543	83.29	116124	12.29	87649	9.28	19191	2.03	8162	0.86	####	2.78
Coronary atherosclerosis	484368	76.45	59620	9.41	46222	7.3	15117	2.39	4729	0.75	####	3.71
Hearing loss	448925	71.63	85129	13.58	54838	8.75	17302	2.76	4107	0.66	####	2.62
Obstructive chronic bronchitis with acute exacerbation	420233	81.86	53028	10.33	21930	4.27	5009	0.98	3342	0.65	9845	1.92
Osteoarthritis	370008	83.82	32983	7.47	21400	4.85	5063	1.15	2375	0.54	9617	2.18
Atrial fibrillation	366335	83.29	31978	7.27	23129	5.26	6110	1.39	2495	0.57	9778	2.22
Congestive Heart failure - unspecified	306597	83.29	85521	18.8	38779	8.52	9174	2.02	2927	0.64	####	2.63
Chest pain	284015	60.63	98564	21.04	56266	12.01	9773	2.09	3251	0.69	####	3.54
Single liveborn birth - vaginal	1381883	51.46	357255	13.3	616199	22.95	142630	5.31	24766	0.92	####	6.05
Single liveborn birth - c-section	661481	51.21	191151	14.8	286062	22.15	65865	5.1	11063	0.86	####	5.89
Vaginal birth after C-section	277488	50.78	75135	13.75	137783	25.21	25559	4.68	4918	0.9	####	4.69

Post-proposal analysis

After the initial submission of the dissertation proposal, the committee requested analyses pursuant to the following questions.

1) A cost comparison with other hospitalizations (e.g. lung diseases, heart failure, etc.)

In order to formulate this comparison, we merged all years of the NIS data set and computed the mean total charge among other DRG's relevant to chronic conditions/disease. Table 62 (below) ranks each DRG based on total charge.

Remembering that the mean charge for IPF patients was $68,926, we can see that very few other chronic condition DRG's eclipse that amount (only five total).

Table 62: Cost comparison by Diagnosis-Related Group

DRG	Frequency	Mean TOTCHG	DRG Description
652	17104	$ 170,760	KIDNEY TRANSPLANT
870	59521	$ 168,462	SEPTICEMIA OR SEVERE SEPSIS W MV 96+ HOURS
853	91818	$ 145,691	INFECTIOUS & PARASITIC DISEASES W O.R. PROCEDURE W MCC
61	5860	$ 82,711	ACUTE ISCHEMIC STROKE W USE OF THROMBOLYTIC AGENT W MCC
665	1164	$ 72,491	PROSTATECTOMY W MCC
854	21371	$ 62,150	INFECTIOUS & PARASITIC DISEASES W O.R. PROCEDURE W CC
62	11008	$ 53,350	ACUTE ISCHEMIC STROKE W USE OF THROMBOLYTIC AGENT W CC
177	136201	$ 50,053	RESPIRATORY INFECTIONS & INFLAMMATIONS W MCC
871	577339	$ 48,729	SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC
855	1624	$ 45,215	INFECTIOUS & PARASITIC DISEASES W O.R.

				PROCEDURE W/O CC/MCC
283	24849	$	43,657	ACUTE MYOCARDIAL INFARCTION EXPIRED W MCC
199	10426	$	43,432	PNEUMOTHORAX W MCC
280	132876	$	42,991	ACUTE MYOCARDIAL INFARCTION DISCHARGED ALIVE W MCC
813	37320	$	42,330	COAGULATION DISORDERS
58	3214	$	41,718	MULTIPLE SCLEROSIS & CEREBELLAR ATAXIA W MCC
196	15203	$	41,388	INTERSTITIAL LUNG DISEASE W MCC
302	16051	$	41,093	ATHEROSCLEROSIS W MCC
682	194504	$	40,157	RENAL FAILURE W MCC
757	3416	$	39,935	INFECTIONS FEMALE REPRODUCTIVE SYSTEM W MCC
63	8441	$	39,783	ACUTE ISCHEMIC STROKE W USE OF THROMBOLYTIC AGENT W/O CC/MCC
666	2950	$	39,438	PROSTATECTOMY W CC
186	22588	$	37,859	PLEURAL EFFUSION W MCC
917	97351	$	37,304	POISONING & TOXIC EFFECTS OF DRUGS W MCC
180	73546	$	36,712	RESPIRATORY NEOPLASMS W MCC
291	357775	$	35,604	HEART FAILURE & SHOCK W MCC
193	244724	$	35,401	SIMPLE PNEUMONIA & PLEURISY W MCC
178	115335	$	35,003	RESPIRATORY INFECTIONS & INFLAMMATIONS W CC
304	15366	$	34,877	HYPERTENSION W MCC
197	18617	$	34,819	INTERSTITIAL LUNG DISEASE W CC
913	2578	$	34,811	TRAUMATIC INJURY W MCC
285	5336	$	34,368	ACUTE MYOCARDIAL INFARCTION EXPIRED W/O CC/MCC
175	48617	$	33,647	PULMONARY EMBOLISM W MCC
189	219159	$	33,001	PULMONARY EDEMA & RESPIRATORY FAILURE
637	68948	$	32,791	DIABETES W MCC
59	9834	$	29,134	MULTIPLE SCLEROSIS & CEREBELLAR ATAXIA W CC
281	94053	$	28,332	ACUTE MYOCARDIAL INFARCTION DISCHARGED ALIVE W CC
190	276806	$	28,219	CHRONIC OBSTRUCTIVE PULMONARY DISEASE W MCC
181	69174	$	27,979	RESPIRATORY NEOPLASMS W CC
187	22515	$	27,197	PLEURAL EFFUSION W CC
872	226105	$	26,598	SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W/O MCC
200	27035	$	25,507	PNEUMOTHORAX W CC
683	266314	$	24,943	RENAL FAILURE W CC

176	122261	$	24,874	PULMONARY EMBOLISM W/O MCC
179	50788	$	24,770	RESPIRATORY INFECTIONS & INFLAMMATIONS W/O CC/MCC
191	278712	$	23,810	CHRONIC OBSTRUCTIVE PULMONARY DISEASE W CC
758	9143	$	23,738	INFECTIONS FEMALE REPRODUCTIVE SYSTEM W CC
292	379747	$	23,570	HEART FAILURE & SHOCK W CC
188	33341	$	23,085	PLEURAL EFFUSION W/O CC/MCC
198	10625	$	22,593	INTERSTITIAL LUNG DISEASE W/O CC/MCC
60	17463	$	22,548	MULTIPLE SCLEROSIS & CEREBELLAR ATAXIA W/O CC/MCC
371	224132	$	22,258	MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W MCC
282	89935	$	22,235	ACUTE MYOCARDIAL INFARCTION DISCHARGED ALIVE W/O CC/MCC
194	473496	$	21,926	SIMPLE PNEUMONIA & PLEURISY W CC
667	3964	$	21,582	PROSTATECTOMY W/O CC/MCC
372	128471	$	21,000	MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W CC
886	23556	$	20,694	BEHAVIORAL & DEVELOPMENTAL DISORDERS
638	164020	$	20,231	DIABETES W CC
202	185694	$	20,007	BRONCHITIS & ASTHMA W CC/MCC
69	194781	$	19,984	TRANSIENT ISCHEMIA
885	1201581	$	19,283	PSYCHOSES
201	22567	$	19,072	PNEUMOTHORAX W/O CC/MCC
185	10903	$	18,918	MAJOR CHEST TRAUMA W/O CC/MCC
312	329982	$	18,731	SYNCOPE & COLLAPSE
284	6241	$	18,206	ACUTE MYOCARDIAL INFARCTION EXPIRED W CC
914	23127	$	17,521	TRAUMATIC INJURY W/O MCC
303	105793	$	17,456	ATHEROSCLEROSIS W/O MCC
182	98944	$	17,107	RESPIRATORY NEOPLASMS W/O CC/MCC
184	35641	$	16,934	MAJOR CHEST TRAUMA W CC
192	302544	$	16,782	CHRONIC OBSTRUCTIVE PULMONARY DISEASE W/O CC/MCC
305	110377	$	16,619	HYPERTENSION W/O MCC
293	214206	$	16,485	HEART FAILURE & SHOCK W/O CC/MCC
684	69556	$	16,447	RENAL FAILURE W/O CC/MCC
183	53997	$	16,354	MAJOR CHEST TRAUMA W MCC
759	16798	$	15,502	INFECTIONS FEMALE REPRODUCTIVE SYSTEM W/O CC/MCC
313	534904	$	15,458	CHEST PAIN
195	315142	$	14,575	SIMPLE PNEUMONIA & PLEURISY W/O CC/MCC
918	219488	$	14,347	POISONING & TOXIC EFFECTS OF DRUGS W/O

				MCC
311	44289	$	13,855	ANGINA PECTORIS
639	145753	$	13,696	DIABETES W/O CC/MCC
897	344374	$	12,707	ALCOHOL/DRUG ABUSE OR DEPENDENCE W/O REHABILITATION THERAPY W/O MCC
203	327223	$	12,357	BRONCHITIS & ASTHMA W/O CC/MCC
373	442602	$	9,318	MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W/O CC/MCC

2) The most frequent procedures performed on the IPF population during their treatment

Also requested was to indicate the most commonly performed procedures on IPF patients across the study period (2007-2012). Table 34 provides a list of all procedures whose counts were greater than or equal to 300.

Table 63: Most common procedures among IPF patients

ICD-9 CM Code	Count	Description
3893	2129	OTHER VENOUS CATH (NEC) (Begin 1980)
3324	1922	CLOSED BRONCHIAL BIOPSY
9904	1881	PACKED CELL TRANSFUSION
9604	1512	INSERT ENDOTRACHEAL TUBE
9390	1207	CONT POS AIRWAY PRESSURE (Begin 1988)
9672	1030	CONT MECH VENT 96+ HRS (Begin 1991)
8872	905	DX ULTRASOUNDHEART
9671	857	CONT MECH VENT < 96 HRS (Begin 1991)
8856	647	CORONAR ARTERIOGR2 CATH
3327	630	CLOS ENDOSCOPIC LUNG BX (Begin 1987)
3404	562	INSERT INTERCOSTAL CATH
3220	482	THORAC EXC LUNG LESION (Begin 2007)
3891	470	ARTERIAL CATHETERIZATION
8853	466	LT HEART ANGIOCARDIOGRAM
3995	458	HEMODIALYSIS
3322	431	FIBEROPTIC BRONCHOSCOPY
3323	389	OTHER BRONCHOSCOPY
8741	382	C.A.T. SCAN OF THORAX
3961	378	EXTRACORPOREAL CIRCULAT
3722	377	LEFT HEART CARDIAC CATH

9907	359	SERUM TRANSFUSION NEC
3897	354	CV CATH PLCMT W GUIDANCE (Begin 2010)
966	330	ENTRAL INFUS NUTRIT SUB (Begin 1986)
4516	313	EGD WITH CLOSED BIOPSY (Begin 1988)
0093	306	TRANSPLANT CADAVER DONOR (Begin 2004)
0093	306	OTHER LACRIMAL GLAND OPS
311	305	TEMPORARY TRACHEOSTOMY

3) Most frequent co-morbidities (e.g. diabetes, heart disease, etc.) and smoking y/n?

We examined are the frequencies of AHRQ comorbidities, as found in the severity data set. This required that the severity data set be linked by HCUP unique identifier to the severity data set. The Frequencies Procedure in SAS was used to compute.

Table 64 provides a list of all AHRQ comorbidities in descending order from the highest count. It should be of note that more than half of all IPF patients were hypertensive.

It was also asked to examine smoking status among IPF patients; however, no such variable exists to provide a reasonable interpretation as to whether or not one was a smoker. Therefore, an analysis of smoking and IPF was not completed.

Table 64: AHRQ comorbidities with IPF

AHRQ comorobidities with IPF	**Frequency**	**Percent**
Hypertension (combine uncomplicated and complicated)	10412	55.4%
Chronic pulmonary disease	7593	40.4%
Fluid and electrolyte disorders	5855	31.2%
Diabetes, uncomplicated	4700	25.0%
Congestive heart failure	4336	23.1%
Deficiency anemias	4227	22.5%
Pulmonary circulation disorders	3133	16.7%
Hypothyroidism	2989	15.9%
Renal failure	2818	15.0%
Depression	2285	12.2%
Rheumatoid arthritis/collagen valcular diseases	1689	9.0%
Valvular disease	1648	8.8%

Obesity	1628	8.7%
Weight loss	1541	8.2%
Peripheral vascular disorders	1356	7.2%
Coagulopathy	1234	6.6%
Other neurological disorders	1222	6.5%
Diabetes with chronic complications	810	4.3%
Psychoses	519	2.8%
Liver disease	464	2.5%
Solid tumor without metastasis	452	2.4%
Alcohol abuse	336	1.8%
Metastatic cancer	329	1.8%
Drug abuse	245	1.3%
Chronic blood loss anemia	239	1.3%
Lymphoma	233	1.2%
Paralysis	208	1.1%
AIDS	34	0.2%
Peptic ulcer disease excluding bleeding	6	0.0%

4) Link to a family history of IPF;

Diagnosis codes within these data sets were provided using ICD-9. An extensive search of the Core and Hospital merged files provided no variable or ICD-9 code pertaining to a family history of IPF.

Thus, forced to approximate using the ICD9 code V17.6, which signifies a family history of other diseases of the respiratory system. The ICD-9 code book indicates that approximate synonyms to this code are as follows:

- Family history of chronic obstructive lung disease
- Family history of chronic respiratory condition
- Family history of chronic respiratory disease
- Family history of pulmonary emphysema
- Family history: Bronchitis
- Family history: Bronchitis/COAD
- Family history: Hay fever
- Family history: Occupational lung disease
- Family history: Respiratory disease
- Fhx of chronic respiratory condition

- History of – hay fever.

Being that this is a rather inclusive variable, its best to seek to measure an association between it and being a patient with IPF. However, after reviewing all diagnosis codes (DX1-DX15), one is only able to find 17 total instances where this code is referenced. From this small number, it is impossible to determine any association with IPF.

Post Defense Analysis

Severity and Risk of Mortality among IPF patients

After the completion of the dissertation defense, and preparatory to performing the proposed trial, we were asked to examine disease severity and the risk of mortality among those with IPF. Severity is defined by NIS as "the likelihood of death or organ failure resulting from disease progression and independent of the treatment process. Disease progression is measured using four stages of increasing complexity as follows:

- Stage 1 – no complications or problems of minimal severity
- Stage 2 - problems limited to a single organ or system; significantly increased risk of complications
- Stage 3 – multiple site involvement; generalized systemic involvement; poor prognosis
- Stage 4 - death

Among all IPF patients average severity was 3.019 (SD = 0.75), which indicates that the average IPF patient's disease had progressed to the point of being systemic with a poor prognosis.

In Table 65, we also record the frequencies or risk of mortality among IPF patient. Risk of mortality was assigned the following values within the analysis data set:

0 = no class specified

1= minor likelihood of dying

2 = moderate likelihood of dying

3 = major likelihood of dying

4 = extreme likelihood of dying

Average risk of mortality in our data was found to be 2.79 (SD = 0.78), which indicates that patients were edging upon being at major risk of death.

Table 65: Reported severity and risk of mortality among IPF patients.

DRG Severity	Frequency	Percent
0	8	0.04
1	361	1.92
2	4027	21.43
3	9262	49.29
4	5132	27.31
Risk of Mortality		
0	8	0.04
1	694	3.69
2	5859	31.18
3	8823	46.96
4	3406	18.13

To further the examination of disease severity we performed descriptive statistics on total charge, length of stay, and number of diagnoses among individuals of varying levels of DRG severity. Table 66 displays the results of this comparison. Beginning with the total charge startum, we notice a difference between average costs in Level 1 of the

severity scale. Those whose severity is zero are, for all intents and purposes not statistically significantly different. However, we find that IPF Severity level 1 mean costs are more than $40,000 more per patient than the typical hospital admission in the NIS. This most likely results from the cost to observe and diagnose a person with IPF, as it is not typically of a known original or etiology. An individual may spend a few days undergoing tests for non-threatening IPF, which would certainly escalate the hospital charges for that individual.

In the length of stay stratum, the only real difference stems from the number of days that individuals are in the hospital for zero severity incidences. There were only seven cases of IPF that were determined to be zero severity, which sample is not reliable to make an estimation for the entire IPF group. Our sense is that, of the 35,619 low severity cases in the length of stay stratum, most of these are likely C-section births. We say this because the average length of stay among C-sections is 9.43 days. Having a C-section would certainly constitute a medical event and would result in a potentially extended stay to prevent or mitigate infection.

We immediately notice that there are many more diagnoses per person within the IPF-only group as compared to the NIS/HCUP data set, regardless of the level of severity. Again, this is most likely due to the difficulty in diagnosing IPF and the presence of multiple chronic conditions among IPF patients. Typical patients entering the hospital most likely enter with a particular problem, one that has a relatively clear label and that can be coded rather easily into a field that fits nicely into the current electronic health record system. IPF, as we have discovered, does not behave like a "typical" condition, and therefore radically diverges from that pattern.

Table 66: Total charges and length of stay by severity

Total Charges						
	IPF			**NIS**		
Severity	N	Mean	Std Dev	N	Mean	Std Dev
0	7	$22,091.57	$17,287.88	38744	$29,394.58	$86,230.54
1	357	$59,655.86	$91,171.47	13628918	$17,577.06	$23,495.30
2	3991	$36,364.49	$61,994.48	13951970	$26,660.12	$32,250.46
3	9078	$42,393.70	$60,513.43	8232179	$42,470.11	$55,498.06
4	5031	$134,624.53	$235,925.60	2387622	$114,615.11	$164,438.55
Length of stay						
Severity	N	Mean	Std Dev	**N**	**Mean**	**Std Dev**
0	7	4.29	1.11	35619	7.34	24.81
1	361	3.98	3.69	13894784	2.59	2.97
2	4027	4.41	4.00	14199906	3.98	4.79
3	9262	6.19	5.40	8399786	6.33	7.07
4	5131	13.05	15.29	2436381	13.26	16.28
Number of diagnoses						
Severity	**N**	**Mean**	**Std Dev**	**N**	**Mean**	**Std Dev**
0	8	8.25	1.75	48192	0	0
1	327	7.26	4.04	16945336	1.05	0.24
2	3406	9.74	4.37	17162766	1.41	0.57
3	7744	13.55	4.88	10077332	2.35	0.77
4	4242	17.10	5.70	2899931	3.57	0.61
Risk of mortality						
Severity	**N**	**Mean**	**Std Dev**	**N**	**Mean**	**Std Dev**
0	8	0.00	0.00	0	0.00	0.00
1	327	1.48	0.60	16945336	1.05	0.24
2	3406	2.04	0.52	17162766	1.41	0.57
3	7744	2.76	0.55	10077332	2.35	0.77
4	4242	3.55	0.54	2899931	3.57	0.61

Impact of Age

Finally, we were asked to examine the impact of age on race, health care coverage, and the top three diagnoses. In order to make comparisons between the general populace of the NIS/HCUP data set and only those with a diagnosis of IPF, we will (over

the balance of the document, provide tables that indicate rates and correlations among the main study body and the sub-population of interest. This comparison is made in Table 67.

Table 67: Mean age - all NIS/HCUP and IPF only

Group	N	Mean	Std Dev
All NIS/HCUP	47,084,607	48.36	27.80
IPF Only	18,788	71.18	13.98

It should be interest that the average age for IPF is vastly different than the age of others within the analysis data set. The variability in the age distribution can be explained through at least two key points. First, those who develop IPF are of (at least) advanc(ing) age, which is a plausible argument, as IPF is a degenerative condition that can accompany other age-related diagnoses. Second the full NIS data set contains data points for individuals of all ages, including those being born and those who die due to acute or non-age-related conditions. Two of the most dominant diagnostic codes in the data indicate the birth of an individual, for which age is nearly always less than 1. Therefore, it should be of no surprise that the average age of patients is lower.

We also examined the age distribution among races in the same data sets. IPF patients were much more tightly grouped as it pertains to age distribution than the general population. Whites were the oldest in both groups. Hispanics were the youngest in NIS/HCUP, while blacks occupied that distinction in the IPF group. The IPF group was more narrowly distributed (had a smaller standard deviation) than the rest of the analysis data set.

Tables 69 and 70 indicate the correlations between age and races among both populations. All Pearson correlation coeffients greater than 0.15 are identified in both

tables. The entire correlation matrix is included for completeness sake. Among the NIS/HCUP population, age was positively associated with being white and negatively associated with being Hispanic, meaning that whites were relatively older and Hispanics were relatively younger in the NIS/HCUP data, but not in a staggering way. The same pattern was noticed in the IPF population.

Table 71 illustrates the distribution of age by payor/insurance. As with previous comparisons, the distribution of age is much broader among the NIS/HCUP population. Of particular interest in this comparison is the lower age of individuals in the Medicaid, self-pay, and no coverage groups compared with the Medicare covered sub-population.

Table 72 summarizes the distribution of payor by disease category in preparation for running correlations on the same data. As mentioned before, one should note that more than 72% of the IPF sub-group is on Medicare, while only 37.8% of the general populace has Medicare coverage. Twenty percent of the general NIS/HCUP are on Medicaid, while only 5.6% of all IPF patients are Medicaid-covered. The 5.6% in the IPF group likely represents those disabled persons who are too young to qualify for Medicaid, while the 20% on Medicaid in the NIS/HCUP are likely children born to lower income mothers who also qualify for the entitlement program.

Table 68: Average age by race

Race	All NIS/HCUP			IPF Only		
	N	Mean	Std Dev	N	Mean	Std Dev
White	26,357,623	53.25	26.88	12080	72.87	12.88
Black	5,814,937	43.87	25.43	1437	61.74	15.31
Hispanic	4,973,566	34.82	27.18	1354	66.26	16.33
Asian	1,077,297	40.15	29.28	295	69.09	16.02
Native American	293,693	42.06	27.26	123	67.74	16.31
Other	1,414,674	39.21	28.55	414	68.38	15.21

Table 69: Pearson correlations age and race (NIS/HCUP all)

	AGE	White	Black	Hispanic	Asian/PI	Native Am	Other
AGE		0.19844	-0.06066	-0.16743	-0.04517	-0.01795	-0.0579
White	0.19844		-0.42277	-0.38711	-0.17237	-0.08923	-0.19824
Black	-0.06066	-0.42277		-0.12892	-0.0574	-0.02972	-0.06602
Hispanic	-0.16743	-0.38711	-0.12892		-0.05256	-0.02721	-0.06045
Asian or Pacific Islander	-0.04517	-0.17237	-0.0574	-0.05256		-0.01212	-0.02692
Native American	-0.01795	-0.08923	-0.02972	-0.02721	-0.01212		-0.01393
Other	-0.0579	-0.19824	-0.06602	-0.06045	-0.02692	-0.01393	

Table 70: Age and race correlations among IPF patients							
	AGE	White	Black	Hispanic	Asian/PI	Native Am	Other
AGE		0.16259	-0.19444	-0.09807	-0.01883	-0.01997	-0.03009
White	0.16259		-0.38611	-0.37405	-0.16946	-0.10891	-0.20139
Black	-0.19444	-0.38611		-0.08022	-0.03634	-0.02336	-0.04319
Hispanic	-0.09807	-0.37405	-0.08022		-0.03521	-0.02263	-0.04184
Asian/PI	-0.01883	-0.16946	-0.03634	-0.03521		-0.01025	-0.01896
Native Am	-0.01997	-0.10891	-0.02336	-0.02263	-0.01025		-0.01218
Other	-0.03009	-0.20139	-0.04319	-0.04184	-0.01896	-0.01218	

Table 71: Average age by Payor/Insurance

	NIS/HCUP all			IPF		
Payor	N	Mean	Std Dev	N	Mean	Std Dev
Medicare	17,797,265	72.99	13.69	13,612	75.81	10.63
Medicaid	9,416,092	24.86	21.42	1,061	53.11	15.50
Private Ins	15,496,100	37.03	23.47	3,408	60.75	13.26
Self-pay	2,446,623	37.18	18.92	296	55.33	14.36
No charge	232,783	41.25	17.06	27	54.67	13.48
Other	1,588,932	40.69	22.17	359	62.23	15.51

Table 72: Insurance/Payor by data grouping

Payor/Insurance	NIS/HCUP	IPF
Medicare	37.88	72.55
Medicaid	20.04	5.65
Private Insurance	32.99	18.16
Self-pay	5.21	1.58
No charge	0.5	0.14
Other	3.38	1.91

Table 73: Pearson Correlations for Age and Payor (NIS/HCUP)

	AGE	Medicare	Medicaid	Private Insurance	Self-pay	No charge	Other
AGE		0.69	-0.42	-0.29	-0.09	-0.02	-0.05
Medicare	0.69		-0.39	-0.55	-0.18	-0.05	-0.15
Medicaid	-0.42	-0.39		-0.35	-0.12	-0.04	-0.09
Private Insurance	-0.29	-0.55	-0.35		-0.16	-0.05	-0.13
Self-pay	-0.09	-0.18	-0.12	-0.16		-0.02	-0.04
No charge	-0.02	-0.05	-0.04	-0.05	-0.02		-0.01
Other	-0.05	-0.15	-0.09	-0.13	-0.04	-0.01	

Table 74: Pearson Correlations for Age and Payor (IPF)

	AGE	Medicare	Medicaid	Private Insurance	Self-pay	No charge	Other
AGE		0.54	-0.32	-0.35	-0.14	-0.04	-0.09
Medicare	0.54		-0.40	-0.76	-0.21	-0.06	-0.23
Medicaid	-0.32	-0.40		-0.12	-0.03	-0.01	-0.03
Private Insurance	-0.35	-0.76	-0.12		-0.06	-0.02	-0.07
Self-pay	-0.14	-0.21	-0.03	-0.06		0.00	-0.02
No charge	-0.04	-0.06	-0.01	-0.02	0.00		-0.01
Other	-0.09	-0.23	-0.03	-0.07	-0.02	-0.01	

Tables 73 and 74 illustrate the association between age and payor through the use of Pearson correlation coefficients. Cells with correlations of at least 0.30 in either direction were highlighted in yellow. Those with correlations greater than 0.5 were highlighted green, while those with negative correlations greater than 0.5 were

highlighted red. As can be seen, the two tables are nearly identical, with the only age-related difference being that the age is slightly more negatively associated with private insurance in the IPF data set. However, this difference can easily be explained using the age differences between data sets (discussed above).

Table 75 further illustrates the ages of the top diagnosis within the NIS/HCUP, and clearly demonstrates the aforementioned dominance of birth-related conditions in controlling the age distribution in in the general records. The average recorded age of live births is approximately 0.15 days. On the other hand, virtually all other patients with the most commonly occurring diagnoses in the NIS/HCUP are near unto or above 60 years of age. In our opinion, this is irrefutable evidence of the influence of the presence of birth records on our analysis.

Table 76 lists the most common diagnoses among IPF patients. Note that IPF itself is comprises only 14.06% of those diagnoses. We indicated in the Literature Review section that IPF is more difficult to diagnose due to its idiopathic nature. This is evidenced by the presence of the top diagnosis in this list, being Idiopathic interstitial pneumonia. As we also mentioned before, we considered those having this diagnosis as IPF patients. Therefore, we can conclude that approximately 48% of those in the IPF group had IPF as a primary diagnosis.

Tables 77 and 78 (found below) illustrate the correlations between age and the top diagnoses in both data sets. All Pearson coefficients above 0.20 are highlighted in yellow in both tables. As can be seen in Table 77, age was moderately positively correlated with all of the top diagnoses in the NIS/HCUP data set except for general chest pain. It is our suspicion, however, that were the source of the chest pain to be more refined, that source would likely be correlated

with age (at least moderately). As with the other correlation measures, the entire matrix is included for completeness sake.

Among the IPF group (Table 78), we see no such correlation. However, this is due to the more rigid and tighter distribution of age within the IPF sub-population. We have little doubt that the diagnoses indicated in Table 48 are age-related; however, for the purposes of this table, that correlation is attenuated. We do notice, however, that IPF and Idiopathic interstitial pneumonia are moderately correlated, and that the two are intertwined with pneumonia. Intuitively, acute respiratory failure and obstructive chronic bronchitis are moderately correlated as well.

As a matter of finality, we invoked the CORR procedures on the remaining variables that we used to analyze the NIS and IPF data sets (Tables 79 and 80). All correlations between 0.20 and 0.40, as well as those between -0.20 and -.40 are highlighted in yellow with dark yellow numbering. Correlations greater than 0.40 but less than 1 are highlighted in green with dark green numbering. Those correlations greater than -0.40 in magnitude, but less than -1 are highlighted in light red with dark red numbering. Finally, all correlations of 1 are shown as a black box. As before, we include the entire correlation for completeness sake.

We can see that severity and risk of mortality are highly correlated in both data sets (0.72 in the IPF and 0.71 in the NIS). Number of diagnoses is highly correlated with severity and risk of mortality in the NIS, but not as much in the IPF data set. The only largely negative correlations come from the racial categories. It is intuitive that practitioners do not general label a person as being white, black, and Hispanic at the same time. One usually identifies with one or two of these racial groups, but not all. Therefore, it follows that, if a person identifies with one, they will not identify (generally) with the other(s). Using this logic, we can easily understand why being “white” is negatively correlated with strictly being “black” and that being Hispanic is negatively correlated with being “white” or “black”.

Table 75: Mean ages among primary diagnoses

Diagnosis Category	Primary Diagnosis	N	Mean (years)	Std Dev
Birth-related diagnoses	Single liveborn birth	3,258,304	0.000044	0.02
	Single liveborn birth	1,541,608	0.000057	0.02
	Previous Cesarean Delivery, Delivered, With Or Without Mention Of Antepartum Condition	636,208	29.71	5.63
Non-birth conditions	Pneumonia	1,115,785	61.37	26.35
	Coronary artherosclerosis	758,970	65.45	11.86
	Diseases of the nervous system and sense organs	707,240	67.97	18.36
	Obstructive chronic bronchitis with acute exacerbation	606,026	68.90	12.03
	Congestive heart failure	548,994	72.61	14.92
	Chest pain	539,627	58.99	14.98
	Osteoarthritis	530,117	66.25	10.27
	Atrial fibrillation	515,854	70.26	14.09

Table 76: Age distribution among the top 10 primary diagnoses (IPF)

Diagnosis	Frequency	Percent	Mean	Std Dev
Idiopathic interstitial pneumonia	3256	33.65	68.51	13.80
Pneumonia	1519	15.7	72.47	14.20
IPF	1361	14.06	68.80	13.56
Acute Respiratory Failure	741	7.66	69.56	13.10
Obstructive Chronic Bronchitis	723	7.47	72.61	11.20
Acute and Chronic Respiratory Failure	693	7.16	69.64	12.47
Hearing loss	518	5.35	72.06	12.82
Congestive Heart Failure	402	4.15	76.67	11.49
Atrial fibrillation	242	2.5	77.35	10.47
Subendo Infarction	222	2.29	77.41	9.75

Table 77: Pearson coefficients for age and top diagnoses (NIS)									
	AGE	**Pneumonia**	**Coronary Artherosclerosis**	**CNS**	**Obstructive chronic bronchitis**	**Congestive Heart Failure**	**Chest pain**	**Osteoarthritis**	**Atrial fibrillation**
AGE		0.27	0.25	0.26	0.25	0.26	0.16	0.21	0.23
Pneumonia	0.27		-0.09	-0.09	-0.08	-0.08	-0.08	-0.08	-0.08
Artherosclerosis	0.25	-0.09		-0.07	-0.07	-0.06	-0.06	-0.06	-0.06
CNS	0.26	-0.09	-0.07		-0.06	-0.06	-0.06	-0.06	-0.06
Chronic bronchitis	0.25	-0.08	-0.07	-0.06		-0.06	-0.06	-0.06	-0.05
CHF	0.26	-0.08	-0.06	-0.06	-0.06		-0.05	-0.05	-0.05
Chest pain	0.16	-0.08	-0.06	-0.06	-0.06	-0.05		-0.05	-0.05
Osteoarthritis	0.21	-0.08	-0.06	-0.06	-0.06	-0.05	-0.05		-0.05
Atrial fibrillation	0.23	-0.08	-0.06	-0.06	-0.05	-0.05	-0.05	-0.05	

Table 78: Pearson coefficients for age and top 10 primary diagnoses (IPF)											
	AGE	**Idiopathic interstitial pneumonia**	**IPF**	**Pneumonia**	**Hearing loss**	**Congestive Heart Failure**	**Atrial fibrillation**	**Acute Respiratory Failure**	**Obstructive Chronic Bronchitis**	**Acute and Chronic Respiratory Failure**	**Subendo Infarction**
AGE		-0.11	-0.05	0.06	0.03	0.09	0.08	-0.02	0.04	-0.02	0.08
Idiopathic interstitial pneumonia	-0.11		-0.29	-0.31	-0.17	-0.15	-0.11	-0.21	-0.20	-0.20	-0.11
IPF	-0.05	-0.29		-0.17	-0.10	-0.08	-0.06	-0.12	-0.11	-0.11	-0.06
Pneumonia	0.06	-0.31	-0.17		-0.10	-0.09	-0.07	-0.12	-0.12	-0.12	-0.07
Hearing loss	0.03	-0.17	-0.10	-0.10		-0.05	-0.04	-0.07	-0.07	-0.07	-0.04
Congestive Heart Failure	0.09	-0.15	-0.08	-0.09	-0.05		-0.03	-0.06	-0.06	-0.06	-0.03
Atrial fibrillation	0.08	-0.11	-0.06	-0.07	-0.04	-0.03		-0.05	-0.05	-0.04	-0.02
Acute Respiratory Failure	-0.02	-0.21	-0.12	-0.12	-0.07	-0.06	-0.05		-0.08	-0.08	-0.04
Obstructive Chronic Bronchitis	0.04	-0.20	-0.11	-0.12	-0.07	-0.06	-0.05	-0.08		-0.08	-0.04
Acute and Chronic Respiratory Failure	-0.02	-0.20	-0.11	-0.12	-0.07	-0.06	-0.04	-0.08	-0.08		-0.04
Subendo Infarction	0.08	-0.11	-0.06	-0.07	-0.04	-0.03	-0.02	-0.04	-0.04	-0.04	

Table 79: Remaining correlations for variables analyzed within the IPF data set												
	LOS	**TOTCHG**	**NDX**	**severity**	**Risk of Mortality**	**FEMALE**	**White**	**Black**	**Hispanic**	**Asian/PI**	**Native Am**	**Other**
LOS		0.72	0.26	0.33	0.29	-0.03	0.01	0.02	0.01	0.01	-0.01	0.01
TOTCHG	0.72		0.23	0.25	0.22	-0.05	0.02	0.00	0.04	0.03	0.00	0.02
NDX	0.26	0.23		0.48	0.44	-0.02	0.13	0.00	-0.03	-0.01	0.00	-0.01
severity	0.33	0.25	0.48		0.72	-0.05	0.00	0.02	-0.01	0.02	0.00	0.01
risk_mortality	0.29	0.22	0.44	0.72		-0.05	0.03	-0.02	-0.03	0.01	-0.01	0.00
FEMALE	-0.03	-0.05	-0.02	-0.05	-0.05		-0.05	0.07	0.03	-0.01	0.02	0.00
White	0.01	0.02	0.13	0.00	0.03	-0.05		-0.39	-0.37	-0.17	-0.11	-0.20
Black	0.02	0.00	0.00	0.02	-0.02	0.07	-0.39		-0.08	-0.04	-0.02	-0.04
Hispanic	0.01	0.04	-0.03	-0.01	-0.03	0.03	-0.37	-0.08		-0.04	-0.02	-0.04
Asian/PI	0.01	0.03	-0.01	0.02	0.01	-0.01	-0.17	-0.04	-0.04		-0.01	-0.02
Native Am	-0.01	0.00	0.00	0.00	-0.01	0.02	-0.11	-0.02	-0.02	-0.01		-0.01
Other	0.01	0.02	-0.01	0.01	0.00	0.00	-0.20	-0.04	-0.04	-0.02	-0.01	

Table 80: Pearson Coefficients for severity, risk of mortality, length of stay, total charge, sex, and race variables in the NIS data set													
	Severity	**Risk of Mortality**	**LOS**	**TOTCHG**	**DIED**	**NDX**	**FEMALE**	**White**	**Black**	**Hispanic**	**Asian/PI**	**Native Am**	**Other**
Severity		0.76	0.35	0.33	0.23	0.68	-0.08	0.07	0.03	-0.08	-0.03	-0.01	-0.03
Risk of Mortality	0.76		0.31	0.32	0.29	0.66	-0.10	0.09	0.00	-0.08	-0.02	-0.01	-0.03
LOS	0.35	0.31		0.66	0.08	0.32	-0.04	0.00	0.03	-0.02	0.00	0.00	0.00
TOTCHG	0.33	0.32	0.66		0.12	0.32	-0.07	0.03	0.00	0.00	0.01	0.00	0.00
DIED	0.23	0.29	0.08	0.12		0.15	-0.02	0.02	0.00	-0.02	0.00	0.00	0.00
NDX	0.68	0.66	0.32	0.32	0.15		-0.06	0.14	0.01	-0.11	-0.04	-0.01	-0.04
FEMALE	-0.08	-0.10	-0.04	-0.07	-0.02	-0.06		-0.02	0.01	0.02	0.01	0.00	0.00
White	0.07	0.09	0.00	0.03	0.02	0.14	-0.02		-0.42	-0.39	-0.17	-0.09	-0.20
Black	0.03	0.00	0.03	0.00	0.00	0.01	0.01	-0.42		-0.13	-0.06	-0.03	-0.07
Hispanic	-0.08	-0.08	-0.02	0.00	-0.02	-0.11	0.02	-0.39	-0.13		-0.05	-0.03	-0.06
Asian/PI	-0.03	-0.02	0.00	0.01	0.00	-0.04	0.01	-0.17	-0.06	-0.05		-0.01	-0.03
Native Am	-0.01	-0.01	0.00	0.00	0.00	-0.01	0.00	-0.09	-0.03	-0.03	-0.01		-0.01
Other	-0.03	-0.03	0.00	0.00	0.00	-0.04	0.00	-0.20	-0.07	-0.06	-0.03	-0.01	

References

Adult Lung Transplantation Statistics. https://www.ishlt.org/registries/slides.asp?slides=heartLungRegistry. (Accessed March 15, 2015).

Antoniou K.M., Hansell D.M., Rubens M.B., Marten K., Desai S.R., Siafakas N.M., Nicholson G.A., du Bois R.M., Wells A.U. (2008). Idiopathic Pulmonary Fibrosis: Outcome in Relation to Smoking Status. *Am J Respir Crit Care Med 177,* 190–194.

Agusti, A.G., Roca, J., Gea, J., Wagner, P.D., Xaubet, A., Rodriguez-Roisin, R. (1991). Mechanisms of gas-exchange impairment in idiopathic pulmonary fibrosis. *Am Rev Respir Dis, 143(2*), 219-225.

Azuma, A., Nukiwa, T., Tsuboi, E., Suga, M., Abe, S., Nakata, K., Taguchi, Y., Nagai, S., Itoh, H., Ohi, M., Sato, A., Kudoh, S. (2005). Double-blind, placebocontrolled trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med 2005, 171(9)*, 1040-1047.

Baumgartner, K.B., Samet, J.M., Stidley, C.A., Colby, T.V., Waldron, J.A. (1997). Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med, 155(1),* 242-248.

Carrington, C.B, Gaensler, E. A. and Coutu, R. E. (1978) Natural history and treated course of usual and desquamative interstitial pneumonia," *New England Journal of Medicine, 298 (15),* 801–809.

Coalition For Pulmonary Fibrosis. Facts About Idiopathic Pulmonary Fibrosis. http://www.coalitionforpf.org/facts-about-idiopathic-pulmonary-fibrosis/ (accessed April 1, 2015).

Collard H.R. Tino G., Noble et al. Patient experiences with pulmonary fibrosis. Respir Med 2007;101:1350–1354.

Cottin, V., Cordier, J. (2012). Velcro crackles: the key for early diagnosis of idiopathic pulmonary fibrosis? *Eur Respir J, 40,* 519–521

Coultas, D.B., Zumwalt, R.E., Black, W.C., Sobonya, R.E. (1994). The epidemiology of interstitial lung diseases. *Am J Respir Crit Care Med, 150(4),* 967-972.

Demedts, M., Behr, J., Buhl, R., Costabel, U., Dekhuijzen, R., Jansen, H.M., . . . M, Montanari, M. (2005) High-dose acetylcysteine in idiopathic pulmonary fibrosis. *N Engl J Med, 353(21).* 2229-2242.

Cecilia Garcı´a-Sancho, M., Guillermo Carrillo, F., Pe´rez-Padilla, R., Ferna´ndez-Plata, M.R., Buendı´a-Rolda´n, I., Vargas, M.H., Selman, M. (2010). Risk factors for idiopathic

pulmonary fibrosis in a Mexican population. A case-control study. *Respiratory Medicine, 104(2),* 305-309.

Cox, DR (1958). "The regression analysis of binary sequences (with discussion)". *J Roy Stat Soc B* **20**: 215–242.

Erasmus D.B., Keller C.A., Alvarez F.B. (2008). Large airway complications in 150 consecutive lung transplant recipients. *Journal of Bronchology*, 15 (3), 152–157.

European Medicines Agency; Committee for Orphan Medical Products. Orphan drugs and rare diseases at a glance. (July 3, 2007). Doc Ref. EMEA/290072/2007 http://www.ema.europa.eu/docs/en_GB/document_library/Other/2010/01/WC500069805.pdf (Accessed April 1, 2015).

Fahim, A., Crooks M., Hart, S. (2011). Gastroesophageal reflux and Idiopathic Pulmonary Firosis: A review . Pulmonary Medicine. Aricle ID 634613, 7 pages. doi: 10.11554/2011/634613.

Fernandez Perez, E.R., Daniels, C.E., Schroeder, D.R., St Sauver, J., Hartman T.E., Brtholmai B.J., . . . Ryu, J.H. (2011). Incidence, prevalence, and clinical course of idiopathic pulmonary fibrosis: a population-based study. *Chest. 137(1),* 129-37.

Green, David M.; Swets, John A. (1966). Signal detection theory and psychophysics. New York, NY: John Wiley and Sons Inc.

Fioret, B.A., Mannino, M.D., Roman, J. In-Hospital Mortality and Costs Related to Idiopathic Pulmonary Fibrosis Between 1993 to 2008.

Frankel, S.K., Schwarz, M.I. (2009). Update in idiopathic pulmonary fibrosis. *Curr Opin Pulm Med, 15(5),* 463-9.

Genetics Home Reference – Idiopathic pulmonary fibrosis. http://ghr.nlm.nih.gov/condition/idiopathic-pulmonary-fibrosis. (Accessed April 14, 2015).

Gribbin, J., Hubbard, R.B., Le Jeune, I., Smith, C.J.P., West, J., Tata, L.J. (2006). Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK. *Thorax 2006, 61,* 980-985.

Hansell, A., Hollowell, J., Nichols, T., McNiece, R., Strachan, D. (1999). Use of the General Practice Research Database (GPRD) for respiratory epidemiology: a comparison with the 4th Morbidity Survey in General Practice (MSGP4). *Thorax, 54(5),* 413-419.

Harrell, F.E. (2001). Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer.

Hosmer, D.W., Lemeshow, S. (1999). Applied survival analysis: regression modeling of time to event data. New York: Wiley.

Johnston, I.D.A., Prescott, R. J., Chalmers, J. C., Rudd, R.M., British Thoracic Society study of cryptogenic fibrosing alveolitis: current presentation and initial management. (1997). *Thorax, 52 (1),* 38–44, 1997.

Kim, D.S., Collard, H.R., King, T.E. Jr. (2006). Classification and natural history of the idiopathic interstitial pneumonias. *Proc Am Thorac Soc, 3(4),* 285-292.

King, T.E., Bradford, W.Z., Castro-Bernardini, S., Fagan, E.A., Glaspole, I., Glassberg, M.K. . . . Noble, P.W. (2014), ASCEND Study Group: A Phase 3 Trial of Pirfenidone in Patients with Idiopathic Pulmonary Fibrosis. *N Engl J Med 370(22),* 2083–2092.

King, T.E., Costabel, U., Cordier, J.F., (2000). Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS), *American Journal of Respiratory and Critical Care Medicine, 161(2),* 646–664.

Meltzer, E.B. and Noble, P.W. (2008). Idiopathic pulmonary fibrosis. *Orphanet J Rare Dis, 3: 8* doi:10.1186/1750-1172-3-8.

Meyers, B.F., Lynch, J.P., Trulock, E.P., Guthrie, T., Cooper, J.D., Patterson, G.A. (2000). Single versus bilateral lung transplantation for idiopathic pulmonary fibrosis: a ten-year institutional experience. *J Thorac Cardiovasc Surg 120(1),* 99–107.

Navaratnam, V., Fleming, K.M., West, J., Smith, C.J.P., Jenkins, R.G., Fogarty, A., Hubbard, R.B. (2011). The rising incidence of idiopathic pulmonary fibrosis in the UK. *Thorax, 66,* 462-467.

Nemery, B., Bast, A., Behr, .J, Borm, P.J.A., Bourke, S.J., Camus, P.H., . . . D., Saltini, C. (2001). Interstitial lung disease induced by exogenous agents: factors governing susceptibility. *Eur Respir J, 18,* 30s–42s

Neurohr, C., Huppmann, P., Thum, D., Leuschner, W., von Wulffen, W., Meis, T., Leuchte, H., Behr, J. (2010) Munich Lung Transplant Group: Potential functional and survival benefit of double over single lung transplantation for selected patients with idiopathic pulmonary fibrosis. *Transpl Int, 23(9),* 887–896

Oh, C.K., Murray, L.A., Molfino, N.A. (2012). Smoking and Idiopathic Pulmonary Fibrosis. *Pulmonary Medicine Volume 2012,* Article ID 808260, 13 pages

Olson, A.L., Swigris, J.J., Lezotte, D.C., Norris, J.M., Wilson, C.G., Brown, K.K. (2007) Mortality from pulmonary fibrosis increased in the United States from 1992 to 2003. *Am J Respir Crit Care Med, 176(3),* 277-84.

Organ Procurement and Transplantation Network (OPTN) and Scientific Registry of Transplant Recipients (SRTR): OPTN/SRTR 2011 Annual Data Report. Rockville, MD: Department of Health and Human Services, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation; 2012.

Raghu, G., Brown, K.K., Bradford, W.Z., Starko, K., Noble, P.W., Schwartz, D.A., King, T.E. (2004). A placebo-controlled trial of interferon gamma-1b in patients with idiopathic pulmonary fibrosis. *N Engl J Med, 350(2),* 125-133.

Raghu, G., Weycker, D., Edelsberg, J., Bradford, W.Z., Oster, G. (2006). Incidence and prevalence of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med, 174(7),* 810-6.

Raghu, G,, Collard, H.R., Egan, J.J., Martinez, F.J., Behr, J., Brown, K.K., Colby, T.V., Cordier, J-F., Flaherty, K.R., Lasky, J.A., et al. (2011). An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med, 183,* 788–824.

Ryu, J. H. Colby, T. V., Hartman, T. E., Vassallo, R. (2001). Smoking- related interstitial lung diseases: a concise review. *European Respiratory Journal,* 17 (1), 122–132.

Richeldi, L., du Bois, R.M., Raghu, G., Azuma, A., Brown, K.K., Costabel, U. . . . . Collard, H.R. (2014). INPULSIS Trial Investigators: Efficacy and Safety of Nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med, 370(22),* 2071-82.

Schachna, L., Medsger, T.A., Dauber, J.H., Wigley, F.M., Braunstein, N.A., White, B., . . . Gelber, A.C. (2006). Lung transplantation in scleroderma compared with idiopathic pulmonary fibrosis and idiopathic pulmonary arterial hypertension. *Arthritis Rheum, 54(12),* 3954–3961.

Schwartz, D.A., Helmers, R.A., Galvin, J.R., Van Fossen, D.S., Frees, K.L., Dayton C.S., Burmeister, L.F., Hunninghake, G.W. (1994). Determinants of survival in idiopathic pulmonary fibrosis. *American Journal of Respiratory and Critical CareMedicine,* 149, (2 I), 450–454.

Swigris, J.J., Stewart, A.L., Gould, M.K., Wilson, S.R. (2005). Patients' perspectives on how idiopathic pulmonary fibrosis affects the quality of their lives. *Health Qual Life Outcomes, 3,* 61.

Thabut, G., Christie, J.D., Ravaud, P., Castier, Y., Dauriat, G., Jebrak, G., Fournier, M., Leseche, G., Porcher, R., Mal, H. (2009). Survival after bilateral versus single-lung transplantation for idiopathic pulmonary fibrosis. *Ann Intern Med, 151(11),* 767–774.

Thabut, G., Mal, H., Castier, Y., Groussard, O., Brugiere, O., Marrash-Chahla, R., Leseche, G., Fournier, M. (2003). Survival benefit of lung transplantation for patients with idiopathic pulmonary fibrosis. J Thorac Cardiovasc Surg, *126(2),* 469–475.

Turner-Warwick, M., Burrows, B., Johnson, A., "Cryptogenic fibrosing alveolitis: clinical features and their influence on survival," *Thorax,* vol. 35, no. 3, pp. 171–180, 1980.

Uno, H., Cai, T., D'Agostino, R.B., Wei, L.J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction with censored survival data. Stat Med, 10, 30(10), 1105-17. Doi: 10.1002/sim.4154

US Food and Drug Administration; Definition of Disease Prevalence for Therapies Qualifying under the Orphan Drug Act.http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/AdvisoryCommitteeforPharmaceuticalScienceandClinicalPharmacology/UCM247635.pdf. (Accessed Mar 4, 2015).

Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". *Biometrika***54**: 167–178.

Watters, L.C., Schwarz, M.I., Cherniack, R.M., Idiopathic pulmonary fibrosis. Pretreatment bronchoalveolar lavage cellular constituents and their relationships with lung histopathology and clinical response to therapy. (1987). *American Review of Respiratory Disease*, *135 (3)*, 696–704.