Improving outcomes in interstitial lung disease through the application of bioinformatics and systems biology

Arunon Sivananthan

National Heart & Lung Institute
Imperial College, London

"This dissertation is submitted for the degree of Master of Philosophy."

This is to certify that to the best of my knowledge, the content of this report has no material previously published or written by another person except where due acknowledgement and reference is made.

Contents

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisors, Professor Toby M. Maher, Dr Adam J. Bryne and Dr Philip L. Molyneaux. Their patience and direction have been of great value to me.

Special thanks to Professor Gavin Donaldson who provided additional supervision to me during my studies. His understanding, encouraging and personal guidance have provided a good basis for the present thesis.

Finally, I would like to express my thanks to my parents for their unselfish support and encouragement throughout my life.

List of Tables

List of Figures

Abstract

Idiopathic pulmonary fibrosis (IPF) and chronic obstructive pulmonary disease (COPD) are two distinct respiratory diseases whose features including pathogenesis and progression are not fully understood. However, both clinicians utilise changes in serial pulmonary function measurements to gain an insight into disease severity and control. More accurate prediction of disease progression would be beneficial, particularly for IPF given the variability in its clinical course as an unknown factor at the time of diagnosis.

Home-based, real-time monitoring of disease progression by spirometry has provided an opportunity to optimise the delivery of treatment and reduce the length of clinical trials. Therefore, the potential to understand the mechanisms underlying disease progression and generate effective treatment has been improved. In light of this, the motivation for this project is to understand the mathematical features within daily pulmonary function time series generated by IPF patients. Hopefully, statistical models of pulmonary function time series would aid the identification of significant clinical events such as acute exacerbation.

The mathematical techniques used to identify potentially important features within pulmonary function time series involved the autocorrelation function, critical transitions and detrended fluctuation analysis (DFA). Temporal properties, such as the serial correlation, abrupt changes in trends and complexity, were assessed using time series from the PROFILE clinical trial and London COPD cohort.

Forced vital capacity (FVC) measurements were found to be correlated to the previous day's reading which may inform the sampling rate of lung function during clinical trials. The presence of short-term memory within FVC time series will influence the management of missing data within clinical trials, particularly methods of imputation. Also, FVC time series' exhibit long-term memory and adaptability supporting the role of FVC as a surrogate marker for IPF disease progression.

1 Introduction

1.1 Idiopathic pulmonary fibrosis and chronic obstructive pulmonary disease

Interstitial lung diseases (ILDs) are a heterogeneous group of diseases[1] that affect the tissue and space around the alveoli, termed the interstitium. ILDs are termed as restrictive diseases on the basis that changes in lung architecture result in reduced lung volumes and inadequate oxygenation of the blood. The manifestations of ILD include respiratory symptoms such as dyspnoea, specific chest radiographic abnormalities, changes in pulmonary function tests (PFTs) as well as microscopic patterns of inflammation and fibrosis. Disease in this group vary in terms of prevention, treatment and disease outcome. A notable subset of ILDs are the idiopathic interstitial pneumonia (IIP) on the grounds that the initiating cause of disease is unknown. Idiopathic pulmonary fibrosis (IPF) is the most common form of IIP with features that include chronic, progressive fibrosis associated with irreversible lung function and high mortality.

In contrast, chronic obstructive pulmonary disease (COPD) consists of a group of poorly reversible lung diseases that cause morbidity and mortality by limiting expiratory airflow, hence is classified as obstructive. Patients afflicted with COPD suffer from mucus secretion and varying degrees of obstructive bronchitis and emphysema depending on the individual[2]. The most significant inhalational challenges that cause COPD in industrialised countries is tobacco smoke whereas in developing countries it is environmental pollutants such as those originating from cooking in confined spaces[2].

1.2 Epidemiology of IPF and COPD

The overall incidence of IPF in the USA between 2005 and 2010 was 6.1 new cases per 100,000 person-years[3]. It is more frequent in individuals who are older, male and former smokers[3]. IPF is an irreversible and progressive disease and patients have a median survival time of 2-3 years after diagnosis[3]. However, this overall survival figure obscures the variability in the rate of IPF disease progression between individuals whose decline may take a number of trajectories punctuated by life-threatening acute exacerbations[4]. Acute exacerbations are described as clinically significant respiratory deteriorations with no identifiable trigger. They are unpredictable events with no treatment and are a major cause of morbidity and mortality in IPF[5].

COPD is a major cause of global illness and death that afflicts approximately 10% of the general population[6], and up to 50% of those who are heavy smokers[7]. A rising trend in the prevalence of and mortality from COPD has been observed globally[8], with the World Health Organisation predicting that by 2020, the position of COPD amongst the most prevalent worldwide diseases rising from 12th to 5th and the most fatal diseases rising from 6th to 3rd[9]. This escalation of COPD can be attributed to reduced mortality from other diseases, such as cardiovascular disease in the industrialised countries and infectious diseases in developing countries, as well as cigarette smoking and environmental pollution.

1.3 Pathogenesis of IPF

IPF was initially perceived as a chronic inflammatory disorder that gradually evolved into fibrosis. However, this view point was challenged when immunosuppressive therapy incorporating prednisolone and azathioprine was observed to increase mortality[10,11] rather than improve patient outcome. IPF is now considered to be the result of multiple interacting genetic and environmental risk factors, with progression towards aberrant wound healing being initiated by repetitive local micro-injuries to ageing alveolar epithelium[12] (Figure 1-1).

These micro-injuries trigger abnormal epithelial-fibroblast communication that, in turn, stimulate matrix-producing myofibroblasts to remodel the lung interstitium by an accumulation of the extracellular matrix. Deposition of an altered extracellular matrix by these myofibroblasts destroys the normal alveolar architecture and disrupts gas exchange. Changes to the extracellular matrix composition impacts cell behaviour so much so that a positive loop between fibroblasts and the extracellular matrix encourages fibrosis[13].

Genome-wide association studies[14–16] and studies of familial interstitial pneumonia[3,17–20] indicate that genetic susceptibility is an important factor in the development of IPF. These studies suggest, if not a causal link, an important role of changes in host defence (*MUC5B, ATP11A, TOLLIP*), telomere maintenance (*TERT, TERC, OBFC1*) and epithelial barrier function (*DSP, DPP9*). The *MUC5B* gene encodes a protein that contributes to airway mucous production. The altered expression of the *MUC5B* gene has been localised to the bronchiolar epithelium where elevated injury is the consequence of either reduced mucociliary clearance or impede normal lung repair[16,21].

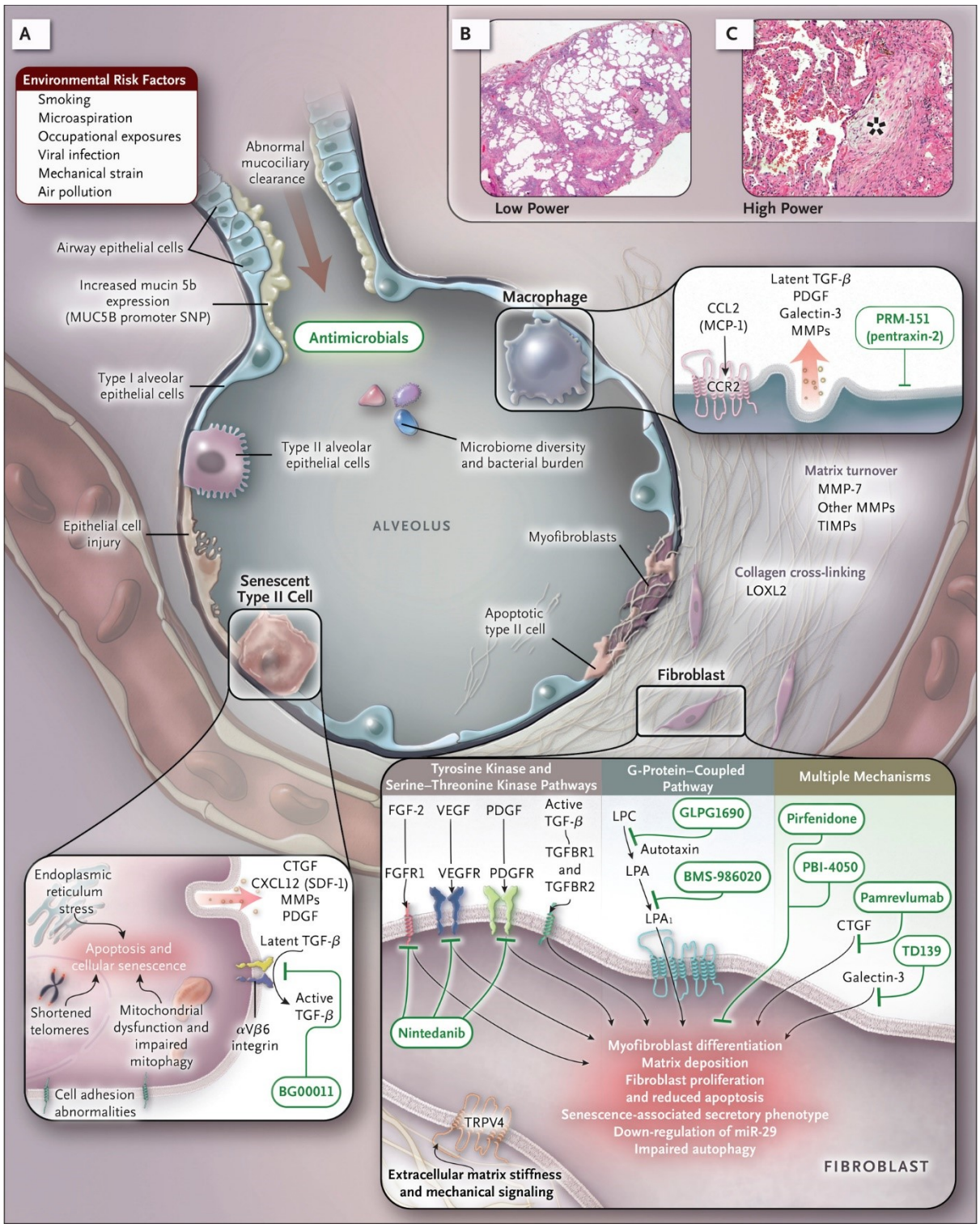


Figure 1-1 ***Pathobiologic features of IPF.*** *A schematic model of IPF pathobiology highlights how the disease is characterised by recurrent epithelial-cell injury, senescent alveolar epithelial cells, microbiome changes and host defence abnormalities (Panel A). Histology indicates the marked patchy fibrosis with the fibroblastic foci (asterisk) forming a prominent feature (Panel B and C) Source: D. J. Lederer and F.J. Martinez et al.*

1.4 Diagnosis, disease progression and treatment of IPF

IPF is defined by the American Thoracic Society (ATS)/ European Respiratory Society (ERS) classification as a "specific form of chronic, progressive fibrosing interstitial pneumonia of unknown cause, occurring primarily in older adults, limited to the lungs, and associated with the histopathologic and/or radiologic pattern of UIP"[3]. The process of diagnosing IPF requires the exclusion of other known causes of ILD and requires collaboration between clinicians, radiologist and pathologists[22]. This multidisciplinary approach combines an understanding of the relationship between histopathologic patterns and subject responsiveness to treatment, access to video-assisted thoracoscopic lung biopsy and high-resolution computerized tomography (HRCT)[22].

Difficulties in obtaining an accurate diagnosis are compounded by the unpredictable and variable evolution of disease (Figure 1-2). These issues are likely to arise as a consequence of IPF itself being the result of the interaction between a multitude of initiating stimuli, epithelial cells, immune cells, fibroblasts and cytokines[23–29]. Biomarkers therefore have the potential to contribute to our understanding of the biological mechanisms of IPF, our ability to stratify patients into precise groups within IPF, predict disease progression and response to treatment.

Acute exacerbations can occur at any time during and may be the presenting feature. The exact incidence of acute exacerbations is unclear but seems to range between 4.4-19% per year with variability due to differences in case definition, method and population[30]. Acute exacerbation are more common amongst males[31] and occur more frequently during winter months[32–34]. Smoking[35] and the administration of proton pump inhibitors for gastroesophageal reflux[36] may reduce the incidence of acute exacerbations. The outcome of patients suffering an acute exacerbation were poor in the immediate- and long-term with either death in hospital or with a shorter median survival[35].

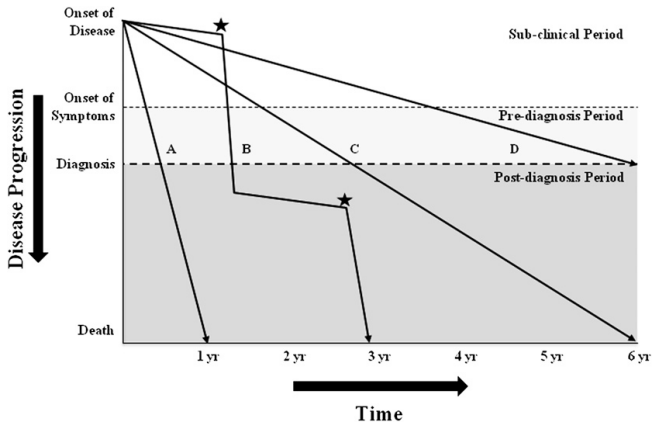


Figure 1-2 ***Examples of the three clinical phenotypes of IPF disease behaviour****. The rate of disease progression to mortality may be rapid (line A), slow (line C and D) or mixed (line B). There is little indication of how to classify the patient at the onset of symptoms. Source: Ley et al. [4]*

It is important to appreciate that the current diagnostic process for IPF relies on aggregating patients with a similar clinical phenotype and failure to meet the diagnostic criteria for other ILDs into a common group. Research into the pathogenesis of IPF would be greatly benefited by the introduction of IPF endotypes, by allowing patients to be grouped with coherent and homogenous genetic and molecular biomarker profiles. In particular, patients who have experienced an acute exacerbation are characterised by a distinct gene expression profile from those who experience stable IPF[37]. Improved understanding of the incidence and pathobiology of acute exacerbation would be benefited by the prospective collection of biological sample and lung function data[38], which has been hindered by the retrospective diagnosis of acute exacerbations.

1.5 Monitoring disease progression of respiratory illness

Systems biology can contribute to improving the diagnosis and management of IPF by, firstly, providing better non-invasive diagnostic markers. In doing so, this could increase the confidence with which clinicians are able to diagnose IPF with HRCT[39] alone, avoiding invasive biopsy procedures which carry a significant morbidity and mortality. Secondly, more descriptive diagnostic markers of IPF permits better stratification of patients which will form a foundation for clinical trials. Thirdly, diagnostic biomarkers may help guide therapy, highlighting specific endotypes which are more likely to respond to therapy and also allowing the assessment of response to any administered anti-fibrotic treatment[40].

Several therapeutic and diagnostic challenges for IPF exist in clinical practice such as identifying those subjects who need lung biopsies to confirm diagnosis[41] or identifying individuals who may respond preferentially to one of the two available pharmacotherapies. Current pharmacotherapy for the treatment of IPF subjects is limited to the anti-fibrotic drugs pirfenidone and nintedanib[42]. These two treatments slow the rate of functional impairment, but are unable to prevent the advancement of the disease[41].

The clinical management and treatment of IPF is complicated by patients experiencing variable disease course and response to anti-fibrotic treatment. More accurate prediction of IPF disease progression would facilitate clinicians to make treatment decisions and counsel patients more appropriately. Traditionally, clinical trials have been designed using serial changes in FVC over 12-months to assess disease progression[43–46] meaning that trials can be long. Improved prediction of disease progression would benefit clinical trials by reducing their duration, allowing cohort enrichment[47], alleviating lead-time bias[4] and accounting for a background of anti-fibrotic therapy[48].

Disease progression tend to be readily evaluated by physiological parameters including worsening symptoms (e.g. worsening dyspnoea)[49,50], worsening physical function[51–53] and the occurrence of acute respiratory worsening requiring hospitalisation[35,43,54]. However, the serial change in FVC is a commonly accepted measure of disease progression in IPF[43,45,55–58], with its predictive power of survival improving with time from six- to twelve-months[58]. Clinical information used to predict the mortality of IPF patients tend to poorly predict future disease progression.

The formal integration of relevant clinical parameters into a classification scheme, or staging systems, that guide management decisions have been adopted for diseases such as lung cancer, HIV/AIDs and COPD[59]. Several baseline features are useful for the staging of IPF, including age, desaturation during the 6-minute walk test or honeycombing on high-resolution computed tomography (HRCT)[55]. Traditional approaches for staging IPF disease progression, for example the gender, age and physiology (GAP) model[60] endeavour to use reliable baseline parameters of IPF disease severity and progression, like single breath diffusing capacity for carbon dioxide (DLco) and FVC, to assess risk of mortality. However, their use in clinical practice is limited because IPF disease progression is not linear and do not reflect distinct biological or clinical phenotypes.

The primary endpoint chosen for monitoring IPF disease progression is important for demonstrating success of any therapeutic agent during future clinical trials and indicating when treatment efficacy changes. FVC is the most commonly accepted primary endpoint in clinical trials of IPF[61] given that it is easy to measure, reliable and reproduce. FVC is a highly relevant measure of lung function thereby reflecting the burden of the fibrotic disease process. The preference for FVC as a primary endpoint despite DLco being a better prognostic indicator[62] is because DLco is difficult to measure, involves a breath hold which may be challenging for symptomatic patients and yields greater intrinsic variability.

Whilst both the baseline FVC and change in FVC are indicative of mortality[45,62,63], the minimal clinically important difference (MCID) is a source of contention. Designating a threshold for a significant change is advantageous over considering FVC as a continuous variable because IPF disease progression can progress in a stepwise manner, it would allow patients who fail to respond to treatment to exit a trial and allow combining a decline in FVC and mortality to assess progression-free survival.

A clinically relevant threshold of a 10% decline in FVC percent-predicted has been established over 24-weeks[3] because it is associated with a fivefold increased risk of mortality[63]. The use of relative change in FVC for the assessment of IPF disease progression is preferred as it adjusts for the different clinical implications any given change in lung function has depending on disease severity and provides a higher prevalence for the decline signal. A marginal threshold of a 5-10% FVC has been associated with a twofold increase in 1-year mortality[63], which is not sufficient to influence management but can guide the assessment of disease progression in the context of symptomatic deterioration or changes in chest radiography[45]. The FVC threshold is calculated from two serial measurements which are traditionally 3 months apart, however there is no evidence to support whether this is the optimal period.

IPF disease progression is unpredictable because of variability between individuals and the interspersion of life-threatening acute exacerbations and episodes of infection. This means that the low frequency of serial measurements for surrogate markers such as FVC and DLco is not sufficient to capture enough information to accurately calculate the rate of disease progression. Therefore any future staging system would need to integrate baseline information with longitudinal parameters such as categorical decline in FVC of greater than 10% predicted[11]. An innovative advancement is the feasibility assessment of home-based spirometry[64] that examined whether daily observations of lung function can provide high-resolution monitoring

of IPF disease progression. High frequency monitoring has the potential to detect subtle changes in lung function prior to the establishment of irreversible fibrosis, acute exacerbations and early identification of individuals with rapid-progressive disease.

High-frequency measurements also contribute to minimising the inaccuracy introduced to endpoint analysis by missing data associated with trial dropouts and deaths. In addition, the alleviated frequency of observations allows for improved precision for estimates of disease trajectory thereby decreasing sample size and/or duration of clinical trials. Given that future early-phase clinical trials will be conducted on a background of anti-fibrotic therapy that slow disease progression, the benefits of daily home spirometry will contribute to overcoming the difficulties of identifying incrementally beneficial therapies.

1.6 Home monitoring of respiratory disease progression

Despite the theoretical benefits, conducting home spirometry are associated with limitations from resource implications and inconvenience to patients. Daily home spirometry measurements consistently underestimate lung function compared to hospital-based observations obtained under the supervision of a respiratory physiologist with hospital-based lung function equipment[64]. As a testament to the benefit of longitudinal data, the predictive value of serial FVC measurements for IPF disease progression remain because the underestimate was constant for all time points.

Any residual concerns relating to feasibility and data integrity when considering primary endpoints derived from home-based assessments were alleviated by an investigation of weekly home spirometry[65]. However, an outstanding question relates to the optimum frequency for home spirometry and whether the additional data points provided by daily spirometry further reduce trial numbers enough to justify patient inconvenience. In order to fulfil the potential of home spirometry, electronic appliances or personalised devices that permit automated data acquisition and data processing need to be introduced.

A home monitoring system that is wireless and real-time, rather than reliant on paper-based collection, would allow for quality control of measurements and prompt timely responses to FVC decline or non-adherence. The integration of home monitoring with real-time wireless home spirometry has been demonstrated as feasible and readily received by patients[66]. This is encouraging given that it allows for the continual monitoring of a patient population with

progressive breathlessness and declining mobility whilst reducing the burden of hospital visits. Real-time uploading and automated email alerts would allow for the quality review of measurements and detection of FVC decline.

The ability to predict the course of IPF disease progression using physiological markers can be improved with molecular and genetic biomarkers[47]. The need for prognostic biomarkers is compelled by the introduction of the sole effective therapies for IPF, pirfenidone and nintedanib[67], because they reduce the change in physiological parameters being used to assess mortality risk or disease progression. Biomarker discovery for IPF has been hampered by a variety of redundant molecular pathways that contribute to the aberrant inflammation and wound repair observed amongst patients[68]. Thus, further motivation for biomarker discovery are the elucidation of relevant disease mechanisms, identification of therapeutic targets for treatment and assessment of therapeutic responses during clinical trials[69].

1.7 IPF clinical trials

The ASCEND (Assessment of Pirfenidone to Confirm Efficacy and Safety in Idiopathic Pulmonary Fibrosis) pirfenidone trial[70] and the two INPULSIS™ nintedanib trials[71] are significant in that they have clarified the treatment effectiveness of the first available IPF therapies. Patient management has been transformed by both these drugs having been approved worldwide as IPF treatments[70,71]. Nintedanib is a tyrosine kinase inhibitor that suppresses multiple signalling receptors involved in fibrosis pathogenesis, including fibroblast growth factor receptor[72,73]. Pirfenidone is an orally administered pyridine with combined anti-inflammatory, antioxidant and antifibrotic actions, however the precise mechanism of action remains ambiguous[74,75].

Regardless of their individual mechanisms of action, both drugs exhibit similar reductions on the rate of decline in FVC over 1 year[40,70,71]. Neither drug demonstrated any survival benefit during the respective trials, however a trend towards declining mortality was perceived. Therefore, the safety profile and tolerability of these drugs will guide patient and provider choice.

Whilst the ASCEND and INPULSIS clinical trials represent major breakthroughs for IPF patients, extrapolating their findings to patients who are outside the recruitment criteria should be done with extreme caution. These studies do not provide any indication as to the

effectiveness of these drugs for patients with severe disease (FVC <50% of the predicted value) or with an acute exacerbation. In addition the effectiveness of these drugs beyond 1-year, in combination with each other or efficacy for pulmonary fibrosis of other ILD forms has yet to be characterised.

The PROFILE (Prospective Observation of Fibrosis in the Lung Clinical Endpoints) study[76] is the largest longitudinal study which recruited patients diagnosed with either IPF or idiopathic non-specific interstitial pneumonia to two coordinating centres (Nottingham, UK and Royal Brompton Hospital, UK). This study has established that dynamic changes in biomarkers of extracellular matrix turnover of the lung are able to predict IPF disease progression[76] and that epithelium-derived proteins in the serum are able to identify disease course and risk of death[77] in treatment-naïve patients.

1.8 Diagnosis of COPD

The diagnosis of COPD relies on the patient history, a physical examination, chest radiography and the use of spirometry to prove airflow obstruction. The spirometric contribution to COPD diagnosis relies on establishing that the post-bronchodilator ratio of forced expiratory volume in 1 second (FEV_1) to FVC as smaller than 0.70 to indicate significant airflow obstruction[78,79]. Having established a diagnosis, disease severity can be assessed using percentage predicted post-bronchodilator FEV_1 as it is objective, reproducible, corresponds well with disease severity and prognosis[80,81]. In addition, serial FEV_1 measurements aids the management of COPD by allowing the progress of disease to be monitored and guide treatment options as the disease enters different stages of the disease.

The serial assessment of physiological and clinical outcomes in IPF and COPD is crucial to the design and interpretation of clinical trials, requiring a different combination of longitudinal markers that reflect their unique aetiology. Already noted is the use of FVC as a measure of the fibrotic burden for IPF patients compared to FEV_1 as an indication of airflow obstruction for COPD disease[82]. Both diseases have the potential to benefit from the use of biomarkers to assess disease heterogeneity and predict disease progression. Whilst inflammatory markers may seem to correlate with disease severity and risk of exacerbation in COPD[83], proteins expressed by the lung epithelium may reflect disease progression for IPF patients[67]. However, both types of respiratory disease benefit from similar methods to assess serial exercise capacity, the six-minute walk test, and health status, the St. George's Respiratory Questionnaire.

1.9 Mechanisms of COPD

COPD is characterised by airflow obstruction that is the consequence of an aberrant inflammatory response in the airways and alveoli. The reduction in the airway diameter and corresponding increased resistance to flow[84] are because of the unusual presence of inflammatory cellular infiltrates that cause the airway to thicken by remodelling its architecture (Figure 1-3)[85]. In addition, inflammatory infiltrates within the alveolar walls are associated with alveolar destruction and enlarged air spaces. T-cell mediated inflammation that can persist years after the cessation of smoking is a key component of COPD[86,87].

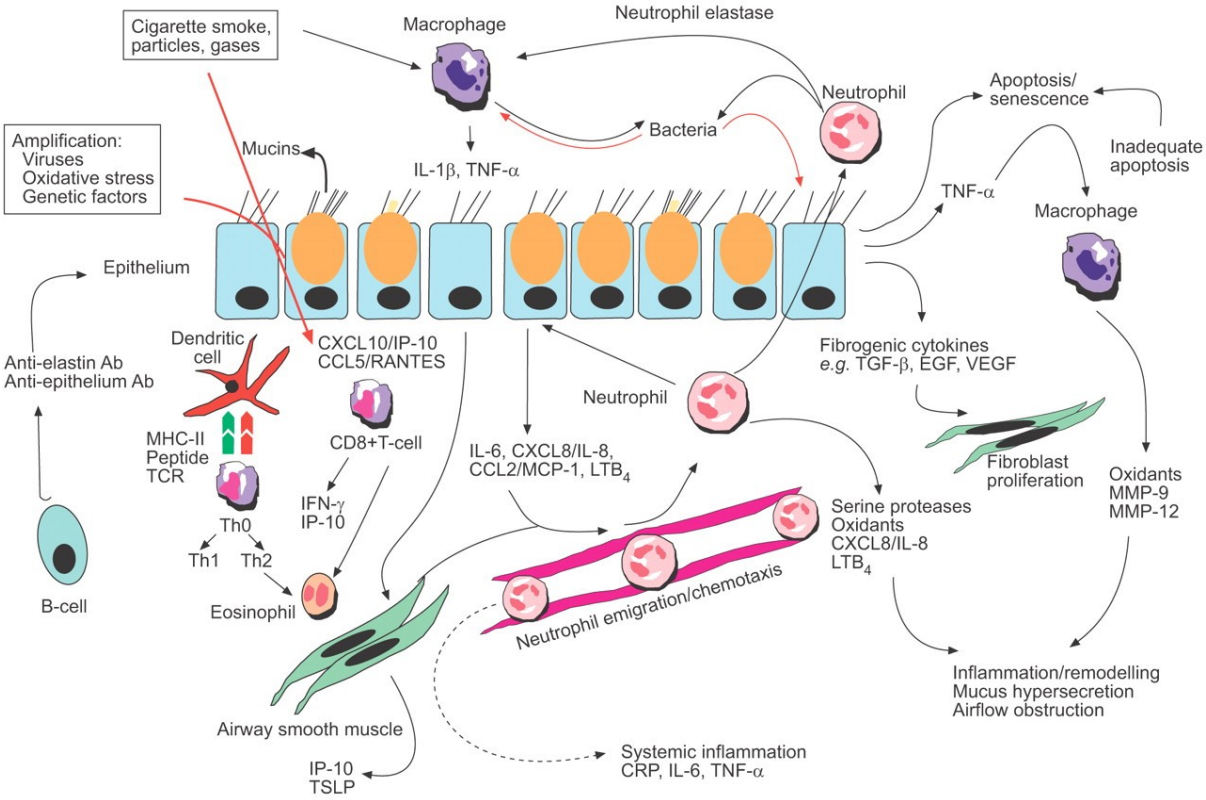


Figure 1-3 ***Summary of inflammatory and cellular interactions linking chronic cigarette exposure to the chronic inflammation of COPD.*** *. Amplification signals released by inflammatory cellular infiltrates are important in augmenting the inflammatory responses that underpin COPD. Source: K.F. Chung & I.M. Adcock [85]*

Cigarette smoke contains xenobiotic compounds and free radicals that injure the lung epithelium, the extent is proportional to their concentration[88,89]. Epithelium-cell injury stimulates an innate immunity cascade that ultimately cause alveolar macrophages and neutrophils to secrete proteolytic enzymes and reactive oxygen species to further damage the

lung[90]. Eventually, cytokines and chemokines produced by macrophages and dendritic cells coordinate the inflammatory response required to activate the adaptive immune system[91,92].

CD8+ and CD4+ cells activated by antigen-bearing dendritic cells are recruited to the lung in a tissue-specific manner. The severity of COPD is proportional to the amount that structural cells of the airways and pulmonary arteries express the chemoattrative ligands for type 1 helper and cytolytic T cells[93–95]. The lung destruction in COPD is the consequence of CD8+ cell, and possibly protease, induced cell death of epithelial and endothelial cells which is not replaced by cell proliferation[96]. B cells have been identified in the lymphoid follicles of the airways and parenchyma of COPD patients[97,98] are likely to promote immune and complement deposition and therefore any tissue injury and airway remodelling. Taken together, tissue destruction perpetuates the cellular mechanisms contributing to COPD by generating additional antigenic material.

Due to the destruction of the alveolar walls, the total respiratory surface is reduced meaning that the gas exchange of oxygen and carbon dioxide is impaired. This is compounded by increased airway resistance resulting in breathing difficulties, air trapping in the lung, and hyperinflation of the lung. Air trapped in the lungs means that there is impaired regional ventilation corresponding to a ventilation-perfusion (V/Q) mismatch.

1.10 Approaches to time series analysis

The analysis of longitudinal spirometric observations to monitor the rate of IPF or COPD disease progression can be considered to be an integral stage in a home telemonitoring workflow. An established home telemonitoring workflow would be an automated process involving the transmission of data regarding a patient's lung function from their home to their health care provider[99] coupled to the continuous acquisition and processing of data with decision support. Therefore, home telemonitoring has the potential to provide high-frequency monitoring of patients, identification of early symptoms and prompt responses to acute exacerbations.

The development of techniques in physiology, like home spirometry, will provide us with insights into biological processes by generating data. However, in order to fully understand the processes and their dynamic interplay within respiratory diseases, the development of statistical tools that will aid the utilisation of this information is required. Traditional statistical methods that provide summary measures like the mean, standard deviation or absolute change can be

useful to conceptualise the time series. However, time series methods will permit us to account for the time-ordered nature of home spirometric data thereby allowing us to characterise any periodicity and correlations within the data. In doing so, time series analysis provides us with tools required to characterise simple and complex biological behaviours, ultimately allowing us to predict outcomes.

This will be instrumental in the provision of timely care and therapy as it would help detect individuals who are experiencing rapidly progressive disease or suffering from an acute exacerbation. The use of interrupted time series analysis for the data processing of longitudinal data would be important in determining the efficacy of therapies and guide when treatment strategies need to be changed. Indeed, maintaining the FVC of IPF patients close to their baseline levels corresponds with a better outlook for individuals[43,63,100]. An initial step in the interrogation of longitudinal data is exploratory data analysis that seeks to establish the patterns in systematic variation between groups of patients and the features of random variation that discriminate individual patients.

Exploratory data analysis provides the foundation for determining the features of spirometric longitudinal data pertinent to development of statistical models of IPF disease progression. A popular statistical method for modelling and forecasting of time series is the ARIMA model. It is a category of model with the acronym for AutoRegressive Integrated Moving Average that incorporates a range of different temporal structures that exist within time series data, including autocorrelation, trends or seasonal variation. This forms the basis as to why time series models possess greater predictability and wider applicability than non-temporal techniques[101]. ARIMA models have been employed in a range of applications from managing healthcare resources like the use of hospital beds during epidemic incidences of the severe acute respiratory syndrome (SARS)[102] and haemorrhagic fever with kidney fever[103] to predicting and investigating antimicrobial resistance[104–106].

The modelling and analysis of cardiovascular and respiratory time series can be categorised either into linear mechanistic models[107,108] or nonlinear descriptive indices[109–111]. Whilst linear techniques have the advantage of unveiling the individual relationships between observed variables, non-linear indices of complexity capture the dynamical behaviours within specific underlying mechanisms. The aberrant wound healing process that drives IPF disease progression and inferred by the loss of lung volume, is being driven by multiple redundant pathways[68]. In the longer term, the use of non-linear time series techniques will aid

understanding the contribution of specific biological pathways to IPF disease progression. The eventual incorporation of serum-based biomarkers[76] into any statistical analysis will allow us to derive inferences regarding the physiology underpinning the IPF disease process by helping to account for biological variability and effects of intervention that contribute to the noise within the medical time series.

1.10.1 Detrended Fluctuation Analysis

The respiratory system is a complex system whose physiological phenotype arises from the non-linear interaction of environmental stimuli with a complex web of immunological, mechanical and inflammatory components[112]. These comprise of structural and regulatory feedback loops that operate at varying temporal and spatial scales thereby allowing adaptation to the stresses of everyday life[113]. Therefore, a healthy, stable system is associated with physiological parameters that fluctuate continuously under non-equilibrium steady-state conditions to maintain adaptability to external or internal stimuli[114]. Hence, analysis of these fluctuations may harbour information about the adaptability of the physiological system, but it is yet unknown whether any given disease state is associated with increased regularity or irregularity[115].

Fractals are used to describe the relationship between spatial or temporal patterns within complexity analysis, where a smaller structure resembles a larger scale form[116], a phenomenon described as self-similarity. Detrended Fluctuation Analysis (DFA) is a statistical technique for analysing self-similarity within a time series by providing a quantitative parameter termed the scaling exponent, which represents the correlation properties within the time series. The DFA method has been used to characterise the long-range correlations within many fields of research like cardiac dynamics[117,118] and bioinformatics[119]. Furthermore, this technique can be used to identify different states of the same system and so distinguish between healthy and sick individuals based on heart inter-beat intervals[120] or the efficacy of therapeutics for asthmatics[121] or chronic obstructive pulmonary disease[122].

The scaling exponent, as quantified by DFA, measures the long-range correlations within the physiologic time series of lung function such as peak expiratory flow (PEF)[121,122], thereby providing an indication of the temporal history of the disease. Successive time points are considered to have no relationship with each other when the scaling exponent, as calculated by DFA, is 0.5. However, if the scaling exponent is greater than 0.5, long-range correlations exist

within the time series. The associated system can be described as having a memory where the current situation is influenced by its temporal history.

Fluctuation analysis, and specifically changes in the scaling exponent, can provide insight as to the optimal timing of regularly given drugs for chronic disease, e.g. the regular administration of short-acting β_2 agonists for asthma tends to drive the internal regulation of airway tone towards a random process and therefore a less stable system[121]. The scaling exponent could potentially be used as a surrogate marker for any pathological condition, given that it is important to establish how the scaling exponent changes with the risk of each pathological condition. For example, whilst higher values of the scaling exponent are associated with decreased risk of severe asthma episodes[121,123], they are associated with an increased exacerbation frequency in COPD[122].

1.10.2 Critical transitions

An acute exacerbation exhibits the typical characteristics of a critical transition in that there is a qualitative, rapid change in the disease progression and the crossing of a threshold resulting in a new 'plateau' (Figure 1) from which disease progression continues as before[124]. The study of critical transitions in complex systems such as ecosystems and societies suggest that as a system approaches a tipping point, a set of circumstances arise that exaggerate a minor perturbation towards an alternative state[125]. A reduction in a system's ability to recover is termed "critical slowing down" which can be equated to an inability of the IPF lung to heal from injury.

Whilst much progress into the mathematical modelling of tipping points has been made with respect to catastrophic meltdowns in financial markets, their use within the financial field has been limited by stochastic modelling of parameters relating to investor and market behaviours[126]. This means that abrupt transitions in financial markets are characterised as a stochastic not critical transitions on the basis that they can occur away from the tipping point. The Lehman Brothers bankruptcy is an example of how stochastic transitions can occur without much warning because of the way in which financial institutions set interest-rate swaps in maturities. This meant that financial institutions became interdependent on each other to cater for their commitments[127].

The extent to which the financial institutions are dependent on each other reflects the likelihood that a stochastic transition will occur should one of them default. In contrast to critical transitions, stochastic transitions tend to occur without early warning signals meaning that predicting a transition is difficult e.g. a financial institution defaulting on a loan. However, the modelling of an acute exacerbation within the IPF lung as a critical transition suggest that there are indicators of the fragility of the lung to the injury and that its ability to restore lung architecture to normal is diminished.

Modelling an acute exacerbation as a critical transition suggest that early warning signals, called dynamical network biomarker (DNB)[128], have the potential to anticipate the transition. The pre-disease state is characterised by little or no symptoms but the expression of these DNBs may indicate that a critical transition in the IPF lung is more likely. It is likely that these DNBs reflect molecular pathways that make the IPF lung more susceptible to periods of dramatic decline e.g. Krebs von den Lungen-6 (KL-6)[129] as a marker of alveolar epithelial cell damage. High-throughput OMICS data can be obtained from a small number of samples for each individual and correlations between the variables can be made based on the high dimensions.

In addition, the fragility of the lung can be measured by parameters that assess lung function i.e. spirometry on a regular basis. The data collected allows many samples for each individual to be collected that have low dimensions. This information is invaluable as it permits a non-invasive method of assessing the health of lung and provides an opportunity to assess how an IPF patient's disease is progressing as it is happening. Therefore, it is possible to pinpoint an exacerbation to a narrow window of time about its occurrence.

1.11 Motivation for study

The motivation for this study is to develop a statistical model that will contribute to monitoring patient wellbeing by detecting individuals with rapid progressive disease or those suffering acute exacerbation. Time series generated by the IPF patients performing home spirometry on a continual basis will form the basis of the algorithm been developed. Any statistical model would benefit from the increasing size of a longitudinal data set to establish short- and long-term trends in changes in lung function. Therefore, this statistical model represents an improvement on the contemporary approach to classifying IPF disease progression which utilises two serial observations twelve-months apart[44].

1.12 Study hypothesis

Whilst individual variables are associated with IPF patient mortality, these variables are limited in their ability to accurately predict prognosis in isolation[4,53]. However, deterministic analysis has the potential to overcome the prognostic limitations of these variables by examining the temporal patterns in airway function by utilising all points within a time series. We hypothesize that the clinical course of IPF can be classified by deterministic methods applied to longitudinal lung function data.

1.13 Objectives of the study

The main aims of this study are set to provide a better understanding of the temporal features within the longitudinal lung function data of IPF patients and allow better, more timely classification of IPF disease progression. Therefore the four objectives of this thesis are

1. to gain an understanding of the physiological parameters impact survival and clinical classification in patients with IPF.
2. to identify the temporal components within the longitudinal lung function measurements of IPF patients.
3. to assess the variability of the IPF airway calibre.
4. to compare and validate the modelling approach to best identify changes in IPF disease progression.

2 Methods

2.1 PROFILE IPF cohort

2.1.1 Patients

50 patients were identified as a subgroup of the PROFILE (Prospective Observation of Fibrosis in the Lung Clinical Endpoints) study[76] from the Interstitial Lung Disease unit at the Royal Brompton Hospital, London who had a consensus diagnosis of IPF according to current international criteria[3]. Ethical approval was given by the Royal Free Hospital and Medical School Research Ethics Committee, and all patients provided signed, informed consent. Patients were recruited into the PROFILE study within 6 months of being diagnosed with IPF. Clinical assessment was undertaken at baseline and at 3-monthly intervals, whilst full hospital-based lung function testing (spirometry, plethysmography, and gas transfer) was undertaken at baseline, 6- and 12-monthly intervals. Patients were monitored until death or 1st April 2018.

2.1.2 Spirometry

Patients participating in the study were given a portable hand-held Micro spirometer (CareFusion, Kent, United Kingdom) that provides a digital read-out of FEV_1 and FVC in litres. The accuracy of the Micro spirometer is factory calibrated to ±3% for FVC within the range of 0.1-9.99L. Patients were given an hour dedicated instruction as how to perform spirometry which was reinforced a month later. Each patient was required to perform a single spirometry manoeuvre at the same time of day and record the reading in a dedicated diary. FVC measurements were recorded in litres to two decimal places. Prior to entry into an electronic database, measurements were reviewed by study staff. Patients were requested to contact study staff should their FVC measurements fall by 10% of baseline over 3 or more consecutive days. Spirometry conducted in a hospital environment were obtained in accordance with current international standards such that the best value of three technically adequate forced expiratory manoeuvres was recorded[130].

2.2 London COPD cohort

2.2.1 Patients

32 COPD patients enrolled in the London COPD cohort and who contributed at least 1 year of data between January, 1 1996 and December, 31 2002 were analysed. Inclusion criteria comprised of a post-bronchodilator FEV_1 <80% predicted for age, height and sex and a

FEV_1/forced vital capacity (FVC) ratio <0.7[122]. Exclusion criteria were applied to patients with significant respiratory disease except COPD or the inability to complete diary cards.

Ethics approval was acquired from the East London and City Research Ethics Committee as well as the Royal Free Hospital NHS Trust Ethics Committee. In addition, all patients gave written informed consent.

This patient cohort has been the subject of previous publications pertaining to exacerbations[131–134] and time series complexity[122] but the comparison of complexity between restrictive and obstructive respiratory diseases is novel.

2.2.2 Recruitment

At recruitment, lung function (FEV_1 and FVC) measurements were obtained using a spirometer (Micro Medical Ltd, Chatham,Kent, UK)[132]. A history of reversibility to salbutamol after withdrawing bronchodilators, the number of exacerbations during the previous year and smoking habits were recorded.

2.2.3 Monitoring and diagnosis of exacerbation

At recruitment, patients were trained how to record in diary cards changes in respiratory symptoms. Major symptoms were classified as dyspnoea, sputum purulence and volume. Minor symptoms were categorised as coryza, wheeze, sore throat and cough. In addition, patients manually recorded the best of three daily morning post-bronchodilator peak expiratory flow (PEF). The mean PEF and coefficient of variation was calculated over the subsequent 365 days. PEF is expressed as a percentage of predicted value and calculated using the equations of Garcío-Río *et al.*[135].

An exacerbation was diagnosed on the basis of the patient experiencing two or more consecutive days of an increase in either two major symptoms or a combination of major and minor symptoms[131,132]. In addition, exacerbations were identified when patients were admitted to hospital without associated changes in symptoms being recorded or during enquiry at clinic visits about primary care treatment[131,132]. The ratio of the number of exacerbations to the period (in years) of observation gave the exacerbation frequency. The exacerbation duration was defined by the number of days after the exacerbation onset that symptoms were still

experienced. Two consecutive days that the patient was symptom free defined the end of an exacerbation. An exacerbation was capped with a maximum duration of 100 days.

2.3 Statistical analysis

Data was analysed with R statistical computing platform version 3.5.1. Normally distributed data were expressed as mean ± SEM and comparison made using the t-test.

2.3.1 Exploratory Time Series Analysis

The lung function measurements (forced vital capacity (FVC) and forced expiratory volume in 1 second (FEV_1)) obtained from the PROFILE and London COPD cohorts form univariate time series given that they are time-ordered observations occurring at daily intervals and originate from a single individual. Time series can be characterised by three basic features consisting of variation, autocorrelation and stationarity[136]. There are four components of time series, comprising of trend, seasonality, cycles and irregular variation with the length of individual time series' needing to be long enough to capture phenomena of interest such as an acute exacerbation (change in trend) and/or seasonality.

Trend, seasonality and cycles represent systematic variability or regular patterns within the time series that can be modelled over different periods of time. If a statistical model accounts for all the systematic variance within a time series, the remaining component should be completely random or white noise. Autocorrelation explores the effect of previous states on attempts to incorporate this source of variation into the statistical model by examining the Pearson correlation of the variable with itself at various time intervals or lags[136]. Time series must be stationary, or have no trend, to assess the presence of autocorrelation.

The augmented Dickey-Fuller (ADF) statistical test evaluates the null hypothesis that a given time series is non-stationary, whilst the Durbin-Watson statistical test assesses whether the residuals of a regression model are autocorrelation[136]. A non-stationary time series can be made stationary by a process of differencing that calculates the differences between consecutive observations and thereby eliminates the trend. Differencing can be achieved using the diff function within the R statistical computing platform. Evaluation of autocorrelation between observations at different lags, or time intervals, is done by using the complete and partial autocorrelation functions together. The distinction between the complete and partial

autocorrelation functions as measures of correlation of time series separated by k time units (y_t and y_{t-k}) is whether they adjust for the presence of the other terms at shorter lags (y_{t-1}, y_{t-2},..y_{t-k-1}).

Heteroscedasticity is a property concerning statistical variability within any given individual time series. The presence of this property can invalidate statistical tests of significance performed during regression analysis as they assume that modelling errors are uncorrelated and uniform. The Levene's test is utilised to assess homogeneity of variance by testing the null hypothesis that population variances are equal. If the p-value is less than the level of significance then the null hypothesis can be rejected and the time series has heteroscedasticity.

2.3.2 Detrended fluctuation analysis

FVC time series of 38 IPF patients and the entire COPD patient cohort, having a minimum length of 75 days were assessed, with detrended fluctuation analysis (DFA)[121]. The first step of the DFA algorithm is the integration of the entire time series according to:

$$y(k) = \sum_{i=1}^{k} (x(i) - \bar{x})$$

where $x(i)$ is the *i*th FVC measurement and $\bar{x}$ is the average FVC measurement. Next, the integrated time series is divided into boxes of equal length, *n*. Within each window of length *n*, a least-squares linear regression is fit to the data to represent the local trend. The y-coordinate of the straight line segments is denoted by $y_n(k)$. The entire time series was detrended by subtracting the local trend, $y_n(k)$, from the data. The second step of the DFA algorithm aims to calculate the root-mean-square fluctuation of this integrated and detrended time series as follows:

$$F(n) = \sqrt{\frac{1}{N} \sum_{i=1}^{k} [y(k) - y_n(k)]^2}$$

The DFA algorithm is repeated over all time scales (box sizes) to determine the relationship between *F(n)* and box size. This relationship can visualised on a double log plot of *F(n)* against

n, and a summary statistic referred to as the scaling exponent (α) obtained from gradient of a straight line fit according to

$$F(n) \propto n^{\alpha}$$

2.3.3 Imputation of missing values

Missing data is a characteristic that features within the spirometric time series of patients to varying extents. The univariate, FVC time series that are the focus of this study provide a challenge to standard imputation algorithms that use inter-attributable correlations to estimate values for missing data[137]. Effective univariate algorithms exploit individual time series characteristics meaning that imputation must be done on an *ad hoc* basis. The Kalman filter is the most effective method for dealing with missing data that can be implemented in R using the zoo package[138].

2.3.4 Critical Transitions

An acute exacerbation is defined as being an abrupt, accelerated decline in lung function (>10% FVC) over a period of 10- to 14-days[64]. Serial FVC measurements provide an important insight into disease progression and can provide an indication of an acute exacerbation at its initial stages. Information about the drivers or conditions at which an acute exacerbation (critical threshold) is currently unknown and difficult to acquire. Leading indicators can act as generic early warning signals of patients who approach an acute exacerbation and are based on common mathematical properties that have the potential to be used in real-time[124].

As a system approaches a critical transition, the rate of return to equilibrium following a small perturbation declines[139], termed "critical slowing down" [140]. Critical slowing down causes an increase in short-term memory or correlation at low lags of a system prior to a transition that can be measured by autocorrelation[140]. Autocorrelation can be estimated according to

$$\rho_1 = \frac{E[(z_t - \mu)(z_{t+1} - \mu)]}{\sigma_z^2}$$

An alternative measure of critical slowing down is variance because a system close to a transition would tend to drift more widely around its stable state. Variance can be measured by the standard deviation as defined by

$$SD = \frac{1}{n-1}\sum_{t=1}^{n}(z_t - \mu)^2$$

An acute exacerbation has been defined as an acute clinically significant deterioration of unidentifiable cause associated with pronounced architectural distortion, decline in lung volumes and impaired gas exchange[141]. Therefore, the pre- and post-transition state of FVC measurements will form two distinct distributions. The calculation of a composite index[142] allows a comparison of the correlation between the local, contemporary group of FVC measurements (PCC_{in}) with the that of FVC measurements since baseline (PCC_{out}) whilst accounting for the noise or standard deviation of the FVC time series according to

$$I = SD\frac{PCC_{in}}{PCC_{out} + \varepsilon}$$

where ε is a comparatively small positive constant to avoid zero division. The correlation of the local group is calculated from most recent FVC measurement and the two previous readings whilst the correlation of the group of observations since baseline will increase with time upon each calculation.

2.3.5 Piecewise linear regression

Each pair of observation within each patients' spirometric time series can be described by an explanatory time variable and response composite index calculation. The relationship between time (x-axis values) and composite index (y-axis values) can be described by two linear regression functions[143] as described by

$$y_1 = a_1 + b_1 \cdot x_i \qquad x_i \leq x_0$$
$$y_2 = a_2 + b_2 \cdot x_i \qquad x_i > x_0$$

where (x_0,y_0) represents the join point, $a_{1,2}$ are the intercepts of the respective regression lines and $b_{1,2}$ are the slopes of the respective regression lines.

The response of the composite index calculation can be described by two linear regression functions with a join point (x_0) that is defined by a point where the regression lines intersect.

An iterative approach can be employed to ascertain the joint point in a finite number of steps using statistical approaches. In each iteration, the time series is partitioned into two groups containing successive observations and the parameters of each linear regression is estimated. The residual sum of squares is calculated for each linear regression function and added together to assess the quality of fit at each iteration.

For successive iterations an observation is shifted between the two parts of the time series with the parameters and residual sum of squares being recalculated each time. The iteration corresponding to the least residual sum of squares relates to the parameters of the optimal fit. Having selected the optimal fit, the angle between two regression lines at the point of intersection can be ascertained from

$$tan_{\alpha} = \frac{m_2 - m_1}{1 + m_1 m_2}$$

where m1 and m2 are the gradients of the respective regression lines and α denotes the angle at the point of intersection.

2.3.6 Receiver operating characteristic

The ability of the statistical model to classify patient disease progression with increasing time was estimated using receiver operating characteristic (ROC) curves. To estimate the classification accuracy using standard ROC methods, the gold standard or true disease progression status was available from clinical follow-up. Sensitivity (i.e. true positive rate) and specificity (i.e. true negative rate) are the fundamental measures that contribute to ROC analysis. Given the principal physiologic medical test result used to monitor IPF disease progression, forced vital capacity (FVC), is a continuous measurement, a threshold value of 10% baseline over 12 months is used to classify patients[44].

Individuals with a change in FVC measurement equivalent to or greater than 10% of their baseline reading over 6 months are classified as suffering rapidly progressive disease whilst those with a change in FVC corresponding to less than 10% baseline over 6 months are classified as experiencing steady illness. A 2X2 contingency table is formed containing counts of the four possible combinations of classification and true disease may be formed, namely true or false negatives and true or false positives. The accuracy of classification is assessed each

day for the 50 IPF patients of the PROFILE cohort in terms of the probability a rapid progressing patient is classified as positive (sensitivity) and the probability a patient experiences steady disease is classified as negative (specificity).

2.4 Survival Analysis

The prognostic value of FVC change modelled as a continuous value becomes meaningful sometime between 28 days and 3 months[64]. Successful prediction of clinical outcomes for IPF patient needs to account for non-linear lung function decline experienced by those suffering acute exacerbations. Hence, survival analyses were performed using R statistical computing platform version 3.5.1 supplemented with functions from the survival and survminer libraries.

2.4.1 Kaplan Meier survival estimate

The Kaplan-Meier estimator[144] is a non-parametric statistic that permits the survival function of the patient cohort to be determined and reflects the probability that an individual patient will survive past a given time. Preparation for Kaplan-Meier analysis involves the construction of a table of patient variables containing three essential input parameters, serial time, status at serial time and study group. The table is sorted by ascending survival times starting with shortest times for each group and including censored values. Censored patients are omitted from the point of omission to prevent them from influencing the proportion of surviving patients. The survival probability at a particular time, S(t), is given by

$$S(t) = p_1 \times p_2 \cdots p_t$$

where p_t is the proportion of all patients surviving past a certain time point.

A survival object can be compiled from the column of survival times and the variable indicating whether a patient is censored by the surv function. This survival object can then be interpreted by the survfit function to fit the Kaplan-Meier curves. A comparison of Kaplan-Meier curves can be achieved by the log-rank statistical test with the null hypothesis that the survival curves of the two populations do not differ.

2.4.2 Cox Proportional-Hazards Model

In contrast the hazard function describes the probability of hazard or event should the individual survived up to certain time point. The hazard function is important when considering the influence of covariates when comparing the survival of patient groups. Explanatory variables, or covariates, may be predictive of an outcome or a factor that might need adjusting to account for interactions between variables. The parameters for covariates are; age at diagnosis as under 65years, 65-74years and over 74years; percentage predicted FVC and FEV above 80, 60-80 and below 60; percentage predicted DL_{CO} groups comprise above 55 (mild disease), 36-55 (moderate disease) and below 35 (severe disease)[59]; and gender.

A forest plot is a useful approach to visualise the hazard ratios (HR) for all the covariates that were included in the Cox Proportional-Hazards model. The Cox proportional hazards model was built using the coxph function and visualised using the ggforest function. A hazard ratio of greater than 1 denotes an increased risk of death compared to that less than 1 inferring a reduced risk of death, should the patient be met by a specific condition.

3 Results chapter 1: Temporal features of longitudinal lung function data

3.1 Introduction

The chapter hypothesis is that there are temporal features within FVC time series of IPF patients. This chapter aims to identify the physiological parameters that impact the survival and clinical classification of IPF patients. In addition, the parameters of the temporal components (e.g. autocorrelation and seasonality) within the longitudinal lung function measurements of IPF patients are explored within this chapter. Finally, the variability of the IPF airway calibre is investigated using the deterministic method, DFA.

Idiopathic pulmonary fibrosis (IPF) is a progressive diffuse parenchymal lung disease with uncertain aetiology[3,4,145]. The urgency for an accurate diagnosis of IPF has increased with the development of two effective anti-fibrotic therapies, coupled with the limitations of traditional therapeutic combinations[3,11]. The diagnosis of IPF is dependent on a thorough clinical evaluation that is responsible for excluding alternate causes of disease, such as chronic hypersensitivity pneumonitis. Optimal accuracy for the diagnosis of various forms of pulmonary fibrosis, including IPF, requires a multi-disciplinary review with the inclusion of pulmonary, radiology and pathology teams[146].

Whilst the prognosis of IPF is poor, the trajectory of IPF disease progression varies from slow progression to acute deterioration and death[145,147] (Figure 1). In addition, patients can experience periods of acute respiratory decline in the absence of infection or other identifiable cause, referred to as an acute exacerbation[38,148]. IPF disease course has traditionally been assessed during clinical trials by measuring clinical endpoints every 3-4 months by trained study technicians in an office-based setting[65]. The main endpoints measured during IPF clinical trials include; change in forced vital capacity (FVC) as a primary endpoint[149] and, as secondary endpoints, symptom severity, quality of life and survival time[150]. Given that IPF clinical course can vary with time, increased measurement frequency of individual predictors ought to improve analytical precision and reduce sample size requirements of clinical trials[151] but require study subjects to return to the study centre where lung function technicians can perform procedures.

Whilst IPF management decisions can be adequately informed by FVC measurements every three to six months, increased frequency of observations has the potential to detect subtle changes in lung function before irreversible fibrotic changes afflict the patient. The use of daily

home spirometry to monitor IPF disease progression is attractive as it has the potential to overcome the resource implications and inconvenience to patients that accompany increased observation frequency. This has been addressed in a sub-study of the PROFILE (Prospective Observation of Fibrosis in the Lung Clinical Endpoints) cohort[64,76]. Whilst daily home-based FVC measurements tended to be lower than their hospital-based equivalents, the trends between the more frequent longitudinal time series mirror hospital-based observations[64].

The potential to control the quality of measurements, or respond to FVC decline or non-adherence can be addressed by employing bluetooth-enabled spirometers that transmit data via a secure encrypted connection[66]. Therefore the home monitoring experiences of IPF patients can be extended by replacing paper-based collection with real-time data transmission, thereby providing direct access to data for both patient and healthcare provider[66]. Real-time monitoring allows patients to be prompted in the event of either bothersome side-effects, an FVC decline of >10% over 3 consecutive days or failure to perform spirometry or record symptoms.

The feasibility of obtaining a high-frequency of FVC measurements using home-spirometry by IPF patients has been established by a number of studies[64–66]. High-frequency measurements will make positive contributions for the future development of IPF clinical trials by reducing trial numbers and helping to evaluate the efficacy of additional drugs with cohorts whose rate of FVC decline is dampened by pre-existing anti-fibrotic therapy. This is because high-frequency measurements offset the impact of biological and measurement variability within lung function assessments and permit more accurate longitudinal trajectory estimates of outcomes[152,153].

In contrast to IPF, Chronic Obstructive Pulmonary Disease (COPD) is a common respiratory condition characterised by airflow limitation, a consequence of pathologic changes to the lung parenchyma[154], airways[155] and pulmonary vasculature[156]. For COPD, the FEV_1/ FVC is useful for assessing airway obstruction in patients at the point of diagnosis, and longitudinal FEV_1 is informative for monitoring disease progression since FEV_1 has previously been shown to decline with time[130].

The ability to monitor progression of IPF and COPD by spirometry is important as it allows a non-invasive measurement of the patient's ability to breathe[157]. The clinical importance placed on the decline of an IPF patient's exercise capacity is because it is deemed a reflection of the architectural destruction of the lung as a consequence of fibrosis[64]. Therefore, accurate

interpretation of spirometry is essential in assessing how rapidly the patient's lung function is declining and whether the patient is suffering a complication. Timely identification of complications may allow early administration of therapy with favourable clinical outcomes. Indeed, maintaining the FVC of IPF patients close to their baseline levels seems to correspond with a better outlook for the wellbeing of those individuals[43,63,100].

Serial changes in pulmonary function tests (PFTs) over a 6 to 12 month period for IPF have a better prognostic value than baseline observations[44,45,149]. Understanding IPF disease progression from PFTs requires threshold values to define a significant decline from baseline, which are currently set at 10% for FVC and 15% for diffusing capacity of the lung for carbon monoxide[3]. Despite home disease monitoring being a feature of asthma self-management and detection of acute rejection for lung transplant recipients[158], the issue of how home disease monitoring can be optimised for the benefit of IPF patients is a challenge that has yet to be addressed.

An issue that still remains regarding home monitoring is the optimum frequency of FVC measurements and whether the additional data provided by daily spirometry can further reduce trial numbers. The next step for IPF home monitoring would involve the development of algorithms that can classify patients on the basis of the clinical form of IPF being experienced allowing therapy to be tailored to individual disease progression[64]. At the time of IPF diagnosis, how an individual's clinical course will proceed is unknown. This may be addressed promptly by statistical models that are able to interpret the features of FVC longitudinal time series and appropriately stratify a patients' clinical course.

The data generated by home-based spirometry requires the application of specialised techniques that can identify patterns in successive observations taken at equally spaced intervals. Time series methods are unique to generalised statistical approaches in that they account for internal structures (autocorrelation, trend or seasonal variation) that may be present within the data[159,160]. An important feature that needs to be considered during time series analysis whether there is dependence between various points within the time series and the size of the time interval at which this influence exists.

Short- and long-term dependence can be quantified by autocorrelation functions or detrended fluctuation analysis respectively (DFA)[161]. Signals generated by complex biological systems harbour fluctuations that exhibit long-range correlations [162], termed self-similarity[163]. DFA has

been used to demonstrate that irregular airflow is decreased in asthmatic patients[164], self-similarity in peak expiratory flow is indicative of increased risk of unstable airway function[121] and is able to distinguish between atopic and non-atopic asthma[165].

Healthy breathing dynamics reveal complex patterns of variation that result from environmental stimuli interacting with an array of immunological, mechanical and inflammatory components forming the respiratory system[112] with the purpose of optimising gas exchange[166]. An understanding of the non-linear behaviour underpinning the respiratory system has the potential to provide both insight into pathology and tools for clinical assessment[161].

DFA assesses self-similarity by comparing the extent of fluctuation about a trend within the physiologic time series using a number of different sized windows. The gradient, termed the scaling exponent, of a double log plot of window size versus fluctuation is indicative of long-term correlation such that a value of 0.5 translate to successive points are unrelated to each other. However, a scaling exponent of between 0.5 and 1 suggest that the related time series exhibits long-range correlations and that past events influence the current situation.

Autocorrelation is defined as the correlation of a particular signal with itself at various time intervals[167]. Correlation is a statistical method employed to understand the strength of the relationship between any given pair of variables. The evaluation of autocorrelation requires the comparison of the complete and partial autocorrelations depicted at various intervals of time or lags. If the autocorrelation function exhibits an exponential decay or falls below the level of significance at a certain lag, the two time-points can be deemed to be independent. For long-term memory processes, the dependence is stronger resulting in the autocorrelation function decaying in a power-like manner.

An important feature of time series data that needs to be evaluated prior to the assessment of autocorrelation is whether it is stationary. A stationary time series is a stochastic process whose properties don't change when shifted in time. Longitudinal forced vital capacity measurements of IPF patients can be anticipated to be non-stationary on the basis that the disease is characterised by irreversible decline in lung function, however slow this may be[3,100]. Nonetheless, any time series can be converted to stationary process by taking the first-order difference, achieved by replacing each value in a time series with the difference between it and

the previous value[160]. The original time series is referred to as an integrated process of order 0, whilst the first-order difference time series is referred to an integrated process of order 1.

Predicting the course of IPF from past trends within each patient FVC time series requires mathematical models that understand the temporal dynamics of the features present. Incorporating parameters that can form a general equation to model the FVC time series of each patient is complicated by the intrinsic variability of FVC, the influence of co-morbidities and dose interruptions. Furthermore, the rate of FVC decline during the current year predicts mortality, but not pulmonary function, in the subsequent following year[168]. Taken together, the ultimate model of fibrotic disease progression will not be linear and incorporate parameters that account for the influence of age, co-morbidities and type of treatment.

The hypothesis for this chapter is that there is information within the time series' of IPF lung function that reflect the disease progression within the patients' lung. Therefore, the aims of this chapter are to understand the physiological parameters that impact the survival and clinical classification of IPF patients. The identification of the temporal components (e.g. autocorrelation and trend) within the longitudinal lung function measurements is the second aim of this chapter as these components could influence how IPF is classified and disease progression monitored. Establishing the variability of the IPF airway calibre is the final aim of this chapter which could contribute additional information about the state of the lung.

3.2 Results

3.2.1 Patient Characteristics

50 patients were recruited as a sub-study of the PROFILE clinical trial (90% men, median±IQR; age, 66.81±10.89yrs; height, 1.8±0.13 m, weight 85.8±24.4kg, FVC 67.85±20.58% predicted and FEV_1 73±19.65% predicted). Detailed baseline characteristics are presented in Figure 3.1. The study was performed prior to the availability of anti-fibrotic therapy meaning that knowledge of disease decline did not influenced therapy. The majority of patients within the study did not smoke (former smokers; 68% or non-smokers; 28%) (Figure 3.1 (b)).

3.2.2 Forced vital capacity (FVC) time series

Ten patients were monitored at home for the entire 490-day study. For those patients remaining in the study, the median duration of the study was 151 days (range, 14-486 days), with thirty-eight subjects dying during the clinical trial follow-up and two discontinuing due to disease progression, intolerance to the procedure, or response to seeing their decline reflected in spirometry readings. The distribution of values within each patient time series varies with no correlation to baseline FVC values (Figure 3.2 (a)). In addition, patient compliance varies between individuals with no correlation to baseline FVC (Figure 3.2 (b)), with the mean proportion of daily observations completed during participation was 81.5±70.8%. The distribution of missing values varied on an intra- and inter-subject basis.

However, inspection of individual patient graphs of longitudinal FVC observations indicate IPF disease progression consistent with that previously described. Therefore, it is possible to utilise this information to classify patient disease behaviour. Within this dataset, there are patients who have experienced a rapid rate of disease progression associated with a projected annualised FVC decline exceeding 95% of baseline. In addition, there are four patients who has experienced an accelerated 14-day period of lung function decline corresponding to a loss of >10% FVC. The symptoms, signs and radiographic findings for this subject are consistent with an acute exacerbation.

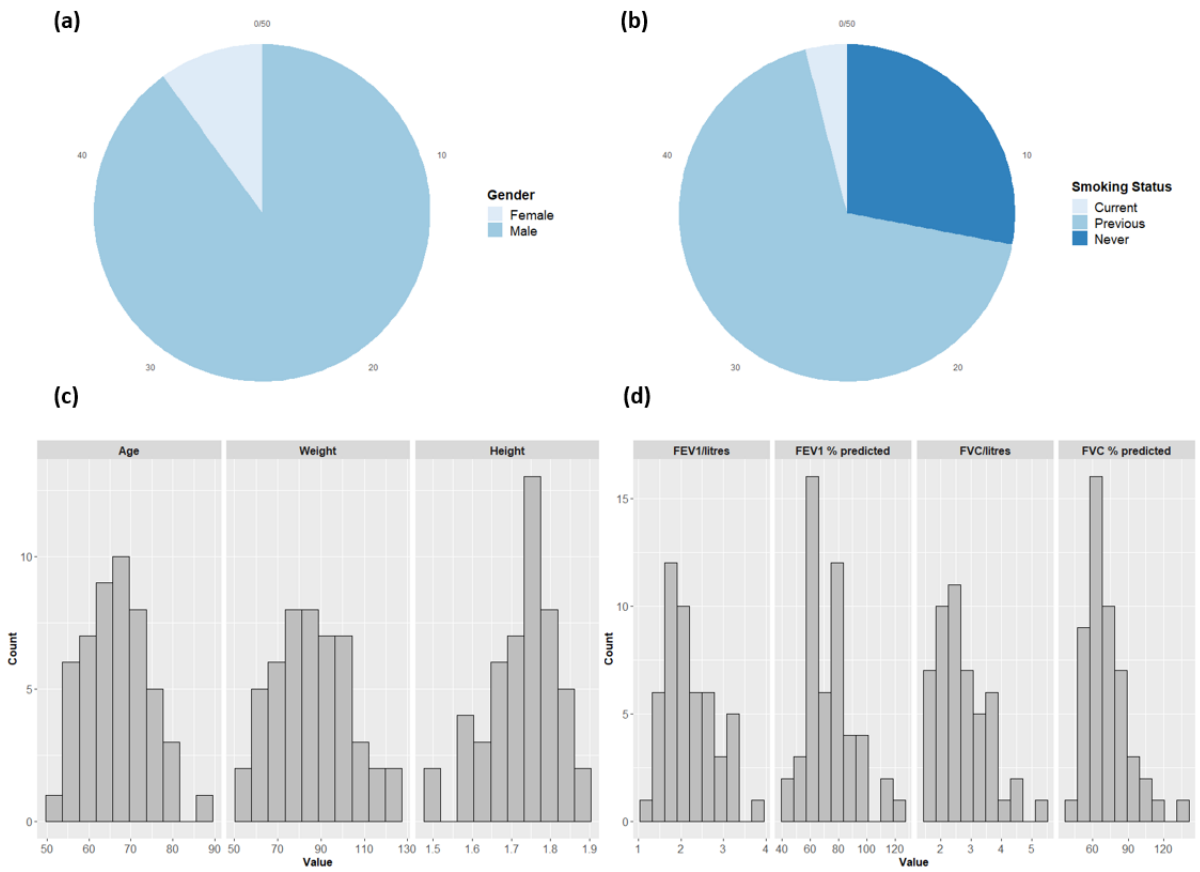


Figure 3-1 | ***Baseline characteristics of study participants, recruited as a subgroup of the PROFILE (Prospective Observation of Fibrosis in the Lung Clinical Endpoints) clinical trial.*** *50 patients were recruited from the interstitial lung disease unit at the Royal Brompton Hospital, London, the majority of whom were men (a). The majority of patients have previously smoked, with a minority currently smoking (b). Physical attributes at recruitment such as age, weight and height of the cohort that may influence interpretation of spirometry are indicated (c). Clinical assessment and full hospital-based lung function testing were undertaken at baseline (d).*

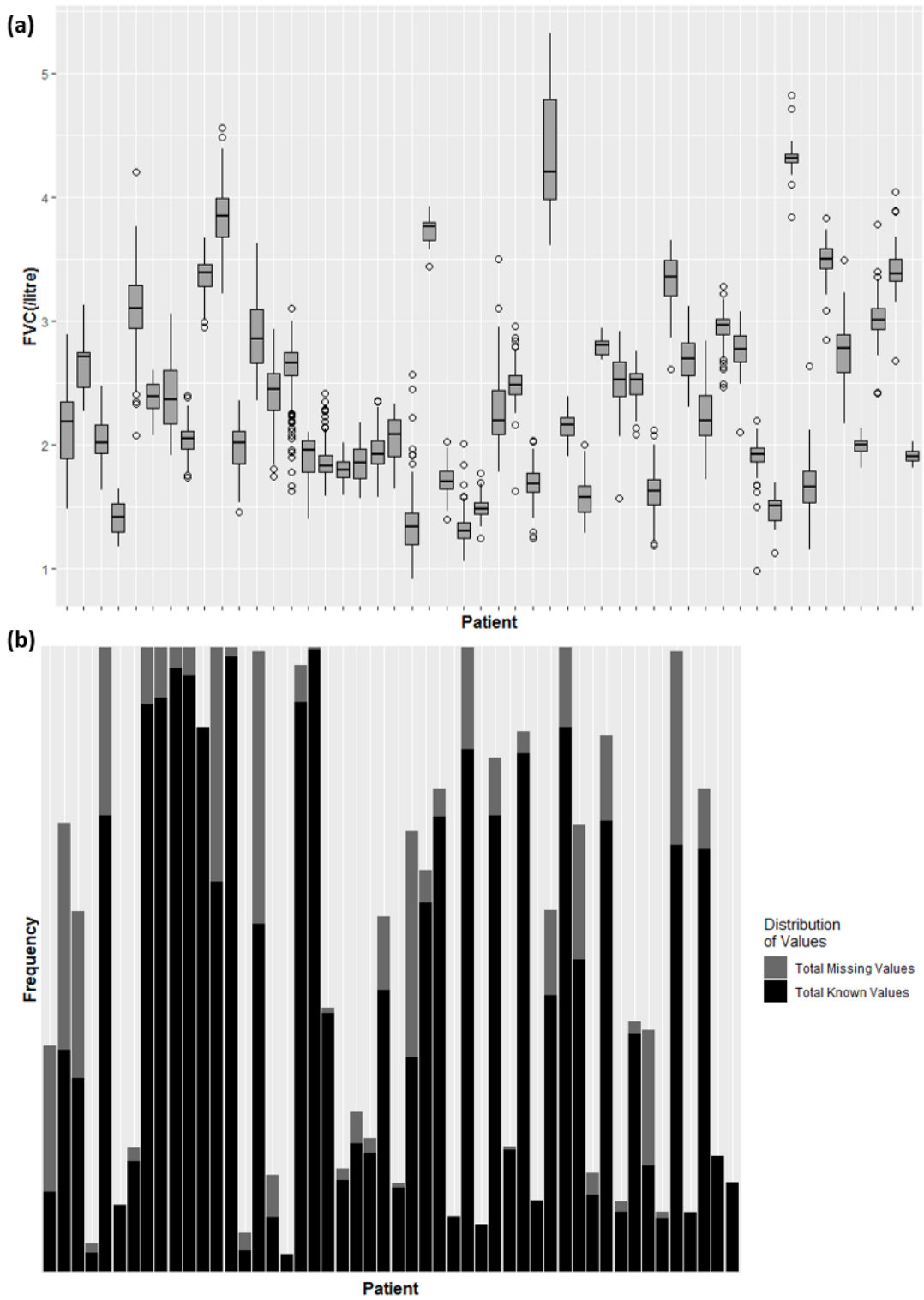


Figure 3-2 | ***Distribution of Forced Vital Capacity (FVC) measurements.*** *Patients were provided with a portable hand-held Micro spirometer (CareFusion, Kent, United Kingdom). Each Micro spirometer provides a digital read out of* FEV_1 *and FVC measurements to an accuracy ±3% in a range of 0.1-9.99L FVC. Participants were asked to perform a single spirometry reading at approximately the same time of day and record the FVC measurement in a dedicated diary. The distribution of FVC measurements vary between individuals with outliers were classified as observations outside the 95% confidence interval (a). Each individual time series were characterised by different amounts of missing values (b).*

3.2.3 Survival analysis

There may be a number of individual clinical variables that may predict survival of patients diagnosed with IPF. The rate of change of FVC when considered as a continuous variable is predictive of subsequent mortality at 3, 6 and 12 months when adjusting for age, sex and baseline FVC[64]. In addition, the consideration of several baseline features may be pertinent in evaluating whether a patient is subject to an increased risk of mortality[59]. Understanding the risk of the smoking status on the survival of IPF patients requires an increased number of current smokers to be able to detect any statistical differences that may exist between the groups (Figure 3.3).

Subjects whose baseline FVC predicted was between 50-80% (HR, 0.070; CI, 0.0063-0.77) or >80% (HR, 0.098; CI, 0.0059-1.64) had a lower risk of death compared to those FVC predicted was <50%. Patients who were diagnosed with a greater proportion of their maximum FVC, potentially had more lung function capacity to lose during the course of the study. Baseline FEV1 percent predicted does not seem to correlate with the risk of mortality from IPF as those who have the most expiratory volume, >80% (HR, 0.490; CI, 0.1118-2.15) and 60-80% (HR, 1.102; CI, 0.4145-2.93), are comparable those with the least FEV_1 predicted <60%. This is consistent with IPF being a restrictive disease and serial FVC changes over 12 months being predictive of subsequent survival[44,45]. Whilst FEV1 and FVC measurements decline in proportion to IPF disease progression (Figure 3.4b), FVC has been identified as a reliable, reproducible marker of IPF disease progression and linked to mortality[56,169]. Hence, the correlation between FEV1 and fibrotic disease progression may not be 1:1.

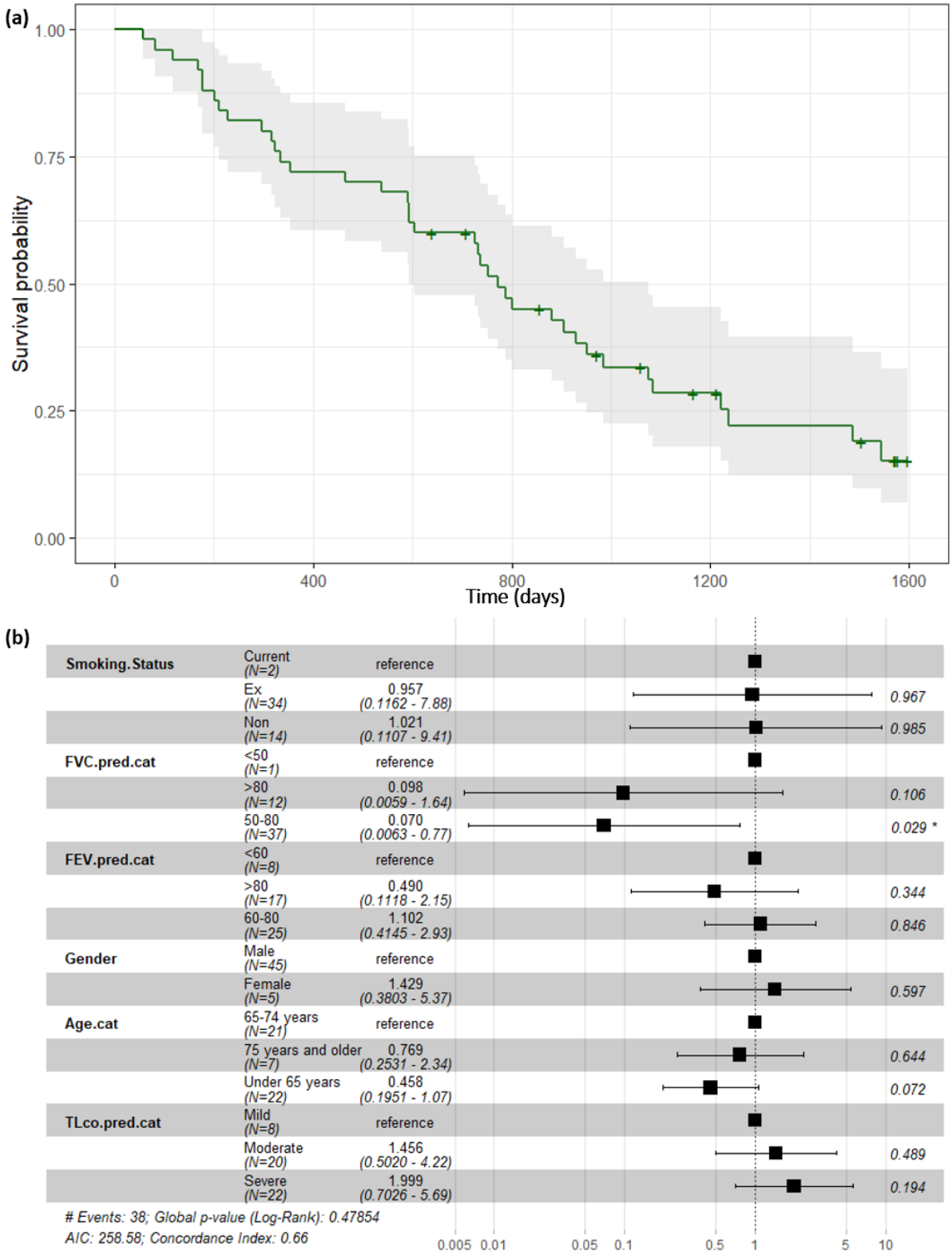


Figure 3-3 | **Comparison of baseline physiological parameters and their impact on survival.** Kaplein-Meier survival curve for patients within this cohort indicated that patients have been censored by 4½ years (a). Multiple variables were associated with altered risk of mortality in idiopathic pulmonary fibrosis (IPF) patients. For the Tlco % predicted category, mild disease is defined as >55%, moderate disease corresponds to 35-55% and severe disease is < 35%[59](b).

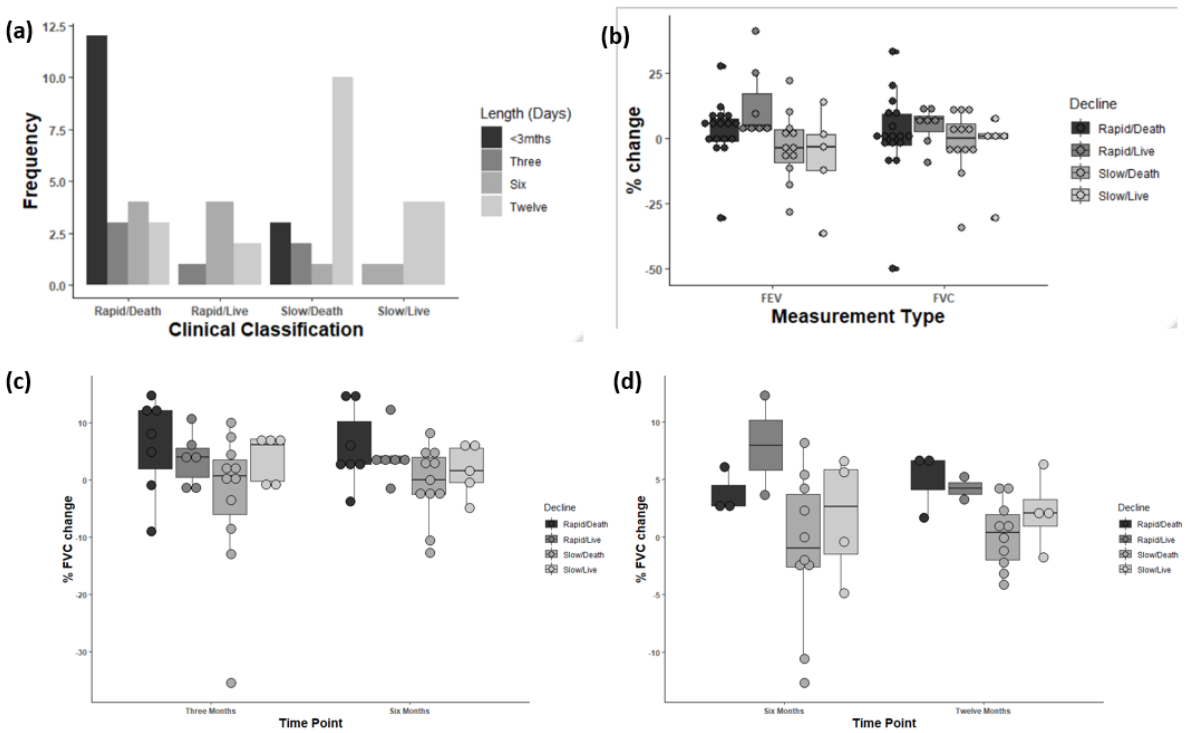


Figure 3-4 | ***Overview of changes in pulmonary lung function categorised by clinical classification.*** *Patients classified with rapid progressive disease were most likely to die within 3 months (p-value = 0.0152; Chi-square test) (a). However, individuals with rapid progressive disease were equally likely to have FVC time series that were 3, 6 or 12 months long (p-value = 0.47; Chi-square test) (a). In contrast, patients with slow progressive disease tended to have FVC time series at least 12 months long (p-value = 0.0002; Chi-square test) (a). Changes in Forced Expiratory Volume in 1 second (FEV_1) and FVC were identical for the period that both pulmonary function tests were performed for each patient (b). The decline in FVC was consistent for each clinical group over 3- (c) and 6-monthly (d) intervals.*

3.2.4 Modelling patterns of behaviour

Monitoring IPF patients is motivated by the need to assess disease progression and inform the administration of therapy more appropriately to subject needs. A categorical change in FVC over 1-year is strongly predictive of IPF mortality when a threshold value of 10% baseline is used[44,45,64]. Linear regression has been utilised to calculate the rate of change in FVC between two observations with baseline as a point reference[64]. Whether a linear regression model based on FVC time series can be informative of IPF disease progression has yet to be determined (Figure 3.5). The evaluation of the linear regression statistical model forms the basis for assessing statistical models generated by alternative approaches.

Linearity, unequal error variances and outlier should be assessed for every patient time series. The rate of FVC decline over three- and six-monthly intervals seem to be constant for patients' classified with slow or rapid disease progression, thereby implying linear trends (Figure 3.4 c and d). The acute exacerbation time series is observed when a IPF patient experiences a short period of unexplained, accelerated FVC decline (Figure 3.5). A linear regression model was performed on the entire time series for a patient who suffered an acute exacerbation (Figure 3.5a). Given that the residuals do not bounce randomly nor form a horizontal band around '0' in a residual versus fitted plot (Figure 3.5b) suggest that the linear assumption is inappropriate, and the variance of error terms is unequal.

The normal Q-Q plot is graphical tool that helps us to determine whether the data forms a normal distribution. If the data comes from a normal distribution, then the normal Q-Q plot should form a straight line, however the extremities of the FVC time series from the acute exacerbation patient curve off (Figure 3.5e). This suggests that the extreme values are more extreme than expected if the data were from a normal distribution. A patient who has experienced an acute exacerbation would be expected to have distinct distributions in their FVC observations given that an acute exacerbation is defined as abrupt, irreversible decline in lung function[170].

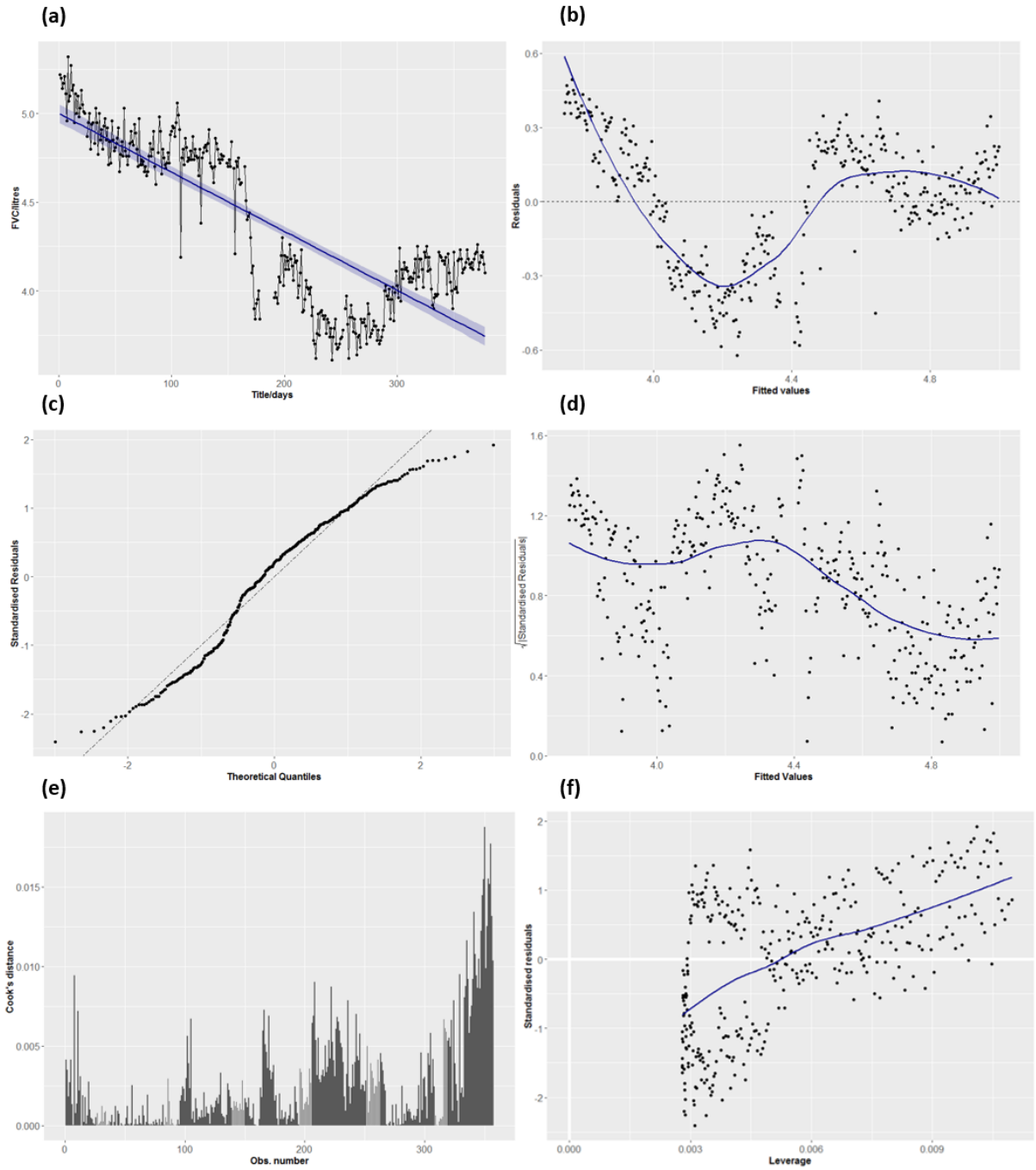


Figure 3-5 | ***Assessment of linear regression models of an individual experiencing an acute exacerbation [PRO1056].*** *Daily FVC measurements for a patient with differing periods of progressive disease to which a linear regression is applied with its confidence interval (a). The residuals vs fitted values is used to detect non-linearity, unequal error variances, and outliers (b). Normal Q-Q plots associated with this linear regression model allows us to determine whether the error terms in the linear regression model is normally distributed (c). The scale-location plot helps us to evaluate whether the residuals are spread evenly within the predictor range, however the line deviates from the horizontal suggesting that there is non-uniform variance in the residuals, i.e heteroscedastic (d). An assessment of influential values on the linear regression model can be made with Cook's distance (e) and residuals vs leverage plots (f). Taken together these two plots suggest that the measurements between 350+ are the most influential and that removing them would have a big impact of the model.*

The spread-location plot helps to assess whether there is equal variance (homoscedasticity) amongst the residuals (Figure 3.5d). Given that the line is not horizontal and the points aren't equally spread underline the unequal variance (heteroscedasticity) observed earlier. Observing unequal variance for linear regression statistical model of a patient who has experienced acute exacerbation is reassuring given that their disease progression is non-linear in that they have experienced an acute period of accelerated decline.

The final consideration when assessing the quality of a statistical model is the assessment of which observations are the most influential within the regression analysis, as evaluated by the Cook's distance (Figure 3.5e) and residual versus leverage (Figure 3.5f) graphs. Cook's distance is metric of the influence of an outlier as a function of residual size and leverage. The uneven distribution in the Cook's distance (Figure 3.5e) and the increase in the standardised residuals (Figure 3.5f) indicate that the more recent observations have the greatest impact on the statistical model being generated and should they be excluded the model would be very different.

Taken together, the linear regression model does not incorporate the features within the FVC time series, particularly a patient who has undergone an acute exacerbation. The general equation for linear regression is given by

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

where y_i is the FVC value, x_i is time and ε_i represents the error or residuals of the statistical model. With residuals of unequal variance and observations of differing influence, the associated statistical model can be considered a poor representation of the loss of lung function because information is lost within the residual terms.

3.2.5 Rolling statistics of univariate time series

The benefit of home-based spirometry is the accumulation of higher resolution information regarding patient disease progression compared to hospital-based spirometry[64,65]. A common approach to model univariate time series is to employ a moving-average model, which can be visualised using rolling statistics (Figure 3.6a). As seen for time series for a patient who experienced an acute exacerbation, the moving-average process is an improvement to a linear regression model (Figure 3.6a). Firstly, the rolling mean provides a better impression that there

may be different periods of rate of change in lung function (Figure 3.6a). Secondly, the rolling window used to calculate the standard deviation gives an impression of whether variance is constant and careful interpretation can indicate which values may be outliers (mean ± SD 4.36 ±0.072L). The size of three measurements for each window is a comprise between detecting different periods of decline in lung function and the variability between measurements. This value should be adjusted to account for differences in the statistical parameters of patients.

Whilst acute exacerbations in IPF are common with a 2-year frequency of 9.6%[31], it is unusual to measure patient lung function during the event. The FVC time series before and after the acute deterioration suggest that the lung function measurements exhibit a binomial distribution (Figure 3.5b). This is consistent with the failure of this time series to adhere to a normal distribution (Figure 3.3c). The peaks within the binomial distribution corresponds to the extended period of observations before and after the period of deterioration. Acute exacerbations may be the clinical event that prompts patient diagnosis with IPF[31] and so the presence of a binomial distribution cannot be expected to be the sole feature of identification.

3.2.6 The autocorrelation function

IPF is defined by an irreversible decline in lung function regardless of its clinical course[4] meaning that FVC time series are characterised by a negative gradient (Figure 3.7a). The original time series contrasts with a modified version that has been made stationary by differencing (Figure 3.7b). This represents an important processing step prior to evaluating whether serial correlation or autocorrelation is existing in the subject time series and requires repeating until stationarity is achieved. An inspection of both the complete and partial autocorrelation functions (Figure 3.6c & Figure 3.7d) indicate that contemporary FVC observations are influenced by measurements of time periods of a day and no greater. All the time series' within the study population required one round of differencing and indicated that any given FVC value was correlated with the corresponding measurement from the previous day.

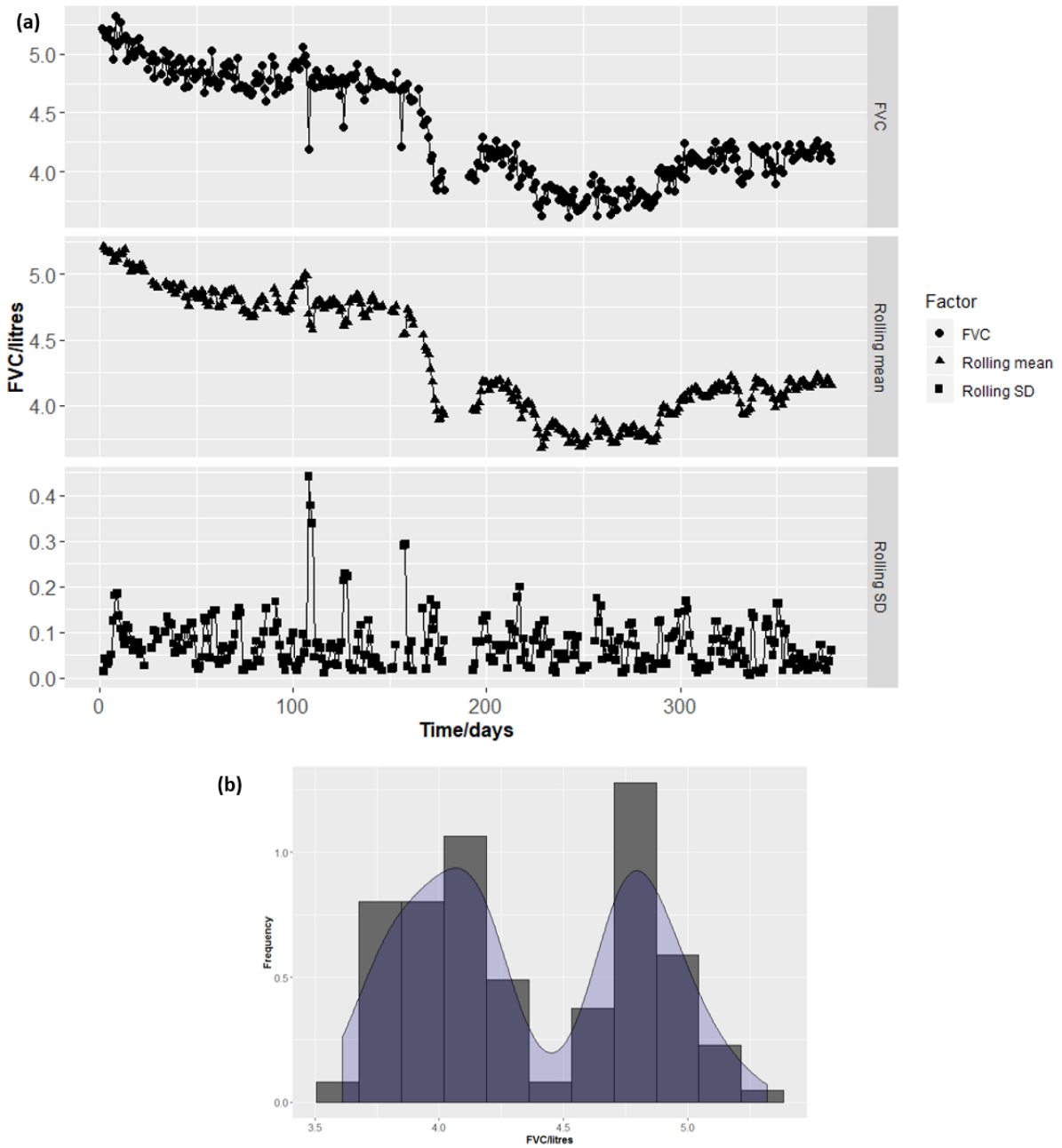


Figure 3-6 | ***Rolling statistics of time series of an example patient experiencing an acute exacerbation [PRO1056].*** *Daily FVC measurements for a patient with differing periods of progressive disease. A rolling mean was applied with a window size of three measurements. The size of window will need to balance the variation within the time series due to random noise, the distribution of missing values and the sensitivity to changes. The rolling mean helps to smoothen out outliers and identify the underlying trend within the time series (a). The time series of a patient who has experienced an acute exacerbation exhibits a binomial distribution (b).*

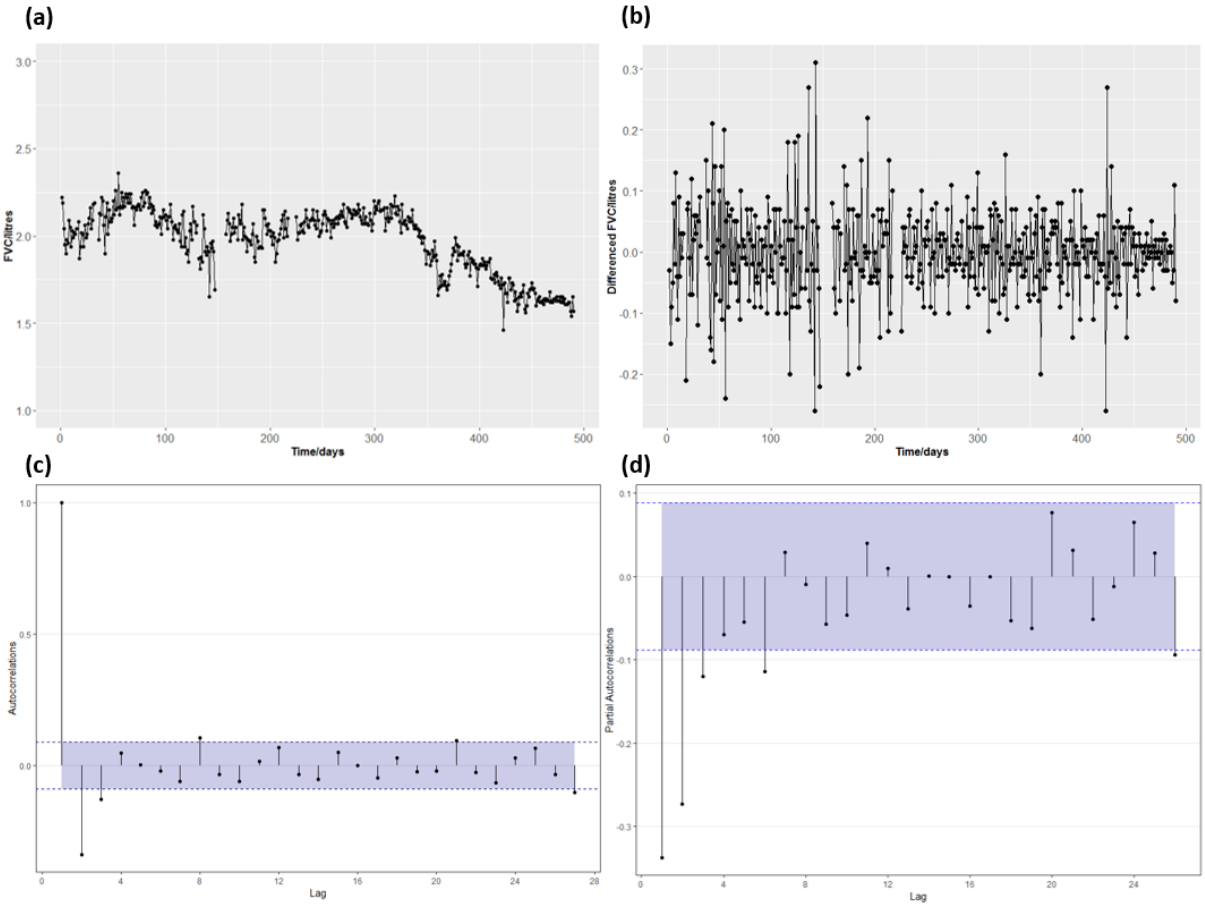


Figure 3-7 | ***Relationship of measurements with previous values for individual examples [PRO1014].*** *Daily FVC measurements for a patient with steady progressive disease as recorded (a) is compared to the differenced time series to make the time series stationary (b). This is an important processing step to assess the degree of correlation, y-axis, with past values of the time series separated by different periods of the lag on the x-axis for autocorrelation (c) and partial autocorrelation (d). The shaded region represents the level of significance within which values can be attributed to chance. The patterns for autocorrelation and partial autocorrelation are important to determine which time periods are correlated with each other.*

3.2.7 Detrended Fluctuation analysis

The self-similarity features within the FVC time series from the IPF cohort was compared to 31 COPD patients recruited from the observational cohort (28 males, mean ± SD age 65.9±9.5). Being an obstructive disease, COPD disease progression is evaluated by changes in FEV_1 and FVC (FEV_1 1.08±0.38 litres; FVC 2.58±0.67) (Figure 3.8a). The survival of COPD patients is not affected by whether individuals were classified as having moderate, severe or very severe on the basis of FEV_1 predicted (Figure 3.8b). Patients with COPD had an increased risk of death if they are male (HR 3.51; CI 0.68-18.3) and had a lower weight (HR 1.36; CI 0.58-3.2) (Figure 3.8c). Whilst FVC does not influence the risk of death (HR 1.02; CI 0.24-4.4), which is consistent with COPD characterised by airway limitation, unusually the risk of death is reduced with a lower FEV_1 (HR 0.69; CI 0.0.22-2.2) (Figure 3.8c).

Detrended fluctuation analysis (DFA) measures self-similarity in the form of a scaling exponent. The variability a time series' is determined by measuring the fluctuation about linear regressions performed within different sized windows (Figure 3.9 a-c). Variability informs the scaling exponent which is calculated by the gradient of a double log plot of fluctuation and window size (Figure 3.9d). The mean scaling exponents of the COPD and IPF cohorts are (0.82±0.15 and 0.85±0.16 respectively; mean ±SD) and are statistically similar (p-value = 0.44; student's t-test) (Figure 3.9e). Baseline physiological parameters of IPF and COPF differed only in terms of weight, BMI, FEV_1 and FEV_1 % predicted.

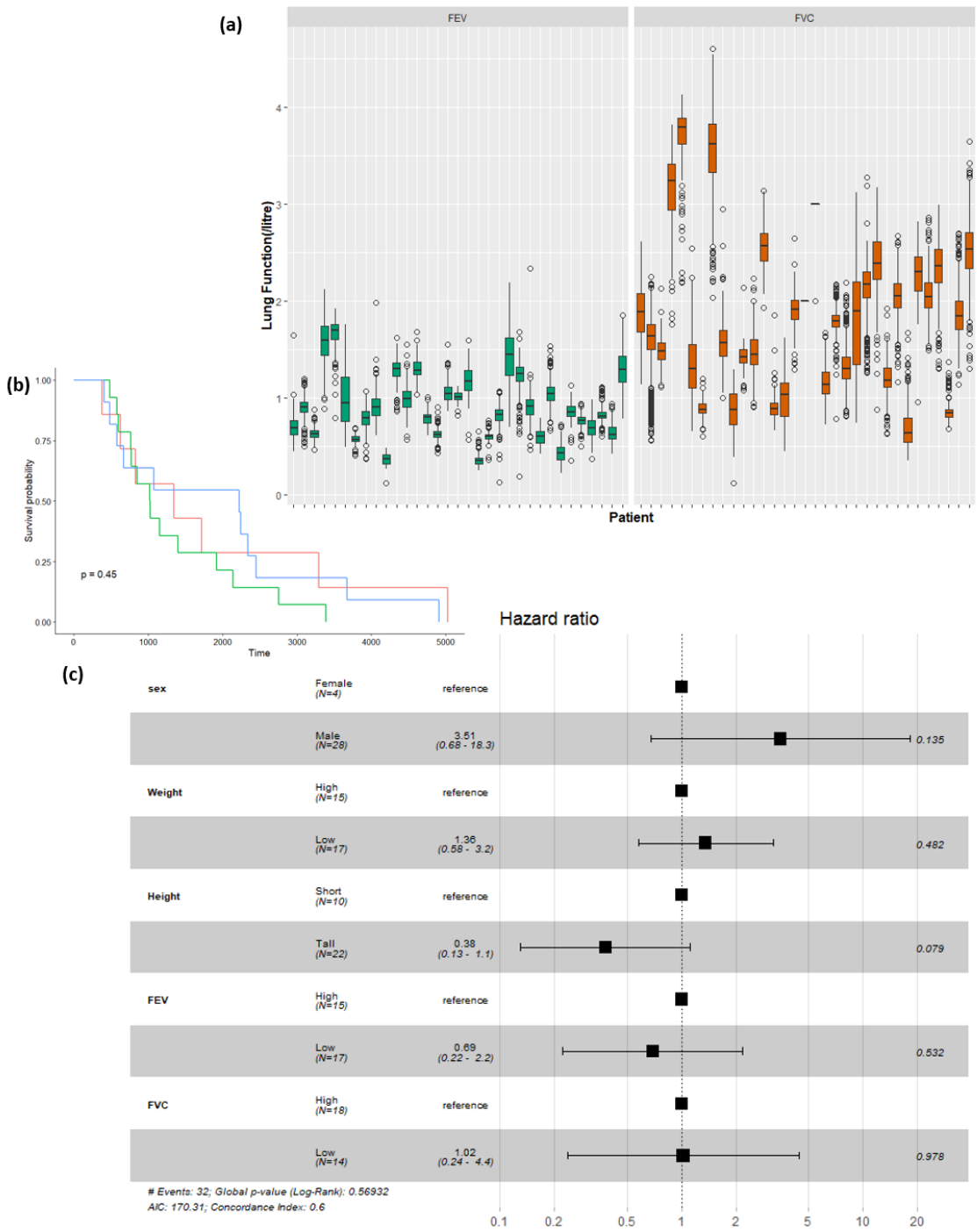


Figure 3-8 | ***Distribution of lung function and comparison of baseline physiological parameters on survival of patients diagnosed with Chronic Obstructive Pulmonary Disease (COPD).*** *32 COPD patients were enrolled into the London COPD Cohort and contributed a minimum of 1 year of data. The distribution of Forced Expiratory Volume in 1 second (FEV_1) and FVC measurements were obtained using a Micro spirometer with observations exceeding the 95% confidence interval indicated (a). Kaplein-Meier survival curve illustrates that survival probability on the basis of moderate (red, 50% < FEV_1 < 80%), severe (Green, 30% < FEV_1 < 50%) and very severe (Blue, FEV_1 < 50%) disease are indistinguishable. Cox hazard analysis compares the influence of multiple baseline variables on survival (c).*

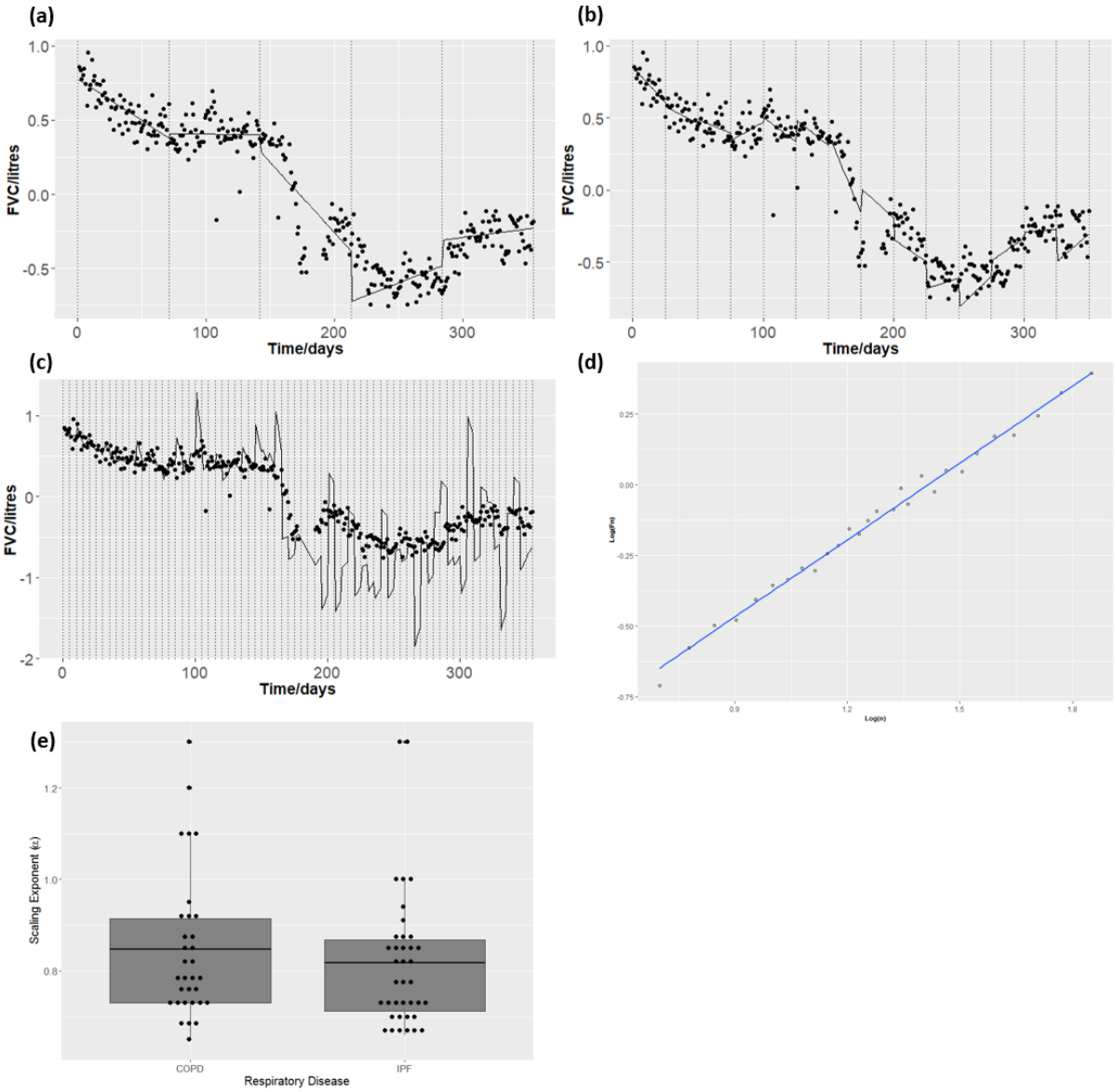


Figure 3-9 | ***Comparison of the variability in the forced vital capacity (FVC) time series of IPF and COPD patients using Detrended Fluctuation Analysis (DFA).*** *Calculation of the scaling exponent, first, requires segmenting the entire patient time series into different sized windows and determining the fluctuation of each window size (a-c). Secondly, the scaling exponent is determined from the gradient of a double log plot of fluctuation against window size (d). The variability of the FVC time series' of 38 IPF and 32 COPD patients enrolled in the PROFILE study and the London COPD cohort respectively for a minimum period of 75 days were compared using the scaling exponent (α). A comparison of the mean scaling exponents of IPF and COPD patients (0.82±0.15 and 0.85±0.16 respectively; mean ±SD) indicate that there is no statistical difference between IPF and COPD (p-value = 0.44; student's t-test) (e).*

Variable	IPF (mean)	COPD (mean)	*p-value*
Height	1.73	1.72	0.9 (NS)
Weight	84.1	73.8	<0.001
BMI	28	24.8	<0.01
FVC	2.76	2.57	0.32
Age	67.5	65.8	0.444
FEV_1	2.24	1.08	0 (signif)
FEV_1 & pred	76.2	2.96	0 (signif)

Table 1 | Physiological parameters at baseline were compared between the IPF and COPD cohorts using the Student's t-test to determine whether distributions were statistically different

3.3 Discussion

The results of this chapter are that several physiological parameters are associated with altered risk of mortality. Changes in lung function correlated risk of mortality and disease progression in IPF patients at 3-monthly intervals. Modelling patterns of IPF disease behaviour must account for the potential non-linearity of the clinical course, i.e. occurrence of an acute exacerbation. Sequential values within the FVC time series are influenced by measurements observed the day before as demonstrated by autocorrelation. Finally, the variability within the FVC time series of IPF and COPD are similar suggesting that both diseases may have common features and defects at play.

IPF is a disease that afflicts patients of older age with individuals of a median age of 66 years at time of diagnosis[3,4]. Inspection of the change of mortality with age are broadly consistent with the assertion that older age confers a poorer prognosis[4] given that those subjects <65 years (HR, 0.458; CI, 0.1951-1.07) at the point of diagnosis have a lower risk of mortality that those who were 65-74 years. The group who were diagnosed with IPF after 75 years (HR, 0.769; CI, 0.2531-2.34) also tended to have a lower risk of mortality compared to the reference 65-74 year group, which may be a consequence of a large confidence interval and too few number of subjects (7 subjects, 14%).

Baseline FVC predicted measurements differs between IPF patients consistent with variable asymptomatic periods prior to diagnosis[4]. Individuals with asymptomatic, early lung fibrosis have been identified within families of individuals with familial pulmonary fibrosis[171], however it is unclear as to how individuals with subclinical IPF should be followed and managed[4]. The frequency and distribution of missing values varied within and between patients, which can be addressed by real-time spirometry[65] and longitudinal data analysis.

Imputation strategies can be employed to account for the non-random nature of missing data and serial trends in lung function before or after missing data[172]. Whilst most missing data in nature are informative, it was surprising that the amount of data did not correlate with clinical classification. However, other symptoms or co-morbidities may correlate with the distribution of missing data within the time series. Whether missing data are biased across treatment groups will have to be addressed in future clinical trials that consider individual or combination therapy.

Participants of the sub-study were required to be recorded spirometry readings in a dedicated diary[64] meaning that patient compliance was evaluated retrospectively. Real-time spirometry has the potential to improve patient compliance by prompting patients to perform their spirometry manoeuvre[65]. In addition, real-time monitoring can enhance the quality of FVC measurements used for analysis by evaluating the manoeuvre during its performance and inform the participant whether a repeated observation is required[65].

Rolling statistics determine trends in the time series based on past information, in particular the average and variance. Whilst rolling statistics offer improvement to a linear regression approach (Figure 3.5a) for modelling changes in the lung function of IPF patients, rolling statistics fail to account for changes that may influence future measures of disease progression. Moving averages would be limited for individuals who experience multiple acute exacerbations as they exhibit high volatility[173]. In addition, determining the optimal window size used to calculate the mean and standard deviation will be difficult to determine on a prospective basis without additional information of individual patient clinical course[173].

The FVC time series for the participant who experienced an acute exacerbation adheres to a binomial distribution (Figure 3.6b), with the two distributions corresponding to pre- and post-acute deterioration. Identifying patients who have experienced an acute exacerbation is significant to understanding disease progression and assessing risk of mortality in a clinical setting[174]. Determining the distribution of lung function observations within a time series is a retrospective approach, meaning its value is limited to identifying patients after the clinically significant event and any therapeutic period.

Longitudinal lung function data are characterised by repeated measurements through time which may be correlated with itself as a function of a delay. The autocorrelation, or Pearson correlation, at a time interval of 1 is significant within the IPF patients meaning that FVC measurements can be predicted using a linear regression between the present and immediately day before values. An important outcome of serial correlation between daily FVC measurements is that it favours daily home spirometry for the home monitoring of IPF disease progression. Time series analysis of daily home spirometry would benefit from an additional parameter autoregression which can contribute to generating a statistical model of IPF disease progression.

The temporal history of longitudinal FVC measurements for IPF patients can be investigated by comparing the scaling exponent with a COPD cohort. Self-similarity for peak expiratory flow (PEF) for this COPD cohort has been characterised previously and compared to other respiratory diseases[121,122,175]. The algorithm generated during this study produces a statistically similar scaling exponent value for PEF in the COPD cohort to previously reported values[122].

Given that the scaling exponent of COPD and IPF FVC time series are above 0.5 suggest that long-range correlations exist and that observations are related to previous values. This implies that the respiratory system of IPF and COPD have a memory meaning, which is consistent with the previously described autocorrelation. COPD and IPF FVC time series have statistically similar scaling exponents that these two respiratory diseases may have common features and defects that influence daily variability in FVC, that are different from asthma.

This chapter can be concluded with a partial acceptance of its hypothesis that there is clinically important information within the longitudinal measurements of physiological parameters relating to IPF disease progression. The design of future clinical trials will require sequential measurements of physiological parameters for IPF patients experiencing steady disease progression and acute exacerbations to allow the development of an optimisation strategy for the calculation of temporal components in real-time. It is yet to be determined how measures of these temporal components change with therapeutic interventions and whether they used as indicators of therapeutic efficacy.

4 Results chapter 2: Disease classification model assembly

4.1 Introduction

The chapter hypothesis is that information within the FVC time series of IPF patients can be used to classify IPF disease progression. Therefore, the principal aim of this chapter is to compare and validate the modelling approaches to best identify changes in the clinical course of IPF.

A central dogma of physiology is the principle that all cells, tissues and organs endeavour to maintain a constant steady-state[162,176]. Contemporary signal processing technologies that obtain continuous time series data from physiological processes including heart rate (HR) and blood pressure (BP) are characterised by non-stationarity and non-linearity, even if the physiological system is described as being at "steady-state" conditions[162,176]. The maintenance of a healthy, optimal environment is the outcome of a continual, dynamic, bi-directional interaction between multiple neural, hormonal and mechanical control systems that operate at a local and global levels[162,176].

Data presented in the first results chapter suggest that the time series of IPF patients contain "hidden information" that can be characterised by applying specialised concepts and techniques from statistical physics[162]. Application of techniques like detrended fluctuation analysis (DFA) imply that long-range correlations exist with the IPF lung that result from the regulation of the complex processes that take place within it[112,162]. If these fluctuations were absent across a range of time scales, the ability of the lung to functionally respond to unpredictable stimuli and stresses would be restricted[162]. Scale-invariance, as measured by the scaling exponent in DFA, relate contemporary measured variables to recently preceding observations and fluctuations of the more distant past.

Whilst DFA has been invaluable in establishing that physiological variability exists within the time series of IPF patients, this has been established retrospectively using time series' with a minimum 75 observations. Therefore, a potential limitation of the DFA technique is its utility for monitoring IPF disease progression on an ongoing process given that clinically significant events like acute exacerbations occur within a 30-day period[38]. An alternative approach to assessing scale invariability is to utilise a visual tool that can be applied to relatively short time periods, termed to Poincaré plot[177].

Poincaré plot analysis has been employed to determine the hidden patterns within the heart rate variability (HRV) thereby helping to assess the heart's adaptability to altering physiological conditions of patients with heart failure, norepinephrine infusion and post-myocardial infarction[178–181]. Patterns formed by physiological variability can be visualised by generating a two-dimensional plot of consecutive points of the corresponding physiological time series (i.e. *lag-1* plot). Visual inspection of the resulting distribution is guided by the standard descriptors, SD1 and SD2, described previously in linear statistics[182]. SD1 and SD2 describe the short- and long-term variability of the physiological time series by measuring the standard deviations perpendicular and parallel to the line of identity (*i.e.* $y = x$ axis), respectively[182].

All statistical analyses are affected to some degree by the presence of missing data. Participants with IPF may discontinue daily spirometry because of their disease progression, technical difficulties or patient distress from witnessing their lung function decline[183]. Enhanced technology including blinding, data storage and remote data access may improve patient compliance by promoting patient participation and monitoring manoeuvre quality[183]. Simple univariate time series algorithms include the last observation carried forward, next observation carried backward and the Kalman filter. The Kalman filter is favoured as an imputation method because it utilises a number of measurements to understand the statistical noise within the time series when generating estimates[160].

A popular statistical method amongst medical researchers for analysing and forecasting time series data is the AutoRegressive Integrated Moving Average (ARIMA) model[101,102,106,184]. The general ARIMA model is a class of statistical model that incorporates three processes involved in univariate time series, autoregression (AR), integration (I) and moving average (MA). Each process for an ARIMA model is specified by integer values representing their order, denoted (p,d,q) respectively.

An autoregression model describes the dependence of current and future values on past time points with the order referring to the lag observations. In addition, the integration process defines the extent of differencing required to make the time series stationary and the moving average denotes the dependency of an observation on the residual error obtained from a moving average. The suitability of the ARIMA statistical model is underlined by the dependency of FVC measurements on observations from the previous day, as shown previously. Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) allows the quality of a set of statistical models to be compared with each other.

The lung can be considered to be a complex system of immunological, mechanical and inflammatory components that fluctuate non-linearly in response to environmental stimuli[112]. Disease progression for some IPF patients may abruptly change, termed an acute exacerbation, resulting in a period of accelerated, irreversible loss in lung function[4]. This clinically significant event may be likened to similar sudden changes, referred to as tipping points, for complex dynamical systems in medicine, global finance or the earth system[124]. Generic early-warning signals may indicate whether tipping points are approaching for asthma attacks, shifts in oceanic circulation and systemic market crashes[185–187].

The tipping points of complex systems refer to a family of catastrophic bifurcations in mathematical models which themselves are simplifications of the stability properties of complex systems[188]. Given that the particular mechanism underpinning a tipping point within a complex system is unknown, recent research activity has sought to find generic indicators of the proximity of a tipping point. In particular, a phenomenon referred to as critical slowing down describes the ability of a complex system, including living systems, to return to equilibrium when subject to small perturbations[188–190]. Therefore, critical slowing down will increase in the proximity of a tipping point because the complex system has a reduced ability to recover from changes. In the context of longitudinal FVC measurements, critical slowing down will correspond with increased variance and temporal autocorrelation due to greater fluctuations[124,191].

The hypothesis of this chapter is that the temporal information within FVC time series' of IPF patients can be utilised to classify IPF disease progression. Therefore, the main aim is to compare and validate the modelling approaches that classify the disease progression of IPF. These modelling approaches will need to account for the variability of lung function between patients and the unknown clinical course at diagnosis.

4.2 Methods

4.2.1 Treatment of outliers and imputation of missing observations

Linear statistics were applied to the original time series to determine the mean and standard deviation. Outliers were classified on the basis of observations falling outside the range specified by the mean ± standard deviation. Measurements above this range were replaced with the sum of the mean and standard deviation. Missing values within time series were imputed using the Kalman filter function provided by the imputeTS package and implemented in the R statistical computing environment[192,193].

4.2.2 Poincare Plot

The Poincare plot is a geometrical demonstration of the time series in a Cartesian plane constructed by pairing values of the time series with measurements immediately before, i.e. lag-1[194]. Based on this construction, a sequence of equal observations will fall along the line of identity, which denotes the 45° diagonal. Assessment of variability within each time series is made by fitting an ellipse to the scattergram, which can be quantified using SD1 and SD2 statistics. Short-term variability, as measured by SD1, is defined by the width of the ellipse or spread of point perpendicular to the line of identity. However long-term variability, denoted by SD2, corresponds to the length of the ellipse or spread of points parallel to the line of identity.

Quantification of the dispersion measured perpendicular and parallel by SD1 and SD2 respectively are related to two basic statistical measures, standard deviation of RR interval (SDRR), and standard deviation of the successive difference of RR interval (SDSD)[195]. Therefore, SD1 and SD2 can be calculated by[195]:

$$SD1^2 = \frac{1}{2}SDSD^2$$
$$= \gamma_{RR}(0) - \gamma_{RR}(1)$$

$$SD2^2 = 2SDRR^2 - \frac{1}{2}SDSD^2$$
$$= \gamma_{RR}(0) + \gamma_{RR}(1) - \overline{RR^2}$$

where $\gamma_{RR}(0)$ and $\gamma_{RR}(1)$ are the autocorrelation functions for the lag-0 and lag-1 intervals of the time series and $\overline{RR}$ is the mean of the time series.

4.2.3 Detection of early-warning signal

The gradual progression of chronic diseases such as IPF may be interrupted by a sudden health deterioration. Disease progression can be considered to occur in three states: a normal, pre-disease and a disease state[196–198]. In the context of IPF, the disease is under control meaning that dynamically the system is highly resilient and robust to perturbations or injury. The pre-disease state is situated at the limit of the disease state, just prior to the transition point, and is characterised by low resilience to perturbations due to its dynamical structure. During the pre-disease state, the respiratory system can be regarded as reversibly sensitive to external stimuli but can potentially collapse through to a disease state when subject to a small perturbation.

The diseased state represents a deteriorated period of high resilience and robustness to external stimuli meaning that the achieving a normal state even with treatment may not be an attainable goal. For many complex diseases, the ability to identify the pre-disease disease state is crucial to prevent acute deterioration but is difficult given that this state shares many characteristics with the normal state[196]. Distinguishing the pre-disease and disease states relies on the identification of dynamical network biomarkers (DNB) which are the first network of genes formed during critical transitions and are related to causal genes in a disease network[196].

The identification of the DNB relies on monitoring fluctuations in gene expression of molecules within and outside the network[197]. Concentrations of members in the DNB will fluctuate more in the pre-disease state whilst behaving collectively together. In order to develop a signal from changes in gene expression to form the following composite index was proposed to evaluate the nonlinear dynamics of the system[197]:

$$\mathrm{I} = \frac{\mathrm{SD}_d \cdot \mathrm{PCC}_d}{\mathrm{PCC}_o}$$

where SD_d is the standard deviation of the dominant group, PCC_d is the average Pearson's correlation coefficient of the leading group and PCC_o is the average Pearson's correlation coefficient of the leading group and others.

The composite index compares variance, as measured by the Pearson's correlation coefficient, between members within DNB and those outside to determine whether these members form a unique group. This principle can be generalised to a univariate time series to compare the

variance of the most contemporary observations with those at baseline. Measurements of FVC during an acute exacerbation should be correlated better with themselves than with observations during the pre- or post-deterioration. Therefore, the composite index can provide an ongoing comparison of the Pearson's correlation coefficient (PCC_d) of the most contemporary FVC observation with that of every observation since baseline (PCC_o).

The numerical value of the composite index is unique to each individual patient as it accounts for the fluctuation observed within each FVC time series. Interpretation of the composite index prior to and during an acute exacerbation can be described as consisting of two sections, each of which is described by two linear regression functions characterised by different slopes. The challenge is determining the point of separation between the two regression functions, which is not known *a prior*[143]. For n pairs of observations (x_i,y_i), $i = 1,\ldots, n$ and assuming that xi are ordered $x_1 \leq x_2 \leq \ldots \leq x_n$ to maintain generality. The relationship between the x and y values can be described by two linear regression functions of the form

$$y_1 = a_1 + b_1 \cdot x_i \quad x_i \leq x_0$$
$$y_2 = a_2 + b_2 \cdot x_i \quad x_i > x_0$$

Where (x_0,y_0) is the coordinates of the join point, a_1 and a_2 represent the intercepts of the first and second regression lines respectively and b_1 and b_2 represent slopes of the first and second regression lines. Estimation of these parameters is based on the method of least squares where the optimal values of the aforementioned parameters correspond to the minimal residual sum of squares from the algorithm. For the first iteration of the algorithm, the first three observations of the data set are collated to estimate the parameters of the first regression function with the remaining observations used for fitting the second. Each subsequent iteration involves shifting an observation to the first part of the data until a range of join points are sampled[143].

An alternative method to interpreting the numerical value of the composite index is proposed in the form of the additive model. A maximum threshold is calculated on a continual basis using the sum of the mean and standard deviation from values within the composite index. Three consecutive composite values that exceed the maximum threshold result in change in the binary output to '1'.

The evaluation of changes in the composite index and the ability of the piecewise and additive model at interpreting them was assessed using simulated spirometric data. The trajectory of

simulated measurements was obtained from the logistic function that was reflected and translated in the y-axis. Variance was introduced by generating random number based on a normal distribution with a median standard deviation of 0.13L (range, 0.04-0.38L)[64].

4.2.4 Autoregressive integrated moving average (ARIMA) model

The Box-Jenkins approach to ARIMA modelling of time series was adopted to take advantage of the associations between sequentially lagged relationships that exist within time series data[199]. ARIMA models are designed to describe, in mathematical terms, how variables vary with time and are formulated using three empirically driven phases: identification, estimation and diagnostic testing[200]. The identification phase involves ascertaining the presence of patterns within the time series, manipulation of the data to achieve stationarity and identification of potential models. An automated approach to selecting the optimal parameters, in terms of fit and parsimony, for each ARIMA model was implemented with the aid of the auto.arima function from the forecast package[201].

The parameters for that define a standard ARIMA model are the order of autoregression (p), the degree of difference (d) and the order of moving average (q). During the search for a suitable model for the order constraints provided within the function, the Akaike and Bayesian information criterion (AIC & BIC), are used to assess the quality of each model relative to the others generated. These two estimators evaluate the amount of information lost in balancing the goodness-of-fit and simplicity of the model. Hence, the model parameters that correspond to the least AIC & BIC values infer the model of the highest quality[202].

The estimation phase involves the 'R' statistical computing environment generating statistical estimates, obtained by the conditional least squares method, for each parameter in the model[103,200]. Diagnosis of the model aims to assess how well the model fits and entails checking that there are no significant autocorrelations amongst the residuals using the Ljung-Box test[103,200].

4.3 Results

4.3.1 Treatment of outliers and imputation of missing observations

FVC measurements were retrospectively adjusted on the basis of exceeding the sum of the mean and standard deviation of the patient time series (Figure 4.1). Outliers were identified in the minority of observations with, at most, ten observations recorded into the diary by the patient being adjusted. Adjustments tended to localise around the trend of the time series as a whole and tended to either be under- or over- estimate local trends within the time series. The Kalman filter utilises the maximum likelihood method to allow variable periods of missing measurements to be imputed[102] (Figure 4.1). Whilst observations tended to be overestimated, successive imputed measurements tended to follow the trends of the time series (Figure 4.1e,f).

4.3.2 Short- and long-term variability of lung function time series

Poincare plots allow a visual representation of the nature of the fluctuations within FVC time series. The ellipse-fitting technique and its quantification by the Poincare descriptors, SD1 and SD2, underpin the variability between patients (Figure 4.2). Variation between patients with regards to their fluctuations may be due to technical and biological sources. Biological variation primarily originates from the pattern of disease progression, with particular consideration to whether a patient experiences an acute exacerbation. The SD1/SD2 ratio represents the ratio between short- and long-term variability which has had mixed results in characterising complex dynamic behaviours[194,203].

Whether Poincare descriptors characterise the complex dynamic behaviours underpinning IPF disease progression was sought by comparing these parameters between clinical groups of disease progression (Figure 4.3). Of the 28 patients who experienced rapid disease progression, 7 individuals survived the entire length of the study. However, 21 subjects endured slow disease progression, of whom 16 died during the course of the study. The SD1 and SD2 parameters did not statistically distinguish between the clinical groups relating to disease progression (Figure 4.3b). Whilst the SD1/SD2 ratio, particularly between rapid/death and slow/death, was approaching statistical difference (Figure 4.3b). The study was under-powered requiring 45 individuals per group to ascertain whether a statistical difference exists.

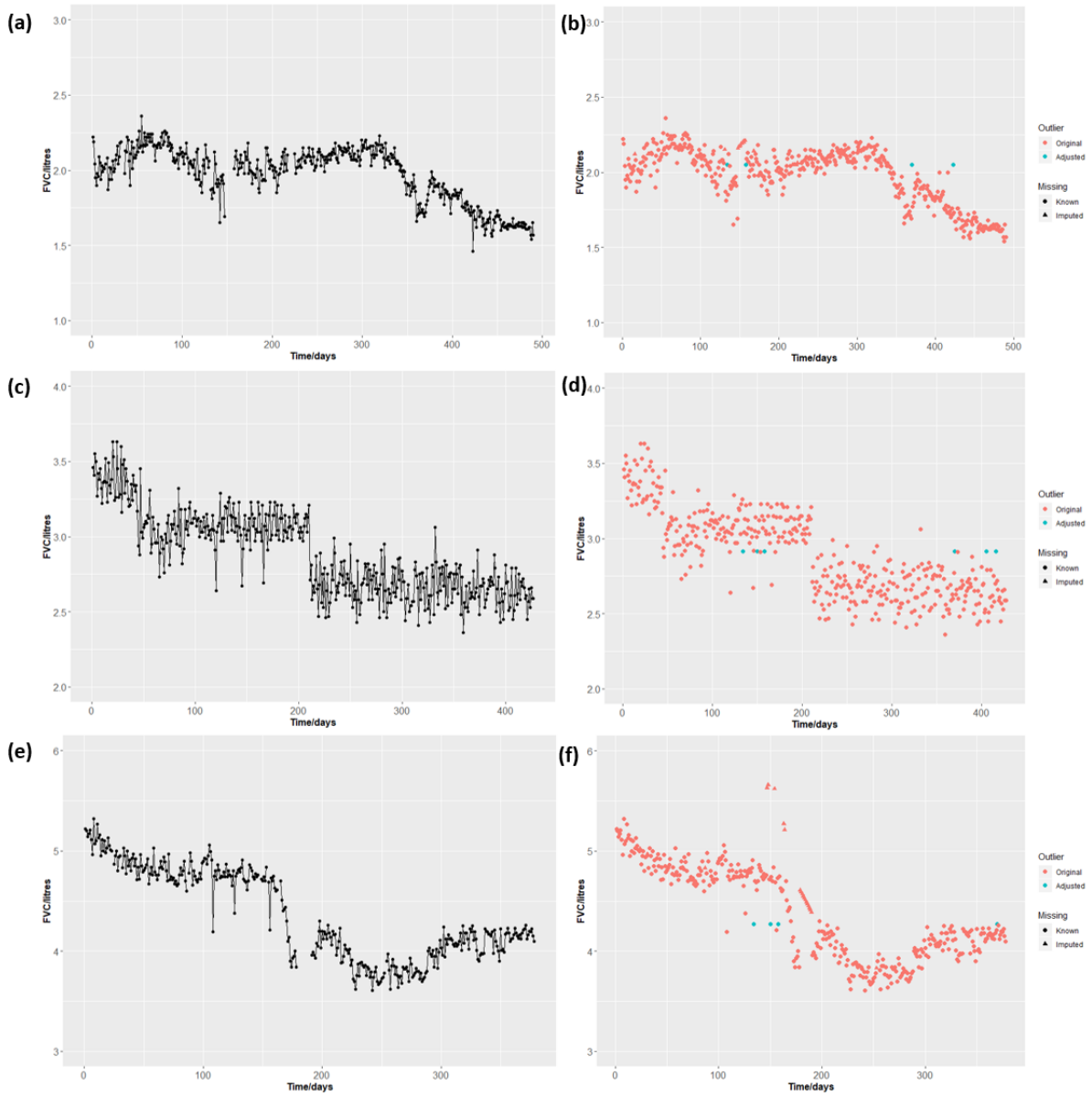
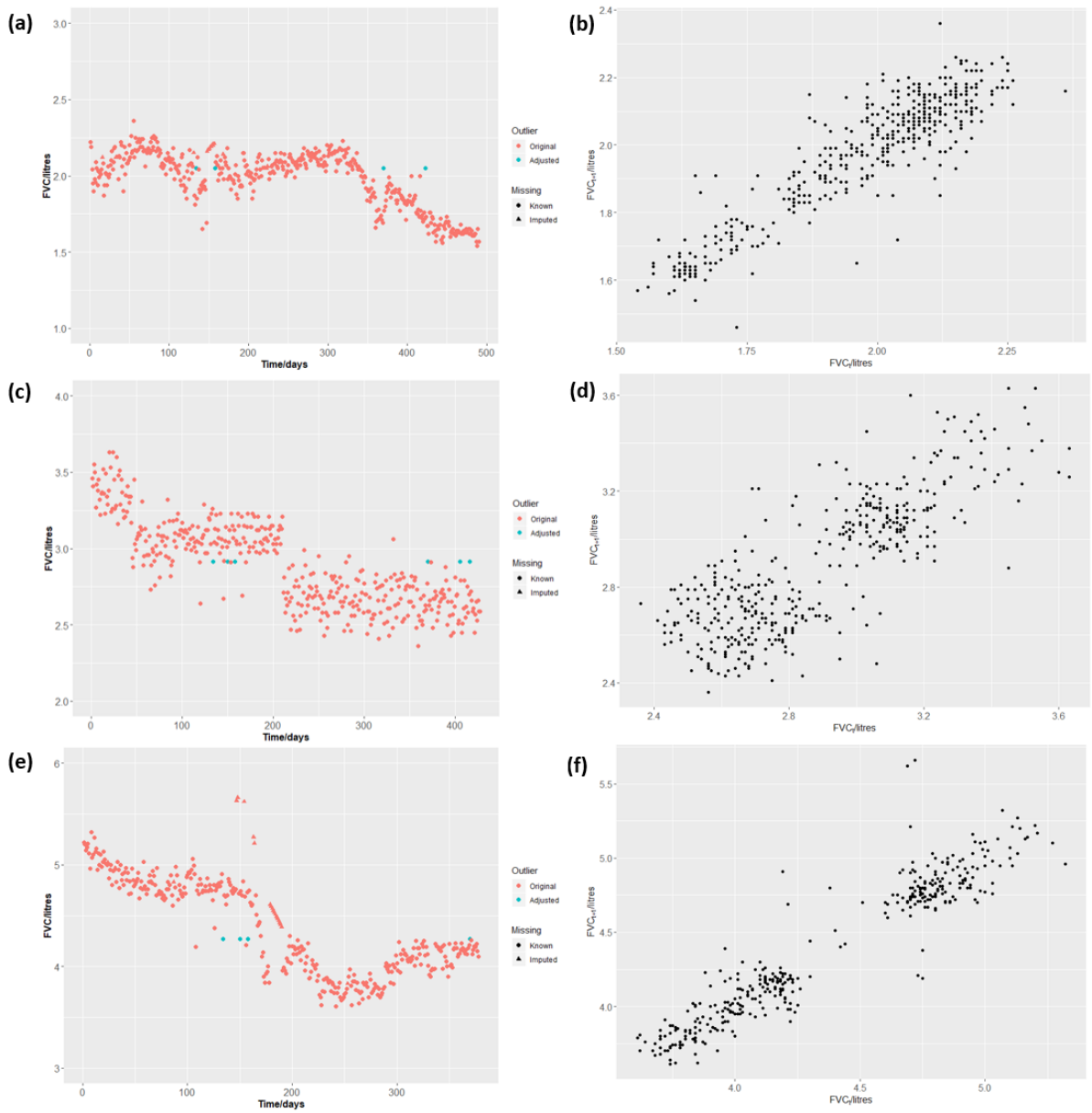


Figure 4-1 | ***Statistical treatment of outlier and missing values within the Forced Vital Capacity (FVC) time series for a subgroup of participants within the PROFILE (Prospective Observation of Fibrosis in the Lung Clinical Endpoints) clinical trial.*** *Outliers were retrospectively classified as exceeding the 95% confidence interval and where replaced with the maximum 95% confidence value for the corresponding time point. Missing values were retrospectively imputed using the Kalman filter. Original time series (a,c,e) are compared with their corresponding adjusted time series (b,d,e) for three selected patients.*



(g)

	Patient 1	Patient 2	Patient 3
Average FVC / litre	1.98	2.88	4.36
SD1	0.05	0.11	0.10
SD2	1.18	1.74	2.70
SD1/SD2	0.05	0.06	0.04

*Figure 4-2| **Poincaré plots for a selection of time series' for participants.** Longitudinal Forced Vital Capacity (FVC) time series for three patient's with adjusted outliers and imputed missing values are shown in (a,c,e) form the basis of standard Poincaré plots (b,d,f). The Poincaré descriptors, SD1, SD2 and the SD1/SD2 ratio, were calculated from each individual Poincaré plot (g). SD1 and SD2 are standard descriptors of linear statistics and characterise the deviation along and perpendicular to the line of identity, respectively. The SD1/SD2 ratio may play an important role in analysing data when non-linear behaviour distinguishes between health and disease.*

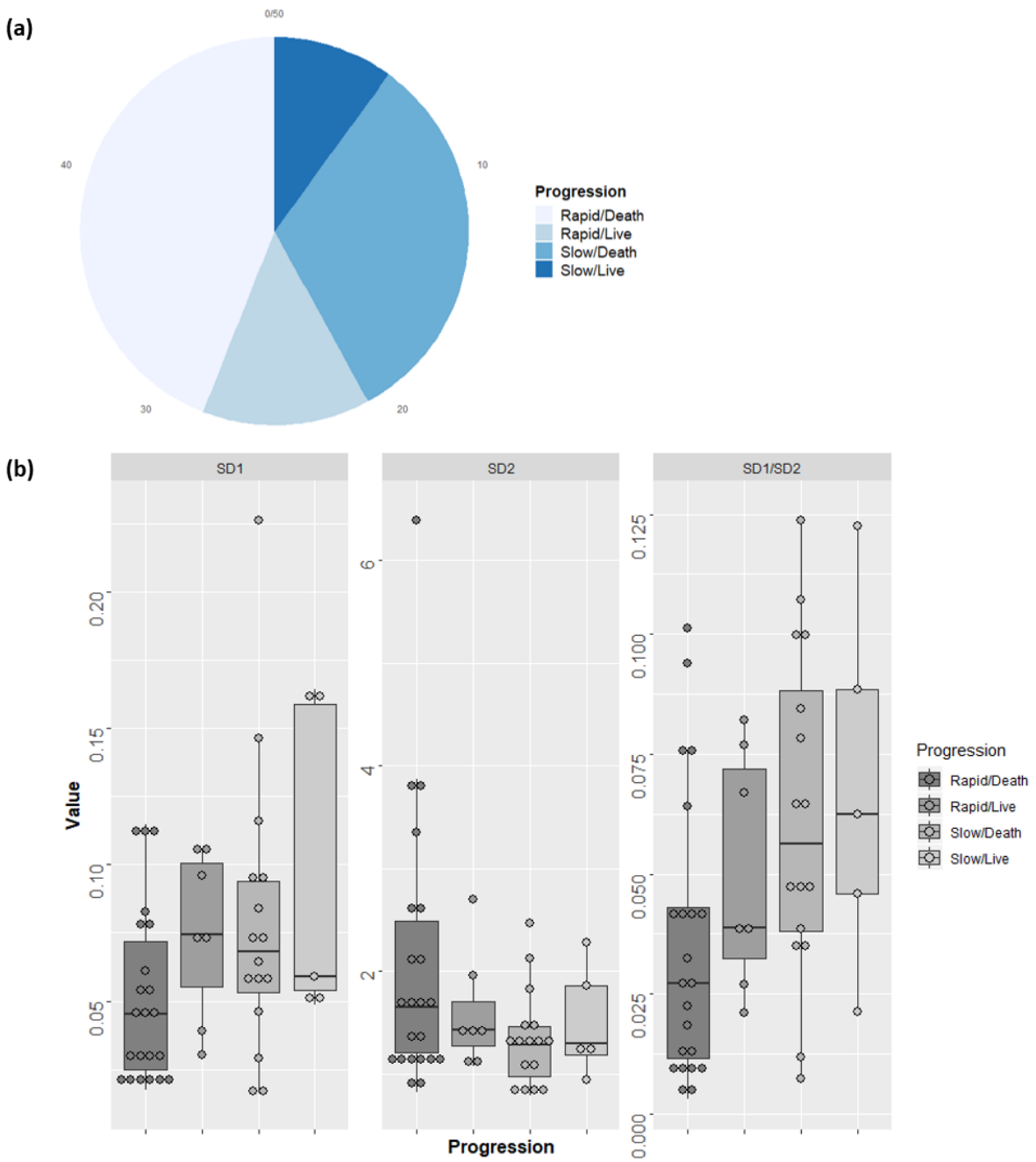


Figure 4-3 | ***Poincare descriptors for the daily spirometry subgroup of the PROFILE cohort.*** *Clinical classification of patient disease progression for the daily spirometry subgroup (a). The threshold for disease progression on the FVC continuous variable is defined as a 10% decline of baseline over 1-year. At the time of censorship, of the patients with slow-progressive disease 5 were alive compared to 16 who had died. In contrast, from the proportion of participants who were classified with rapid-progressive disease 7 participants were alive whilst 22 passed. A one-way ANOVA test suggests that there are significant differences between the disease progression at less than 0.05 significance level for the SD1/SD2 ratio but not for SD1 and SD2. A Tukey multiple pairwise-comparison suggests that a significant difference between the slow/death and rapid/death groups with an adjusted p-value of 0.06. Power analysis assuming an effect size of 0.25 suggest that each group would need 45 individuals to detect a statistically significant difference at the 0.05 confidence level.*

4.3.3 Autoregressive integrated moving average (ARIMA) model

Exploratory data analysis of FVC time series, in particular relating to autocorrelation, suggest that longitudinal lung function measurements of IPF patients are non-stationary and contemporary measurements are correlated to previous observations within a time interval of a day. The most parsimonious ARIMA models were evaluated for each FVC time series and every combination of parameter (p,d,q) up to the order of 5 using an automated algorithm within the auto.arima function[201]. The disease progression of twenty-two patients were modelled by an ARIMA(1,1,1) model and the remaining represented by an ARIMA(1,1,0) model, where each model corresponded to the lowest AIC and BIC of those generated (Figure 4.4a,c,e).

Further diagnostic checking of each ARIMA model involved residual analysis, specifically measurement of the autocorrelation within the residuals of the selected ARIMA model using the Ljung-Box test. The p-value of the Ljung-Box statistic for both the ARIMA(1,1,1) (mean±standard deviation Ljung-Box statistic 0.88±1.99; p-value 0.68±0.31) and ARIMA(1,1,0) (Ljung-Box statistic 0.85±2.45; p-value 0.70±0.27) models exceed the 0.05 level of significance indicating that the residuals are independent and that these models account for all trends present in the data.

4.3.4 Detection of early-warning signal

The value of the composite index remained constant (Figure 4.4) whilst the gradient of lung function did not change. At the beginning of an acute period of deterioration the value of the composite index rose, and varied as the period of deterioration continued. After the period of acute deterioration, the rate of lung function decline returned to original parameters albeit at new, lower absolute level. The initial value of the composite index is dependent on when the spirometry begins, and given that the time of diagnosis varies between patients there will be an issue interpreting changes in composite index.

The piecewise and additive models provide two distinct methods for interpreting the composite on an individual time series basis. An additive model relies on three sequential composite index values exceeding the sum of the mean and standard deviation to generate an exacerbation signal of '1' for the duration of the acute deterioration (Figure 4.5c,d). However, the piecewise model

uses the angle between two adjacent linear regressions to indicate the beginning and end of an exacerbation, which is highlighted by different signs (Figure 4.5e,f).

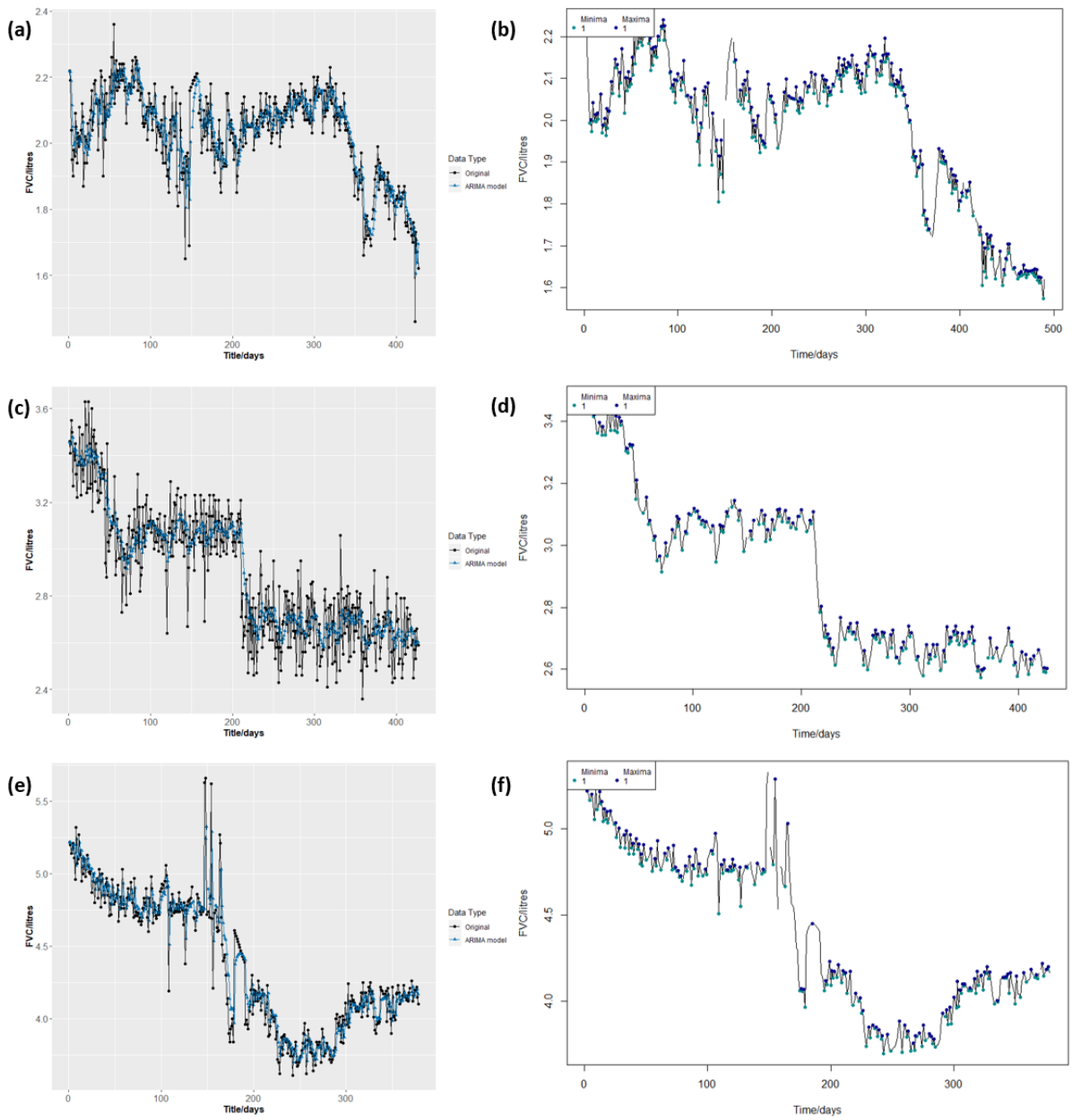


Figure 4-4 | ***Autoregressive Integrated Moving Average (ARIMA) statistical model for a selection of patients' longitudinal FVC measurements.*** *For each time series, the most parsimonious ARIMA model and its corresponding (p,d,q) parameters were chosen by selecting the model with the lowest Akaike and Bayesian Information Criterion for a range of systematically generated ARIMA models (a,c,e). Local maximum and minimum values were calculated from ARIMA fitted values for corresponding time series (b,d,f).*

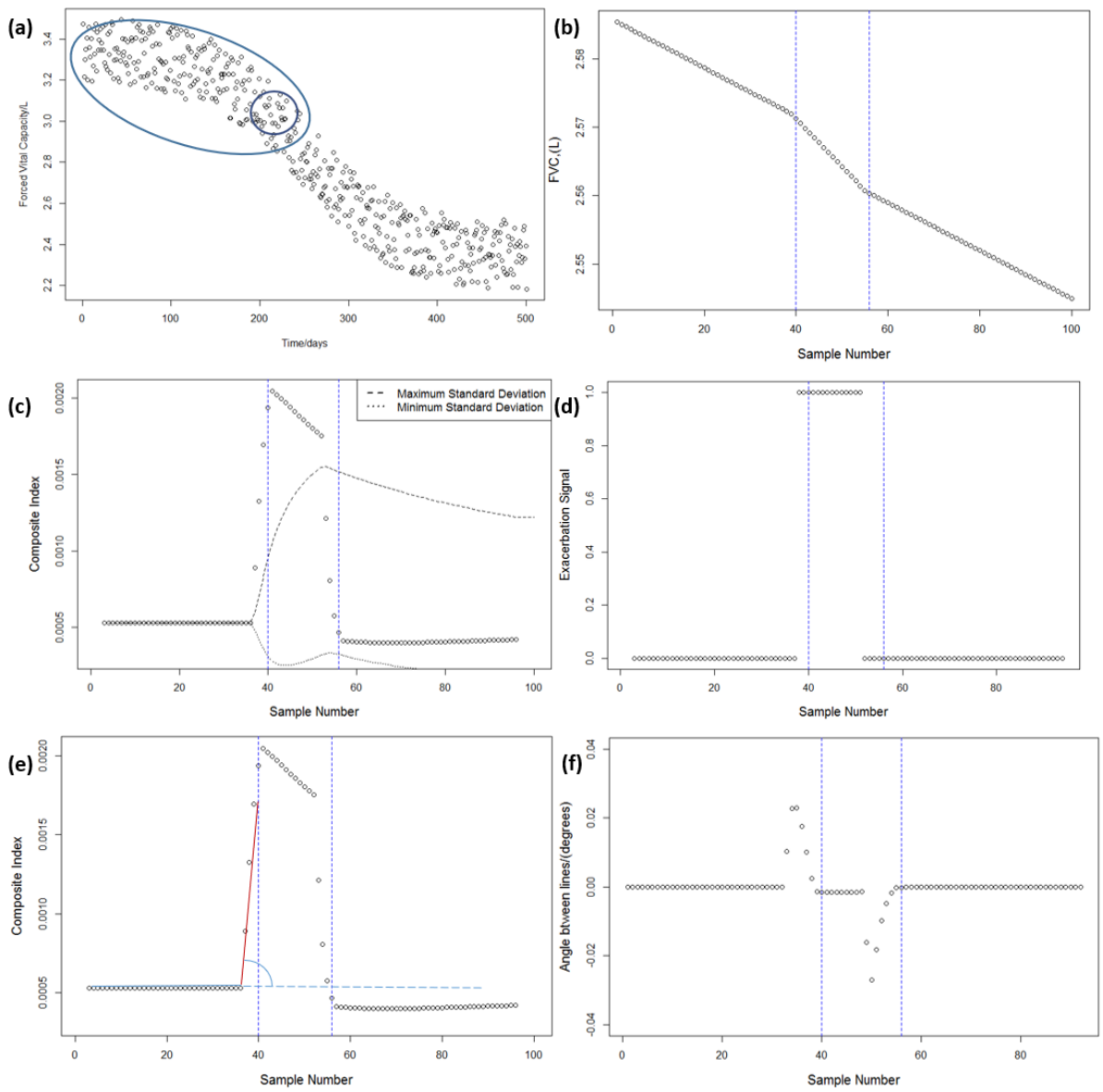


Figure 4-5 | ***Calculation and interpretation of the composite index to assess the presence of a critical transition.*** *Simulated spirometry was used to form the basis for calculating the composite index that may be used to indicate the beginning of an acute exacerbation (a). Two linear regressions (represented by the ovals) are compared in the composite index to contrast the short-term decline in lung function with lung decline since baseline. The numerical value of the composite index needs to interpreted with respect to the baseline reading of each individual patient. The additive (c,d) and piecewise (e, f) models allow the comparison of composite index values for each patient to generate either a binary exacerbation signal for the additive model (d) or no-zero angle between adjacent linear regression gradients for the piecewise model (f). For the additive model the exacerbation signal is generated when three consecutive values exceed the 95% confidence interval (c).*

4.4 Discussion

The results of this chapter are that the imputation of outliers within patients' time series can be improved with techniques like the Kalman filter that account for present trends. In addition, dynamic descriptors, like the Poincaré plot, have the potential to classify IPF patients according to their disease progression. Also, the application of the ARIMA model to FVC time series of IPF patients accounts for all trends present within the data. Finally, treating an acute exacerbation as a critical transition has the potential to provide an early-warning signal to this clinically significant event.

The treatment of outliers is deemed to be important prior to determining the parameters of the ARIMA statistical model given that they reduce the ability of the model to account for trends within time series data. Outliers are likely to be present within the data due to data entry error and can be addressed by the introduction of Wi-Fi-enabled spirometers that can transmit data in real-time[183]. Whilst real-time transmission of data can promote patient compliance and reduce the extent of missing data, there is no guarantee the elimination of missing values. Therefore, dealing with missing data and outliers has the potential to improve ARIMA model's ability to represent and forecast trends in disease progression.

Inappropriate statistical modelling of home spirometry data and treatment of outliers can cause unanticipated technical and analytical issues that prevent potentially beneficial treatment being made available to patients of ILD[204]. Linear regression models do not best account for the information within home spirometry data meaning that large numbers of measurements are required to prevent the prediction of physiologically implausible values, particularly in the presence of outliers[204].

Outliers were identified retrospectively using statistical descriptors, such as mean and standard deviation, which apply to the entire time series. IPF disease progression can be non-linear, particularly during an acute exacerbation[4], meaning that outlier treatment may be benefitted by segmenting the time series into shorter periods of uniform disease progression. Discrete periods of uniform disease progression can be aided by the application of piecewise linear regression to a sequence of localised maximum and minimum values (Figure 4.4).

Both detrended fluctuation analysis (DFA) and Poincare plot analysis are complementary, non-linear methods that can be used to analyse the variability in physiological time series[205]. DFA

provides insight into the long-range correlations and, therefore adaptability, within complex biological systems[122,162,205]. In contrast, Poincare plot analysis permits the quantitative and visual analysis of time series fluctuations and, specifically the evolution of a dynamical system in phase space[205]. A significant consequence of the Poincare plot being defined in phase space, rather than time space for DFA, is that the Poincare plot is influenced by the length of the interval but not the amount of intervals that occur[206].

Whilst this study has utilised DFA to establish that FVC time series from IPF patients contain long-range correlations, a limitation is that this technique requires a minimum of 75 daily observations. Poincare plot analysis has the potential to give a clinically useful insight into changes in patient well-being because it does not require long time series. The insights that Poincare plot analysis can provide regarding the variability of FVC time series and other parameters is yet to be verified. Power analysis indicate that at least 45 patients per clinical progression group are required (Figure 4.3) to determine whether the Poincare descriptors can distinguish patients' disease progression. However, if Poincare descriptors can be used to characterise IPF patient disease progression, this would greatly assist establishing patient prognosis early in their clinical care.

Statistical modelling of IPF disease progression can be achieved using a combination of autoregressive (AR) and moving average (MA) models. Inspection of residuals using the Ljung-Box test indicate randomness and that all systematic trends have been captured within the model. Each FVC time series were characterised by either an ARIMA(1,1,1) or ARIMA(1,1,0) model with unique combinations of coefficients relating to each model. These models characterise their respective time series and so can be used to indicate clinically significant events such as an acute exacerbation. In addition, ARIMA models can be used to forecast disease progression, which may potentially be useful to predict responses to therapy.

The ability to identify patients who experience an acute exacerbation is clinically significant in terms of patient care, treatment and understanding disease mechanisms to develop therapy. Likening an acute exacerbation to a critical transition assists in the development of potential approaches that allow the trends within the time series to be understood. The composite index permits the detection of an acute deterioration by comparing the rate disease progression before and after a deterioration. An assumption is that the patient has been diagnosed prior to an acute exacerbation and has been monitored for a prescribed period of time.

Each patient is diagnosed at different stages during their illness meaning that their baseline FVC will be different[4]. Interpretation of changes in the composite index can be achieved by the either the additive or piecewise models in order to highlight a critical transition. The additive model has the advantage that the entire period of acute deterioration is marked by a '1' signal of binary output. In contrast, the piecewise model indicates the start and end of the acute deterioration with a positive or negative value but not the period. Whilst it is possible to utilise these models in conjunction, how these models perform in identifying a number of cases with an acute exacerbation remains to be seen.

Much of the work has focused on the development of statistical models used to understand IPF disease progression from longitudinal FVC measurements. This is clinically important as it forms the basis for home monitoring of patients and timely administration of care. The statistical models, however, do not directly provide insight into the molecular mechanisms of IPF disease progression. These statistical models can be utilised during clinical trials to aid analysis of changes in biomarker expression during clinically important events and in response to potential therapies. Therefore, these statistical models could help to elucidation of the mechanisms that underpin IPF disease progression.

There is sufficient data within this chapter to partially accept the hypothesis that the temporal information within FVC time series' can be utilised to classify IPF disease progression. However, further work is required to address the application of various analytical approaches and the weighting of physiological parameters for individual patients given their therapeutic background and comorbidities. It would be of particular interest to establish how the Poincaré descriptors reflect the IPF clinical course, the parameters of the ARIMA model and whether the composite index can be used as an early warning signal of an acute exacerbation.

5 General Discussion

Idiopathic pulmonary fibrosis (IPF) is characterised by poor prognosis with each patient experiencing a unique clinical course defined by their rate of deterioration, acute deterioration and death[145,147]. Presently, there is no established approach to combine the individual clinical parameters that do correlate with survival to accurately determine prognosis. The motivation for this study is to apply statistical time series techniques to obtain concrete and reliable information about IPF disease progression.

In chapter 3, I determine common statistical features within the longitudinal forced vital capacity (FVC) measurements of IPF patients characterised by distinct disease progression profiles. Specifically, a linear regression model does not provide an adequate approach to account for all the information within FVC time series. In addition, FVC measurements are correlated to readings taken the day before, as evaluated by autocorrelation. Finally, FVC time series from IPF patients exhibit long-range correlation, or memory, as determined by detrended fluctuation analysis (DFA).

Chapter 4 develops an approach to determine the disease progression on an individual, real-time basis using a suitable statistical model. In particular, the autoregressive integrative moving average (ARIMA) class of statistical model was deemed appropriate given the non-stationary and lag-1 autocorrelation nature of longitudinal FVC time series. The parameters corresponding to the most parsimonious ARIMA model for each patient can be determined on an automated basis. Each ARIMA model forms the basis for characterising IPF disease progression with regards to deterioration rate, acute exacerbation events and responses to therapy.

Even though these findings have been addressed in detail within their respective chapters, general concepts emerge that warrant further consideration. Therefore, during this concluding discussion, I will reflect on the following:

1. Statistical models underpinning systems biology of IPF
2. ARIMA modelling and modelling of IPF disease progression
3. Non-linear dynamical analysis
4. Outliers and missing data

1. Statistical models underpinning systems biology of IPF

Home-based monitoring of physiological variables for complex respiratory diseases has become increasingly practicable due to miniaturisation and falling cost of equipment. In concert with improvements of wireless electronic connectivity, daily surveillance spirometry has the potential advantage for providing a faithful impression of disease behaviour and subsequent outcome[64]. Development of statistical models are necessary to exploit the longitudinal trends captured by more frequent observation, and to account for IPF as a heterogeneous disease with an unpredictable clinical course[183].

Recent technological advances including wireless-enabled spirometers accompanied by ready access to the internet can facilitate real-time feedback on home-based spirometry technique and optimise compliance[65]. Therefore, the lung function datasets generated should be characterised by fewer missing data and trends that are influenced by less technical variability[207], making the development of accurate prognostic statistical models easier. Understanding the features of physiological time series, and suitable statistical techniques, will assist clinicians to gain a better impression of disease behaviour from statistical models based on longitudinal data.

The purpose for any statistical model of lung function data will involve the interpretation of trends relating to IPF disease progression. A recommended approach to the staging of IPF disease in clinical practice is a 10% change in FVC over a 12-month period because this threshold correlates well with mortality[3,45,58]. The benefit of FVC severity threshold in guiding the clinical management of IPF patients centres on dichotomising a continuous variable. The limitation of the FVC severity threshold is its assumption that the clinical course of IPF disease progression is linear and fails to account for clinically significant events such as acute exacerbations that correlate with high-levels of mortality[35,38].

Intra-patient variability of longitudinal FVC measurements may influence the time period over which the FVC severity threshold is calculated, particularly if involving serial local maxima and minima. Identification of local maxima and minima within FVC time series can improve the correlation of serial change with mortality meaning that the FVC severity threshold could be calculated over a shorter period. Removal of noise using statistical techniques may seem to be a more straightforward solution but may be detrimental as the noise of the time series contains important information about the physiology of the lung.

The nonlinear dynamical analysis that I have conducted, in particular the detrended fluctuation analysis (DFA), has indicated that the temporal pattern in FVC time series of IPF patients exhibits long-term memory and adaptability[115]. Whilst the precise aetiology of IPF is unknown, it is believed that the excessive scarring within the lung is the consequence of an aberrant wound response to injury of the lung epithelium[208]. Therefore, I believe that fluctuations of FVC time series is informative of the IPF lung's ability to restore pulmonary lung function. The FVC time series of the acute exacerbation reinforces this belief given that there are two distinct distributions, and therefore two attractors, about which fluctuations in FVC occur.

The aberrant wound healing response of the IPF lung is consequence of the interaction of multiple pathways, cell types and biological processes[68]. In addition, different contributions of cell types and wound healing pathways underpins the heterogeneity observed between IPF patients[39,68,209]. Therefore, a systems biology approach is essential to understanding IPF disease and entails the longitudinal monitoring of multiple genes and proteins[39,210]. Lung function time series can guide the acquisition of samples from IPF patients and help to contextualise -omics data with respect to individual disease progression.

Continual assessment of FVC time series, including nonlinear dynamical measures, can provide a foundation for the analysis of IPF disease mechanisms by providing an clear impression of disease progression including acute exacerbations[124,191,211]. Periodic hospital-based assessment can include sampling of patient serum and lung tissue via bronchoscopy. Retrospective transcriptomic and proteomic analysis can aid the generation of dynamical network biomarkers that may highlight important mediators of disease progression[142,212]. The ultimate aim is the identification of reliable markers of disease activity at an early enough stage to permit the administration of disease-modifying therapy.

Analysis of FVC time series is an important first-step in improving our understanding of the mechanisms of IPF disease progression and the development of effective and timely therapy. Whilst precise FVC measurement is essential to describing IPF clinical history, changes in FVC is the consequence of significant, irreversible modifications to the lung architecture. Development of accurate, non-invasive biomarkers can be invaluable tools to evaluate future disease behaviour, thereby allowing early tailoring of therapy including need for transplant[39]. In addition, biomarkers have the potential to detect the onset of acute exacerbation, possibly permitting their prevention or limiting their severity.

2. ARIMA modelling and modelling of IPF disease progression

The autocorrelation analysis and ARIMA models relating of the FVC time series infer serial dependence between FVC measurements that are no more than a day apart. This is significant because it asserts that the frequency of home-based spirometry should be at least daily. Requiring more frequent measurements would risk undermining the freedom and convenience that remote monitoring offers patients whilst potentially reducing the degree of patient compliance. The ARIMA models generated were deemed to account for all the trends present with their corresponding time series meaning that these models can be applied for the efficient resource management of hospital resources[101], risk assessment of tuberculosis infection[213] and monitoring of intensive care patients[160].

Statistical analysis using the ARIMA modelling has the potential to be a significant resource for the characterising and forecasting of IPF disease progression on an ongoing basis. ARIMA-based forecasting of disease trajectory may be indicative of clinically significant events such as acute exacerbations. A potential limitation of ARIMA-based modelling in predicting an acute exacerbation is the length of the time series prior to the event, given that some IPF patients are diagnosed whilst experiencing acute exacerbations[214]. Interrupted time series analysis, involving ARIMA modelling, to assess the efficacy and response to therapy during clinical trials can determine when treatment begins thereby circumventing this issue[215].

ARIMA modelling utilises the two distinct features of time series, the moving average and autocorrelation, to define the trends present within the associated system. I believe that the propensity of ARIMA statistical models relies on autocorrelation and, at least, daily lung function measurements. In contrast, the difficultly to apply technical analysis within the financial markets is meaningless because either there is no memory present or we lack the mechanistic insights to develop statistical models with appropriate predictive power[127,216–218].

An unexplored facet of ARIMA modelling and its application for IPF disease progression is its ability to identify periodic components including seasonality within time series. Long-term home monitoring of IPF patients has the potential to identify how disease progression changes throughout the year, thereby potentially guiding the management of resources according to patient needs.

3. Non-linear dynamical analysis

Detrended fluctuation analysis (DFA) conducted within this project indicates that FVC time series of IPF patients exhibits long-range correlations and adaptability or memory. The scaling exponent generated by DFA has been demonstrated to vary in the time series of patients with asthma and chronic obstructive pulmonary disease (COPD) in response to therapy[121,122]. Non-linear dynamical analysis can be a useful indicator of treatment efficacy in clinical practice or during clinical trials. Additional investigation is required to examine how the scaling exponent of DFA varies with the two anti-fibrotic agents of demonstrated therapeutic benefit, nintedanib and pirfenidone, with the slowing of disease progression in IPF patients[40,70,219].

The behaviour of dynamical indicators, like the scaling exponent, is dependent on therapy type and disease context, so its characterisation is essential for the interpretation of IPF patients' response to therapy[121,122]. Dynamical indicators may potentially guide the administration of concomitant or switching between pirfenidone and nintedanib monotherapy[220]. DFA has been invaluable in establishing non-linear dynamics within the FVC time series of IPF patient. DFA requires a relatively long time series in order to evaluate the presence of long-range correlations, and therefore may be limited in indicating real-time changes in disease progression[120,121].

Alternative non-linear dynamical systems methods include Lyapunov exponents, recurrence plot, Poincaré plot and approximate entropy, which may be more suitable for real-time monitoring due to using shorter time series[206]. My motivation to using the Poincaré plot technique as a short-term dynamical analysis is its graphical approach to characterisation, meaning that its interpretation may be more accessible to clinicians and patients. The Poincaré plot descriptors provide an indication of the short- and long-term variability of the time series which may be useful in indicating the presence of an acute exacerbation. Further evidence is required to support the usefulness of the Poincaré plot analysis in defining the features of FVC time series' of different types of disease progression, particularly in terms of power.

4. Outliers and missing data

Until recently, there has been limited evidence to suggest that any drug had the ability to change the course of IPF disease progression[221]. The efficacy of pirfenidone and nintedanib has been shown to slow down the course of IPF due to the worldwide cooperative efforts of a few large randomised controlled trials[40,70,71]. These trials have utilised FVC decline as the primary end-point, in conjunction with recommendations from the US Food and Drug Administration (FDA), and have measured mortality to sustain the FVC results[222]. However, lung function parameters and vital status data could not be evaluated for some study participants who have dropped out because of loss to follow-up, withdrawal or death resulting in missing data.

Missing data for the primary endpoint poses obstacles for the evaluation of the magnitude and statistical significance of the treatment effect. However, it is important to note that most missing data are informative in nature and possibly biased across treatment groups, particularly when assessing efficacy outcomes. The success of future clinical trials relies on understanding the main challenges that missing values pose and the limitations of the methods used to handle them.

Imputation strategies employed during the INPULSIS trials required that the mean FVC at each time-point for all non-missing available data[70,71]. This meant that FVC was computed from a decreasing pool of patients over time and assumed that FVC decline was the same between those patients who dropped out and those who remained within the study. An alternative approach is the use of composite end-point, including death or FVC decline, which assigns a similar weighting to weight to death or FVC decline >10%[223]. However, there is a considerable loss of power because there is no natural FVC cut-off and dichotomisation of continuous variable (like FVC).

Employment of distinct imputation strategies between clinical drugs trials could create confusion with results interpretation, e.g. differences in imputation policies between the ASCEND and INPULSIS trials led to confusion between inclusion criteria and recruitment issues[224]. The use of identical imputation strategies is unlikely to avoid false estimates when comparing the effect of drugs between two separate trials. Nonetheless, the optimal approach to determine the best drug is to conduct an adequately powered head to head study.

General summary

The purpose of this project is to characterise the features of the FVC time series of IPF patients and to evaluate methods to define IPF disease progression. Classifying IPF disease progression based on continual, real-time home spirometry will provide important timely information about responses to therapy and changes in patient well-being prior to the onset of potentially irreversible physiological change in the lung. Given that the lung is a complex, dynamical system, the consequence of multiple components interacting together, statistical modelling of a physiological parameter represents a first-step in a long journey to effective treatment of IPF patients.

Identification of an acute deterioration based on FVC time series, even whilst it is occurring, may well be too late for the delivery of meaningful treatment to reverse any decline. FVC time series analysis can form the basis of understanding other biological parameters including biomarkers of serum, bronchoalveolar lavage or lung epithelium in the context of lung physiology[76,77,225,226]. Elucidating biomarkers that correlate with distinct stages of IPF disease progression, including acute exacerbations, can help us to understand how disease develops, provide insight into therapeutic mechanisms and stratify patients according to therapeutic responses.

Lung function measurements are physiological observations that are the consequence of a number of environmental and biological influences interacting together[112]. An acute exacerbation experienced by a subset of IPF patients may be likened to a critical transition on the basis of an abrupt change between two distinct stable physiological states[124,142,227]. IPF patients may experience a susceptible state prior to an acute exacerbation during which a clinical event may be avoided with a suitable intervention[212]. This so-called "pre-disease" state may be identified by dynamical network biomarkers and act as a warning prior to irreversible changes in physiology[212].

There is evidence to partially accept the overall hypothesis of this study that the clinical course of IPF can be classified by deterministic methods applied to longitudinal lung function data. This study is limited as how to integrate the temporal features of physiological function and vital data into a staging system for IPF disease progression. Future studies will need to select enough patients representing all IPF clinical courses, particularly acute exacerbations, to assess the application of analytical approaches.

6 References

1. Behr, J. Approach to the diagnosis of interstitial lung disease. *Clin. Chest Med.* **33**, 1–10 (2012).
2. Barnes, P. J. Chronic obstructive pulmonary disease. *N. Engl. J. Med.* **343**, 269–80 (2000).
3. Raghu, G. *et al.* 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 (2011).
4. Ley, B., Collard, H. R. & King, T. E. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **183**, 431–40 (2011).
5. Maher, T. M. *et al.* Development of a Consensus Statement for the Definition, Diagnosis, and Treatment of Acute Exacerbations of Idiopathic Pulmonary Fibrosis Using the Delphi Technique. *Adv. Ther.* **32**, 929–43 (2015).
6. Vogelmeier, C. F. *et al.* Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. *Am. J. Respir. Crit. Care Med.* **195**, 557–582 (2017).
7. Rennard, S. I. & Vestbo, J. COPD: the dangerous underestimate of 15%. *Lancet (London, England)* **367**, 1216–9 (2006).
8. Russi, E. W. *et al.* Management of chronic obstructive pulmonary disease: the Swiss guidelines. Official Guidelines of the Swiss Respiratory Society. *Swiss Med. Wkly.* **132**, 67–78 (2002).
9. Lopez, A. D. & Murray, C. C. The global burden of disease, 1990-2020. *Nat. Med.* **4**, 1241–3 (1998).
10. Selman, M. *et al.* Idiopathic pulmonary fibrosis: prevailing and evolving hypotheses about its pathogenesis and implications for therapy. *Ann. Intern. Med.* **134**, 136–51 (2001).
11. Idiopathic Pulmonary Fibrosis Clinical Research Network *et al.* Prednisone, azathioprine, and N-acetylcysteine for pulmonary fibrosis. *N. Engl. J. Med.* **366**, 1968–77 (2012).
12. Lederer, D. J. & Martinez, F. J. Idiopathic Pulmonary Fibrosis. *N. Engl. J. Med.* **378**, 1811–1823 (2018).
13. Parker, M. W. *et al.* Fibrotic extracellular matrix activates a profibrotic positive feedback loop. *J. Clin. Invest.* **124**, 1622–35 (2014).
14. Fingerlin, T. E. *et al.* Genome-wide association study identifies multiple susceptibility

loci for pulmonary fibrosis. *Nat. Genet.* **45**, 613–20 (2013).

15. Noth, I. *et al.* Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet. Respir. Med.* **1**, 309–317 (2013).
16. Evans, C. M. *et al.* Idiopathic Pulmonary Fibrosis: A Genetic Disease That Involves Mucociliary Dysfunction of the Peripheral Airways. *Physiol. Rev.* **96**, 1567–91 (2016).
17. Armanios, M. Y. *et al.* Telomerase mutations in families with idiopathic pulmonary fibrosis. *N. Engl. J. Med.* **356**, 1317–26 (2007).
18. Tsakiri, K. D. *et al.* Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7552–7 (2007).
19. Wang, Y. *et al.* Genetic defects in surfactant protein A2 are associated with pulmonary fibrosis and lung cancer. *Am. J. Hum. Genet.* **84**, 52–9 (2009).
20. Nogee, L. M. *et al.* A mutation in the surfactant protein C gene associated with familial interstitial lung disease. *N. Engl. J. Med.* **344**, 573–9 (2001).
21. Nakano, Y. *et al.* MUC5B Promoter Variant rs35705950 Affects MUC5B Expression in the Distal Airways in Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **193**, 464–6 (2016).
22. Demedts, M. & Costabel, U. ATS/ERS international multidisciplinary consensus classification of the idiopathic interstitial pneumonias. *Eur. Respir. J.* **19**, 794–796 (2002).
23. Chambers, R. C. Role of coagulation cascade proteases in lung repair and fibrosis. *Eur. Respir. J. Suppl.* **44**, 33s-35s (2003).
24. Kinnula, V. L., Fattman, C. L., Tan, R. J. & Oury, T. D. Oxidative stress in pulmonary fibrosis: a possible role for redox modulatory therapy. *Am. J. Respir. Crit. Care Med.* **172**, 417–22 (2005).
25. Thannickal, V. J. & Horowitz, J. C. Evolving concepts of apoptosis in idiopathic pulmonary fibrosis. *Proc. Am. Thorac. Soc.* **3**, 350–6 (2006).
26. Keane, M. P., Strieter, R. M. & Belperio, J. A. Mechanisms and mediators of pulmonary fibrosis. *Crit. Rev. Immunol.* **25**, 429–63 (2005).
27. Willis, B. C., DuBois, R. M. & Borok, Z. Epithelial origin of myofibroblasts during fibrosis in the lung. *Proc. Am. Thorac. Soc.* **3**, 377–82 (2006).
28. Mehrad, B. *et al.* Circulating peripheral blood fibrocytes in human fibrotic interstitial lung disease. *Biochem. Biophys. Res. Commun.* **353**, 104–8 (2007).
29. Raghu, G. *et al.* A placebo-controlled trial of interferon gamma-1b in patients with idiopathic pulmonary fibrosis. *N. Engl. J. Med.* **350**, 125–33 (2004).

30. Kim, D. S. Acute exacerbation of idiopathic pulmonary fibrosis. *Clin. Chest Med.* **33**, 59–68 (2012).
31. Kim, D. S. *et al.* Acute exacerbation of idiopathic pulmonary fibrosis: frequency and clinical features. *Eur. Respir. J.* **27**, 143–50 (2006).
32. Olson, A. L., Swigris, J. J., Raghu, G. & Brown, K. K. Seasonal variation: mortality from pulmonary fibrosis is greatest in the winter. *Chest* **136**, 16–22 (2009).
33. Simon-Blancal, V. *et al.* Acute exacerbation of idiopathic pulmonary fibrosis: outcome and prognostic factors. *Respiration.* **83**, 28–35 (2012).
34. Collard, H. R. *et al.* Suspected acute exacerbation of idiopathic pulmonary fibrosis as an outcome measure in clinical trials. *Respir. Res.* **14**, 73 (2013).
35. Song, J. W., Hong, S.-B., Lim, C.-M., Koh, Y. & Kim, D. S. Acute exacerbation of idiopathic pulmonary fibrosis: incidence, risk factors and outcome. *Eur. Respir. J.* **37**, 356–63 (2011).
36. Lee, J. S. *et al.* Anti-acid treatment and disease progression in idiopathic pulmonary fibrosis: an analysis of data from three randomised controlled trials. *Lancet. Respir. Med.* **1**, 369–76 (2013).
37. Konishi, K. *et al.* Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **180**, 167–75 (2009).
38. Collard, H. R. *et al.* Acute exacerbations of idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **176**, 636–43 (2007).
39. Maher, T. M. Beyond the diagnosis of idiopathic pulmonary fibrosis; the growing role of systems biology and stratified medicine. *Curr. Opin. Pulm. Med.* **19**, 460–5 (2013).
40. Noble, P. W. *et al.* Pirfenidone in patients with idiopathic pulmonary fibrosis (CAPACITY): two randomised trials. *Lancet (London, England)* **377**, 1760–9 (2011).
41. Magnini, D. *et al.* Idiopathic Pulmonary Fibrosis: Molecular Endotypes of Fibrosis Stratifying Existing and Emerging Therapies. *Respiration.* **93**, 379–395 (2017).
42. Rogliani, P., Calzetta, L., Cavalli, F., Matera, M. G. & Cazzola, M. Pirfenidone, nintedanib and N-acetylcysteine for the treatment of idiopathic pulmonary fibrosis: A systematic review and meta-analysis. *Pulm. Pharmacol. Ther.* **40**, 95–103 (2016).
43. du Bois, R. M. *et al.* Ascertainment of individual risk of mortality for patients with idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **184**, 459–66 (2011).
44. Richeldi, L. *et al.* Relative versus absolute change in forced vital capacity in idiopathic pulmonary fibrosis. *Thorax* **67**, 407–11 (2012).
45. Zappala, C. J. *et al.* Marginal decline in forced vital capacity is associated with a poor outcome in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **35**, 830–6 (2010).

46. Durheim, M. T. *et al.* Association of hospital admission and forced vital capacity endpoints with survival in patients with idiopathic pulmonary fibrosis: analysis of a pooled cohort from three clinical trials. *Lancet. Respir. Med.* **3**, 388–96 (2015).
47. Ley, B. *et al.* Predictors of Mortality Poorly Predict Common Measures of Disease Progression in Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **194**, 711–8 (2016).
48. Maher, T. M. Precision medicine in idiopathic pulmonary fibrosis. *QJM* **109**, 585–587 (2016).
49. Swigris, J. J. *et al.* The UCSD shortness of breath questionnaire has longitudinal construct validity in idiopathic pulmonary fibrosis. *Respir. Med.* **106**, 1447–55 (2012).
50. Manali, E. D. *et al.* The Medical Research Council chronic dyspnea score predicts the survival of patients with idiopathic pulmonary fibrosis. *Respir. Med.* **102**, 586–92 (2008).
51. du Bois, R. M. *et al.* 6-Minute walk distance is an independent predictor of mortality in patients with idiopathic pulmonary fibrosis. *Eur. Respir. J.* **43**, 1421–9 (2014).
52. Flaherty, K. R. *et al.* Idiopathic pulmonary fibrosis: prognostic value of changes in physiology and six-minute-walk test. *Am. J. Respir. Crit. Care Med.* **174**, 803–9 (2006).
53. Lederer, D. J. *et al.* Six-minute-walk distance predicts waiting list survival in idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **174**, 659–64 (2006).
54. Brown, A. W. *et al.* Outcomes after hospitalization in idiopathic pulmonary fibrosis: a cohort study. *Chest* **147**, 173–179 (2015).
55. Flaherty, K. R. *et al.* Prognostic implications of physiologic and radiographic changes in idiopathic interstitial pneumonia. *Am. J. Respir. Crit. Care Med.* **168**, 543–8 (2003).
56. Collard, H. R. *et al.* Changes in clinical and physiologic variables predict survival in idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **168**, 538–42 (2003).
57. King, T. E. *et al.* Analyses of efficacy end points in a controlled trial of interferon-gamma1b for idiopathic pulmonary fibrosis. *Chest* **127**, 171–7 (2005).
58. Latsi, P. I. *et al.* Fibrotic idiopathic interstitial pneumonia: the prognostic value of longitudinal functional trends. *Am. J. Respir. Crit. Care Med.* **168**, 531–7 (2003).
59. Kolb, M. & Collard, H. R. Staging of idiopathic pulmonary fibrosis: past, present and future. *Eur. Respir. Rev.* **23**, 220–4 (2014).
60. Salisbury, M. L. *et al.* Idiopathic Pulmonary Fibrosis: Gender-Age-Physiology Index Stage for Predicting Future Lung Function Decline. *Chest* **149**, 491–498 (2016).
61. Saketkoo, L. A. *et al.* Connective tissue disease related interstitial lung diseases and idiopathic pulmonary fibrosis: provisional core sets of domains and instruments for use

in clinical trials. *Thorax* **69**, 428–36 (2014).

62. Nathan, S. D. *et al.* Long-term course and prognosis of idiopathic pulmonary fibrosis in the new millennium. *Chest* **140**, 221–229 (2011).
63. du Bois, R. M. *et al.* Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference. *Am. J. Respir. Crit. Care Med.* **184**, 1382–9 (2011).
64. Russell, A.-M. *et al.* Daily Home Spirometry: An Effective Tool for Detecting Progression in Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* (2016). doi:10.1164/rccm.201511-2152OC
65. Johannson, K. A. *et al.* Home monitoring improves endpoint efficiency in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **50**, (2017).
66. Moor, C. C. *et al.* A home monitoring program including real-time wireless home spirometry in idiopathic pulmonary fibrosis: a pilot study on experiences and barriers. *Respir. Res.* **19**, 105 (2018).
67. Chiba, H., Otsuka, M. & Takahashi, H. Significance of molecular biomarkers in idiopathic pulmonary fibrosis: A mini review. *Respir. Investig.* **56**, 384–391 (2018).
68. Maher, T. M., Wells, A. U. & Laurent, G. J. Idiopathic pulmonary fibrosis: multiple causes and multiple mechanisms? *Eur. Respir. J.* **30**, 835–9 (2007).
69. Collard, H. R. *et al.* A new era in idiopathic pulmonary fibrosis: considerations for future clinical trials. *Eur. Respir. J.* **46**, 243–9 (2015).
70. King, T. E. *et al.* A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N. Engl. J. Med.* **370**, 2083–92 (2014).
71. Richeldi, L. *et al.* Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N. Engl. J. Med.* **370**, 2071–82 (2014).
72. Hilberg, F. *et al.* BIBF 1120: triple angiokinase inhibitor with sustained receptor blockade and good antitumor efficacy. *Cancer Res.* **68**, 4774–82 (2008).
73. Chaudhary, N. I. *et al.* Inhibition of PDGF, VEGF and FGF signalling attenuates fibrosis. *Eur. Respir. J.* **29**, 976–85 (2007).
74. Lee, B. S., Margolin, S. B. & Nowak, R. A. Pirfenidone: a novel pharmacological agent that inhibits leiomyoma cell proliferation and collagen production. *J. Clin. Endocrinol. Metab.* **83**, 219–23 (1998).
75. Schaefer, C. J., Ruhrmund, D. W., Pan, L., Seiwert, S. D. & Kossen, K. Antifibrotic activities of pirfenidone in animal models. *Eur. Respir. Rev.* **20**, 85–97 (2011).
76. Jenkins, R. G. *et al.* Longitudinal change in collagen degradation biomarkers in idiopathic pulmonary fibrosis: an analysis from the prospective, multicentre PROFILE

study. *Lancet. Respir. Med.* **3**, 462–72 (2015).

77. Maher, T. M. *et al.* An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicentre PROFILE cohort study. *Lancet. Respir. Med.* **5**, 946–955 (2017).
78. Johannessen, A. *et al.* Post-bronchodilator spirometry reference values in adults and implications for disease management. *Am. J. Respir. Crit. Care Med.* **173**, 1316–25 (2006).
79. Qaseem, A. *et al.* Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the American College of Physicians, American College of Chest Physicians, American Thoracic Society, and European Respiratory Society. *Ann. Intern. Med.* **155**, 179–91 (2011).
80. Anthonisen, N. R., Wright, E. C. & Hodgkin, J. E. Prognosis in chronic obstructive pulmonary disease. *Am. Rev. Respir. Dis.* **133**, 14–20 (1986).
81. Burrows, B. The course and prognosis of different types of chronic airflow limitation in a general population sample from Arizona: comparison with the Chicago 'COPD' series. *Am. Rev. Respir. Dis.* **140**, S92-4 (1989).
82. Wise, R. A. The value of forced expiratory volume in 1 second decline in the assessment of chronic obstructive pulmonary disease progression. *Am. J. Med.* **119**, 4–11 (2006).
83. Zemans, R. L. *et al.* Multiple biomarkers predict disease severity, progression and mortality in COPD. *Respir. Res.* **18**, 117 (2017).
84. Cosio, M. *et al.* The relations between structural changes in small airways and pulmonary-function tests. *N. Engl. J. Med.* **298**, 1277–81 (1978).
85. Chung, K. F. & Adcock, I. M. Multifaceted mechanisms in COPD: inflammation, immunity, and tissue repair and destruction. *Eur. Respir. J.* **31**, 1334–56 (2008).
86. Saetta, M. *et al.* Activated T-lymphocytes and macrophages in bronchial mucosa of subjects with chronic bronchitis. *Am. Rev. Respir. Dis.* **147**, 301–6 (1993).
87. Finkelstein, R., Fraser, R. S., Ghezzo, H. & Cosio, M. G. Alveolar inflammation and its relation to emphysema in smokers. *Am. J. Respir. Crit. Care Med.* **152**, 1666–72 (1995).
88. Sun, W., Wu, R. & Last, J. A. Effects of exposure to environmental tobacco smoke on a human tracheobronchial epithelial cell line. *Toxicology* **100**, 163–74 (1995).
89. Wickenden, J. A. *et al.* Cigarette smoke prevents apoptosis through inhibition of caspase activation and induces necrosis. *Am. J. Respir. Cell Mol. Biol.* **29**, 562–70 (2003).
90. Richards, G. A., Theron, A. J., Van der Merwe, C. A. & Anderson, R. Spirometric abnormalities in young smokers correlate with increased chemiluminescence responses of activated blood phagocytes. *Am. Rev. Respir. Dis.* **139**, 181–7 (1989).

91. Jiang, D. *et al.* Regulation of lung injury and repair by Toll-like receptors and hyaluronan. *Nat. Med.* **11**, 1173–9 (2005).
92. Marshak-Rothstein, A. Toll-like receptors in systemic autoimmune disease. *Nat. Rev. Immunol.* **6**, 823–35 (2006).
93. Freeman, C. M., Curtis, J. L. & Chensue, S. W. CC chemokine receptor 5 and CXC chemokine receptor 6 expression by lung CD8+ cells correlates with chronic obstructive pulmonary disease severity. *Am. J. Pathol.* **171**, 767–76 (2007).
94. Saetta, M. *et al.* Increased expression of the chemokine receptor CXCR3 and its ligand CXCL10 in peripheral airways of smokers with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **165**, 1404–9 (2002).
95. Grumelli, S. *et al.* An immune basis for lung parenchymal destruction in chronic obstructive pulmonary disease and emphysema. *PLoS Med.* **1**, e8 (2004).
96. Calabrese, F. *et al.* Marked alveolar apoptosis/proliferation imbalance in end-stage emphysema. *Respir. Res.* **6**, 14 (2005).
97. Hogg, J. C. *et al.* The nature of small-airway obstruction in chronic obstructive pulmonary disease. *N. Engl. J. Med.* **350**, 2645–53 (2004).
98. van der Strate, B. W. A. *et al.* Cigarette smoke-induced emphysema: A role for the B cell? *Am. J. Respir. Crit. Care Med.* **173**, 751–8 (2006).
99. Paré, G., Jaana, M. & Sicotte, C. Systematic review of home telemonitoring for chronic diseases: the evidence base. *J. Am. Med. Inform. Assoc.* **14**, 269–77 (2007).
100. Nathan, S. D. *et al.* Long-term course and prognosis of idiopathic pulmonary fibrosis in the new millennium. *Chest* **140**, 221–9 (2011).
101. Sato, R. C. Disease management with ARIMA model in time series. *Einstein (Sao Paulo).* **11**, 128–31
102. Earnest, A., Chen, M. I., Ng, D. & Sin, L. Y. Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Serv. Res.* **5**, 36 (2005).
103. Liu, Q., Liu, X., Jiang, B. & Yang, W. Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. *BMC Infect. Dis.* **11**, 218 (2011).
104. López-Lozano, J. M. *et al.* Modelling and forecasting antimicrobial resistance and its dynamic relationship to antimicrobial use: a time series analysis. *Int. J. Antimicrob. Agents* **14**, 21–31 (2000).
105. Hsueh, P.-R., Chen, W.-H. & Luh, K.-T. Relationships between antimicrobial use and antimicrobial resistance in Gram-negative bacteria causing nosocomial infections from 1991-2003 at a university hospital in Taiwan. *Int. J. Antimicrob. Agents* **26**, 463–72

(2005).

106. Aldeyab, M. A. *et al.* Modelling the impact of antibiotic use and infection control practices on the incidence of hospital-acquired methicillin-resistant Staphylococcus aureus: a time-series analysis. *J. Antimicrob. Chemother.* **62**, 593–600 (2008).

107. Porta, A. *et al.* Model-based assessment of baroreflex and cardiopulmonary couplings during graded head-up tilt. *Comput. Biol. Med.* **42**, 298–305 (2012).

108. Mukkamala, R., Mathias, J. M., Mullen, T. J., Cohen, R. J. & Freeman, R. System identification of closed-loop cardiovascular control mechanisms: diabetic autonomic neuropathy. *Am. J. Physiol.* **276**, R905-12 (1999).

109. Korhonen, I. Multivariate closed-loop model for analysis of cardiovascular dynamics. *Methods Inf. Med.* **36**, 264–7 (1997).

110. Costa, M., Goldberger, A. L. & Peng, C.-K. Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* **89**, 068102 (2002).

111. Ivanov, P. C. *et al.* Multifractality in human heartbeat dynamics. *Nature* **399**, 461–5 (1999).

112. Frey, U. & Suki, B. Complexity of chronic asthma and chronic obstructive pulmonary disease: implications for risk assessment, and disease progression and control. *Lancet (London, England)* **372**, 1088–99 (2008).

113. Goldberger, A. L., Peng, C.-K. & Lipsitz, L. A. What is physiologic complexity and how does it change with aging and disease? *Neurobiol. Aging* **23**, 23–6

114. Suki, B., Bates, J. H. T. & Frey, U. Complexity and emergent phenomena. *Compr. Physiol.* **1**, 995–1029 (2011).

115. Frey, U., Maksym, G. & Suki, B. Temporal complexity in clinical manifestations of lung disease. *J. Appl. Physiol.* **110**, 1723–31 (2011).

116. Goldberger, A. L., Rigney, D. R. & West, B. J. Chaos and fractals in human physiology. *Sci. Am.* **262**, 42–9 (1990).

117. West, B. J. *et al.* Fractal fluctuations in cardiac time series. *Physica A* **270**, 552–66 (1999).

118. Ivanov, P. C. *et al.* Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis. *Nature* **383**, 323–7 (1996).

119. Peng, C. K. *et al.* Statistical properties of DNA sequences. *Physica A* **221**, 180–92 (1995).

120. Peng, C. -K. K., Havlin, S., Stanley, H. E. & Goldberger, A. L. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos* **5**, 82–7 (1995).

121. Frey, U. *et al.* Risk of severe asthma episodes predicted from fluctuation analysis of airway function. *Nature* **438**, 667–70 (2005).
122. Donaldson, G. C., Seemungal, T. A. R., Hurst, J. R. & Wedzicha, J. A. Detrended fluctuation analysis of peak expiratory flow and exacerbation frequency in COPD. *Eur. Respir. J.* **40**, 1123–9 (2012).
123. Frey, U. Predicting asthma control and exacerbations: chronic asthma as a complex dynamic model. *Curr. Opin. Allergy Clin. Immunol.* **7**, 223–30 (2007).
124. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461**, 53–9 (2009).
125. Scheffer, M. *et al.* Anticipating critical transitions. *Science* **338**, 344–8 (2012).
126. Guttal, V., Raghavendra, S., Goel, N. & Hoarau, Q. Lack of Critical Slowing Down Suggests that Financial Meltdowns Are Not Critical Transitions, yet Rising Variability Could Signal Systemic Risk. *PLoS One* **11**, e0144198 (2016).
127. Quax, R., Kandhai, D. & Sloot, P. M. A. Information dissipation as an early-warning signal for the Lehman Brothers collapse in financial time series. *Sci. Rep.* **3**, 1898 (2013).
128. Liu, R. *et al.* Identifying critical transitions of complex diseases based on a single sample. *Bioinformatics* **30**, 1579–86 (2014).
129. Guiot, J., Moermans, C., Henket, M., Corhay, J.-L. & Louis, R. Blood Biomarkers in Idiopathic Pulmonary Fibrosis. *Lung* **195**, 273–280 (2017).
130. Miller, M. R. *et al.* Standardisation of spirometry. *Eur. Respir. J.* **26**, 319–38 (2005).
131. Seemungal, T. A. *et al.* Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **157**, 1418–22 (1998).
132. Donaldson, G. C., Seemungal, T. A. R., Bhowmik, A. & Wedzicha, J. A. Relationship between exacerbation frequency and lung function decline in chronic obstructive pulmonary disease. *Thorax* **57**, 847–52 (2002).
133. Donaldson, G. C. *et al.* Longitudinal changes in the nature, severity and frequency of COPD exacerbations. *Eur. Respir. J.* **22**, 931–6 (2003).
134. Quint, J. K. *et al.* Predictive accuracy of patient-reported exacerbation frequency in COPD. *Eur. Respir. J.* **37**, 501–7 (2011).
135. García-Río, F., Pino, J. M., Dorgham, A., Alonso, A. & Villamor, J. Spirometric reference equations for European females and males aged 65-85 yrs. *Eur. Respir. J.* **24**, 397–405 (2004).
136. Jebb, A. T., Tay, L., Wang, W. & Huang, Q. Time series analysis for psychological research: examining and forecasting change. *Front. Psychol.* **6**, 727 (2015).
137. Greenland, S. & Finkle, W. D. A critical look at methods for handling missing covariates

in epidemiologic regression analyses. *Am. J. Epidemiol.* **142**, 1255–64 (1995).

138. Moritz, S., Sard  , A., Bartz-Beielstein, T., Zaefferer, M. & Stork, J. Comparison of different Methods for Univariate Time Series Imputation in R. *CoRR* **abs/1510.0**, (2015).
139. Wissel, C. A universal law of the characteristic return time near thresholds. *Oecologia* **65**, 101–107 (1984).
140. Dakos, V. *et al.* Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLoS One* **7**, e41010 (2012).
141. Hambly, N., Cox, G. & Kolb, M. Acute exacerbations of idiopathic pulmonary fibrosis: tough to define; tougher to manage. *Eur. Respir. J.* **49**, (2017).
142. Liu, R., Chen, P., Aihara, K. & Chen, L. Identifying early-warning signals of critical transitions with strong noise by dynamical network markers. *Sci. Rep.* **5**, 17501 (2015).
143. Vieth, E. Fitting piecewise linear regression functions to biological responses. *J. Appl. Physiol.* **67**, 390–6 (1989).
144. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
145. Martinez, F. J. *et al.* The clinical course of patients with idiopathic pulmonary fibrosis. *Ann. Intern. Med.* **142**, 963–7 (2005).
146. Chung, J. H. & Lynch, D. A. The Value of a Multidisciplinary Approach to the Diagnosis of Usual Interstitial Pneumonitis and Idiopathic Pulmonary Fibrosis: Radiology, Pathology, and Clinical Correlation. *AJR. Am. J. Roentgenol.* **206**, 463–71 (2016).
147. Fernández Pérez, E. R. *et al.* Incidence, prevalence, and clinical course of idiopathic pulmonary fibrosis: a population-based study. *Chest* **137**, 129–37 (2010).
148. Silva, C. I. S. *et al.* Acute exacerbation of chronic interstitial pneumonia: high-resolution computed tomography and pathologic findings. *J. Thorac. Imaging* **22**, 221–9 (2007).
149. Karimi-Shah, B. A. & Chowdhury, B. A. Forced vital capacity in idiopathic pulmonary fibrosis--FDA review of pirfenidone and nintedanib. *N. Engl. J. Med.* **372**, 1189–91 (2015).
150. Ryerson, C. J. *et al.* Depression and functional status are strongly associated with dyspnea in interstitial lung disease. *Chest* **139**, 609–616 (2011).
151. Vittinghoff, E., Glidden, D. V, Shiboski, S. C. & McCulloch, C. E. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. (Springer Science & Business Media, 2011).
152. Ard, M. C. & Edland, S. D. Power calculations for clinical trials in Alzheimer's disease.

J. Alzheimers. Dis. **26 Suppl 3**, 369–77 (2011).

153. Dodge, H. H. *et al.* Use of High-Frequency In-Home Monitoring Data May Reduce Sample Sizes Needed in Clinical Trials. *PLoS One* **10**, e0138095 (2015).
154. Castaldi, P. J. *et al.* Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. *Am. J. Respir. Crit. Care Med.* **188**, 1083–90 (2013).
155. Hersh, C. P. *et al.* Paired inspiratory-expiratory chest CT scans to assess for small airways disease in COPD. *Respir. Res.* **14**, 42 (2013).
156. Estépar, R. S. J. *et al.* Computed tomographic measures of pulmonary vascular morphology in smokers and their clinical implications. *Am. J. Respir. Crit. Care Med.* **188**, 231–9 (2013).
157. Liou, T. G. & Kanner, R. E. Spirometry. *Clin. Rev. Allergy Immunol.* **37**, 137–52 (2009).
158. Belloli, E. A. *et al.* Longitudinal Forced Vital Capacity Monitoring as a Prognostic Adjunct after Lung Transplantation. *Am. J. Respir. Crit. Care Med.* **192**, 209–18 (2015).
159. Zeger, S. L., Irizarry, R. & Peng, R. D. On time series analysis of public health and biomedical data. *Annu. Rev. Public Health* **27**, 57–79 (2006).
160. Imhoff, M. & Bauer, M. Time series analysis in critical care monitoring. *New Horiz.* **4**, 519–31 (1996).
161. Peng, C. K. *et al.* Quantifying fractal dynamics of human respiration: age and gender effects. *Ann. Biomed. Eng.* **30**, 683–92 (2002).
162. Goldberger, A. L. *et al.* Fractal dynamics in physiology: alterations with disease and aging. *Proc. Natl. Acad. Sci. U. S. A.* **99 Suppl 1**, 2466–72 (2002).
163. Eke, A. *et al.* Physiological time series: distinguishing fractal noises from motions. *Pflugers Arch.* **439**, 403–15 (2000).
164. Veiga, J., Lopes, A. J., Jansen, J. M. & Melo, P. L. Airflow pattern complexity and airway obstruction in asthma. *J. Appl. Physiol.* **111**, 412–9 (2011).
165. Shirazi, A. H. *et al.* Quantifying memory in complex physiological time-series. *PLoS One* **8**, e72854 (2013).
166. Raoufy, M. R. *et al.* Classification of Asthma Based on Nonlinear Analysis of Breathing Pattern. *PLoS One* **11**, e0147976 (2016).
167. Höll, M. & Kantz, H. The fluctuation function of the detrended fluctuation analysis --- investigation on the AR(1) process. *Eur. Phys. J. B* **88**, 126 (2015).
168. Schmidt, S. L. *et al.* Predicting pulmonary fibrosis disease course from past trends in pulmonary function. *Chest* **145**, 579–585 (2014).
169. Jegal, Y. *et al.* Physiology is a stronger predictor of survival than pathology in fibrotic

interstitial pneumonia. *Am. J. Respir. Crit. Care Med.* **171**, 639–44 (2005).

170. Ryerson, C. J., Cottin, V., Brown, K. K. & Collard, H. R. Acute exacerbation of idiopathic pulmonary fibrosis: shifting the paradigm. *Eur. Respir. J.* **46**, 512–20 (2015).
171. Steele, M. P. *et al.* Clinical and pathologic features of familial interstitial pneumonia. *Am. J. Respir. Crit. Care Med.* **172**, 1146–52 (2005).
172. Thabut, G., Crestani, B., Porcher, R. & Richeldi, L. Missing data in IPF trials: do not let methodological issues undermine a major therapeutic breakthrough. *Eur. Respir. J.* **46**, 607–14 (2015).
173. Brockwell, P. J., Davis, R. A. & Calder, M. V. *Introduction to time series and forecasting*. **2**, (Springer, 2002).
174. Leuschner, G. & Behr, J. Acute Exacerbation in Interstitial Lung Disease. *Front. Med.* **4**, 176 (2017).
175. Frey, U. & Bielicki, J. A. Fluctuation Metrics as Novel Endpoints for Clinical Trials in Asthma. *Am. J. Respir. Crit. Care Med.* **195**, 967–968 (2017).
176. McCraty, R. & Shaffer, F. Heart Rate Variability: New Perspectives on Physiological Mechanisms, Assessment of Self-regulatory Capacity, and Health risk. *Glob. Adv. Heal. Med.* **4**, 46–61 (2015).
177. Kamen, P. W., Krum, H. & Tonkin, A. M. Poincaré plot of heart rate variability allows quantitative display of parasympathetic nervous activity in humans. *Clin. Sci. (Lond).* **91**, 201–8 (1996).
178. Woo, M. A., Stevenson, W. G., Moser, D. K., Trelease, R. B. & Harper, R. M. Patterns of beat-to-beat heart rate variability in advanced heart failure. *Am. Heart J.* **123**, 704–10 (1992).
179. Huikuri, H. V *et al.* Abnormalities in beat-to-beat dynamics of heart rate before the spontaneous onset of life-threatening ventricular tachyarrhythmias in patients with prior myocardial infarction. *Circulation* **93**, 1836–44 (1996).
180. Mäkikallio, T. H. *et al.* Dynamic analysis of heart rate may predict subsequent ventricular tachycardia after myocardial infarction. *Am. J. Cardiol.* **80**, 779–83 (1997).
181. Stein, P. K., Domitrovich, P. P., Huikuri, H. V, Kleiger, R. E. & Cast Investigators. Traditional and nonlinear heart rate variability are each independently associated with mortality after myocardial infarction. *J. Cardiovasc. Electrophysiol.* **16**, 13–20 (2005).
182. Brennan, M., Palaniswami, M. & Kamen, P. Do existing measures of Poincaré plot geometry reflect nonlinear features of heart rate variability? *IEEE Trans. Biomed. Eng.* **48**, 1342–7 (2001).
183. Johannson, K. A. & Ryerson, C. J. What Gets Measured Gets Managed: Daily Home

Spirometry in Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **194**, 926–927 (2016).

184. Petukhova, T., Ojkic, D., McEwen, B., Deardon, R. & Poljak, Z. Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada. *PLoS One* **13**, e0198313 (2018).
185. Venegas, J. G. *et al.* Self-organized patchiness in asthma as a prelude to catastrophic shifts. *Nature* **434**, 777–82 (2005).
186. Lenton, T. M. *et al.* Tipping elements in the Earth's climate system. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1786–93 (2008).
187. May, R. M., Levin, S. A. & Sugihara, G. Complex systems: ecology for bankers. *Nature* **451**, 893–5 (2008).
188. van de Leemput, I. A. *et al.* Critical slowing down as early warning for the onset and termination of depression. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 87–92 (2014).
189. Veraart, A. J. *et al.* Recovery rates reflect distance to a tipping point in a living system. *Nature* **481**, 357–9 (2011).
190. Dai, L., Vorselen, D., Korolev, K. S. & Gore, J. Generic indicators for loss of resilience before a tipping point leading to population collapse. *Science* **336**, 1175–7 (2012).
191. Dakos, V., van Nes, E. H., D'Odorico, P. & Scheffer, M. Robustness of variance and autocorrelation as indicators of critical slowing down. *Ecology* **93**, 264–71 (2012).
192. Moritz, S. & Bartz-Beielstein, T. imputeTS: Time Series Missing Value Imputation in R. *R J.* **9**, 207–218 (2017).
193. R Core Team. R: A Language and Environment for Statistical Computing. (2018).
194. Fishman, M. *et al.* A method for analyzing temporal patterns of variability of a time series from Poincare plots. *J. Appl. Physiol.* **113**, 297–306 (2012).
195. Karmakar, C. K., Khandoker, A. H., Gubbi, J. & Palaniswami, M. Complex correlation measure: a novel descriptor for Poincaré plot. *Biomed. Eng. Online* **8**, 17 (2009).
196. Liu, R. *et al.* Identifying critical transitions and their leading biomolecular networks in complex diseases. *Sci. Rep.* **2**, 813 (2012).
197. Chen, L., Liu, R., Liu, Z.-P., Li, M. & Aihara, K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* **2**, 342 (2012).
198. Achiron, A. *et al.* Microarray analysis identifies altered regulation of nuclear receptor family members in the pre-disease state of multiple sclerosis. *Neurobiol. Dis.* **38**, 201–

9 (2010).

199. Akhtar, S. & Rozi, S. An autoregressive integrated moving average model for short-term prediction of hepatitis C virus seropositivity among male volunteer blood donors in Karachi, Pakistan. *World J. Gastroenterol.* **15**, 1607–12 (2009).
200. Linden, A., Adams, J. L. & Roberts, N. Evaluating disease management program effectiveness: an introduction to time-series analysis. *Dis. Manag.* **6**, 243–55 (2003).
201. Hyndman, R. J. *et al.* Automatic time series forecasting: the forecast package for {R}. *J. Stat. Softw.* **26**, 1–22 (2008).
202. Aho, K., Derryberry, D. & Peterson, T. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* **95**, 631–6 (2014).
203. Goshvarpour, A. & Goshvarpour, A. Poincare indices for analyzing meditative heart rate signals. *Biomed. J.* **38**, 229–34
204. Maher, T. M. *et al.* Pirfenidone in patients with unclassifiable progressive fibrosing interstitial lung disease: a double-blind, randomised, placebo-controlled, phase 2 trial. *Lancet. Respir. Med.* **8**, 147–157 (2020).
205. Voss, A. *et al.* Comparison of nonlinear methods symbolic dynamics, detrended fluctuation, and Poincare plot analysis in risk stratification in patients with dilated cardiomyopathy. *Chaos* **17**, 015120 (2007).
206. Nayak, S. K., Bit, A., Dey, A., Mohapatra, B. & Pal, K. A Review on the Nonlinear Dynamical System Analysis of Electrocardiogram Signal. *J. Healthc. Eng.* **2018**, 6920420 (2018).
207. Vollmer, W. M., McCamant, L. E., Johnson, L. R. & Buist, A. S. Long-term reproducibility of tests of small airways function. Comparisons with spirometry. *Chest* **98**, 303–7 (1990).
208. Wells, A. U. & Denton, C. P. Interstitial lung disease in connective tissue disease--mechanisms and management. *Nat. Rev. Rheumatol.* **10**, 728–39 (2014).
209. Plantier, L. *et al.* Physiology of the lung in idiopathic pulmonary fibrosis. *Eur. Respir. Rev.* **27**, (2018).
210. Studer, S. M. & Kaminski, N. Towards systems biology of human pulmonary fibrosis. *Proc. Am. Thorac. Soc.* **4**, 85–91 (2007).
211. Trefois, C., Antony, P. M. A., Goncalves, J., Skupin, A. & Balling, R. Critical transitions in chronic disease: transferring concepts from ecology to systems medicine. *Curr. Opin. Biotechnol.* **34**, 48–55 (2015).
212. Liu, R., Aihara, K. & Chen, L. Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes. *Quant. Biol.* **1**, 105–114

(2013).

213. Wah, W. *et al.* Time series analysis of demographic and temporal trends of tuberculosis in Singapore. *BMC Public Health* **14**, 1121 (2014).

214. Sakamoto, K. *et al.* Acute exacerbation of idiopathic pulmonary fibrosis as the initial presentation of the disease. *Eur. Respir. Rev.* **18**, 129–32 (2009).

215. Bernal, J. L., Cummins, S. & Gasparrini, A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int. J. Epidemiol.* **46**, 348–355 (2017).

216. Sornette, D. Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes, and human birth. *Proc. Natl. Acad. Sci. U. S. A.* **99 Suppl 1**, 2522–9 (2002).

217. Turcotte, D. L. & Rundle, J. B. Self-organized complexity in the physical, biological, and social sciences. *Proc. Natl. Acad. Sci. U. S. A.* **99 Suppl 1**, 2463–5 (2002).

218. Su, R.-Q., Ni, X., Wang, W.-X. & Lai, Y.-C. Forecasting synchronizability of complex networks from data. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **85**, 056220 (2012).

219. Richeldi, L. *et al.* Design of the INPULSIS™ trials: two phase 3 trials of nintedanib in patients with idiopathic pulmonary fibrosis. *Respir. Med.* **108**, 1023–30 (2014).

220. Hagmeyer, L., Treml, M., Priegnitz, C. & Randerath, W. J. Successful Concomitant Therapy with Pirfenidone and Nintedanib in Idiopathic Pulmonary Fibrosis: A Case Report. *Respiration.* **91**, 327–32 (2016).

221. Travis, W. D. *et al.* An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. *Am. J. Respir. Crit. Care Med.* **188**, 733–48 (2013).

222. Paterniti, M. O. *et al.* Acute Exacerbation and Decline in Forced Vital Capacity Are Associated with Increased Mortality in Idiopathic Pulmonary Fibrosis. *Ann. Am. Thorac. Soc.* **14**, 1395–1402 (2017).

223. Demedts, M. *et al.* High-dose acetylcysteine in idiopathic pulmonary fibrosis. *N. Engl. J. Med.* **353**, 2229–42 (2005).

224. Hunninghake, G. M. A new hope for idiopathic pulmonary fibrosis. *N. Engl. J. Med.* **370**, 2142–3 (2014).

225. Wakamatsu, K. *et al.* Prognostic value of serial serum KL-6 measurements in patients with idiopathic pulmonary fibrosis. *Respir. Investig.* **55**, 16–23 (2017).

226. Hesselstrand, R. *et al.* Biomarkers from bronchoalveolar lavage fluid in systemic sclerosis patients with interstitial lung disease relate to severity of lung fibrosis. *Respir. Med.* **107**, 1079–86 (2013).

227. Chen, P., Liu, R., Li, Y. & Chen, L. Detecting critical state before phase transition of complex biological systems by hidden Markov model. *Bioinformatics* **32**, 2143–50 (2016).