UNIVERSITAT DE BARCELONA

Multi-Level Integrated Analysis of Chronic Obstructive Pulmonary Disease (COPD) heterogeneity

Guillaume Noell

MULTI-LEVEL INTEGRATED ANALYSIS OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD) HETEROGENEITY

Guillaume Noell

MSc in Bioinformatics for Health Sciences

PhD co-directors:
Dra. Maria Rosa Faner Canet
Dr. Àlvar Agustí

PhD program in biomedicine

Department of medicine

University of Barcelona (UB)

January 15th 2019

Thesis co-director	Thesis co-director	Doctoral student
Dra. Maria Rosa Faner Canet	Dr. Àlvar Agustí	Guillaume Noell

Acknowledgements

First, I would like to express my utmost gratitude to Dra. Maria Rosa Faner and Dr. Àlvar Agustí for their guidance, hard work and the trust that have put in me from the very start of the PhD. I sincerely feel privileged to have had them as my supervisors. Without sacrificing to the pursuit of scientific excellence, they have mastered the art of encouragement, positive reinforcements, honest feedback and human connection which I valued immensely and made my PhD trouble-free. They are experts in their field and exemplary researchers, always looking out to foster positive collaborations between scientists with the utmost consideration for scientific truth and the progress of medicine. I will miss working with them and will remember their supervision as an example to follow for the remainder of my professional career.

Besides my directors, I would like to thank the members of my thesis defense committee: Dr. Ferran Sanz Carreras, Dr. Josep Roca Torrent, and Dra. Esther Barreiro.

I would also like to thank my colleagues and lab mates for their unwavering enthusiastic support, assistance and kindness.

Last but not least, I am most happy to thank my family, my parents, brothers, sister and grandparents. They have supported me wholeheartedly throughout my studies and life choices, shaped my core values, and always provided me with the best material and emotional conditions. They encouraged me to aim high in whichever path I personally chose to pursue. I owe them much of my accomplishments.

Table of Contents

Abbreviations

A1PI or A1AP: Alpha1–Proteinase Inhibitor
ACO: co-occurrence of Asthma and Chronic Obstructive Pulmonary Disease
AUC: Area Under the Score
BAFFs: B cell Activating Factor belonging to the tumor necrosis factor (TNF) Family
BMI: Body Mass Index
COPD: Chronic Obstructive Pulmonary Disease
CRP: C-Reactive Protein
DLCO: Diffusing Capacity of the Lung for Carbon Monoxide
DPPIV: Dipeptidyl Peptidase IV
ECM: ExtraCellular Matrix
ECOPD: Exacerbations of Chronic Obstructive Pulmonary Disease
ELLF: Early adulthood Low peak Lung Function
ENLF: Early adulthood Normal peak Lung Function
FDR: False Discovery Rate
FEV1: Forced Expiratory Volume in 1^{st} second
FFMI: Fat-Free Mass Index
FOC: Framingham Offspring Cohort
FVC: Forced Vital Capacity
FiO2: Fraction of Inspired Oxygen
GDF-15: Differentiation Factor-15
GOLD: Global Initiative for Obstructive Lung Disease
HDACs: Histone Deacetylases
HGP: Human Genome Project
HR: Hazard Ratio
IgE: Immunoglobulin E
KCO: Carbon Monoxide transfer coefficient
LAMAs: Long-Acting Muscarinic Antogonists
LFs: Lymphoid Follicles
LVRS: Lung Volume Reduction Surgery
MLDNA: Multi-Level Differential Network Analysis
MM: Module Modularity
MMPs: Matrix-Mettaloproteinases
NCDs: Non-Communicable Diseases
NF-κB: transcription factor nuclear factor
NT-proBNP: N-Terminal pro B-type Natriuretic peptide
PAFI: PaO2/FiO2
PDE4: Phosphodiesterase 4
PFT: Pulmonary Function Test
PPMs: Potential Pathogenic Microorganisms

PaO2: Partial Pressure of Oxygen
ROC: Receiver operating Characteristic
SAA: Serum Amyloid A
SABA: Short-Acting inhaled Beta-Agonists
SAMA: Short-Acting Muscarinic Antagonist
SLPI: leukocyte Protease Inhibitor
SP-D: Human surfactant Protein D
TNF-α: Tumour Necrosis Factor-alpha
VmaxFRC: maximal expiratory flows at Functional Residual Capacity
iNOs: inducible Nitric Oxide synthase
mMRC: modified British Medical Research Council

Introduction

1. NON-COMMUNICABLE DISEASES

1.1 Global Prevalence

Non-Communicable Diseases (NCDs) are chronic diseases that result from a combination of genetic, physiological, environmental and behavioral factors [1]. The main types of NCDs are cancer, cardiovascular and cerebrovascular diseases, chronic obstructive pulmonary disease (COPD), asthma and metabolic diseases (diabetes). NCDs are a major global health problem of the 21st century [1]. They are estimated to represent 63% of global annual deaths according to the World Health Organization (WHO) [2, 3]. They are known to be by-and-large preventable with the appropriate management of their principal risk factors at an individual level throughout life: tobacco smoking, alcohol abuse, physical inactivity and unhealthy dieting. Specifically, the WHO estimates that up to 40% of cancers and 75% of heart diseases, stroke and type 2 diabetes could be prevented. Unfortunately, 80% of NCDs deaths occur in low- and middle-income countries [2] where individuals lack preventive information, early detection, access to healthcare and the economic resources to minimize the risk factors or afford treatment.

1.2 NCDs are Complex and Heterogeneous Conditions

NCDs are caused by complex gene-environment interactions that develop over years or decades (thus are associated with aging) and often co-exist in the same individual as they share risk factors [4] and pathological mechanisms (leading to what is known as multimorbidity) [5]. These cooccurrences lie at the heart of NCDs and make clear-cut singular diagnostics difficult. Their pathobiology is also complex, heterogeneous and may lead to unspecific symptoms. For most NCDs, current available treatments are not able to cure the condition, but rather only alleviate symptoms and slow the disease progression.

NCDs often share major risk factors [4]. Therefore, multimorbidity may be explained by the hypothesis that the progressive abnormal transformation of a biological system (e.g. metabolic or respiratory) that lead to a dysfunctional long-lasting state with observable symptoms is likely to also affect other parts of the organism in its course or to be caused by a

common denominator (e.g. systemic inflammation or impaired immune response, or common susceptibility genes [6]).

2. BIOMEDICAL RESEARCH OF COMPLEX DISEASES

2.1 Historical Perspective

The continuous ageing of the general population worldwide over the last two centuries [7] has caused an increase in the overall incidence of NCDs since they are more prevalent in older individuals. Life expectancy in fact rose from a worldwide average of 32 years in 1850, to 48 years in 1950 and is now, as of 2018, over 70 years (Figure 1), and is associated to three cooccurring factors: the worldwide expansion of modernization and industrialization, general lifestyle improvements in high-income countries (such as overall reduced tobacco smoking [8] and less physical strenuous jobs), and unprecedented progress in experimental medicine.

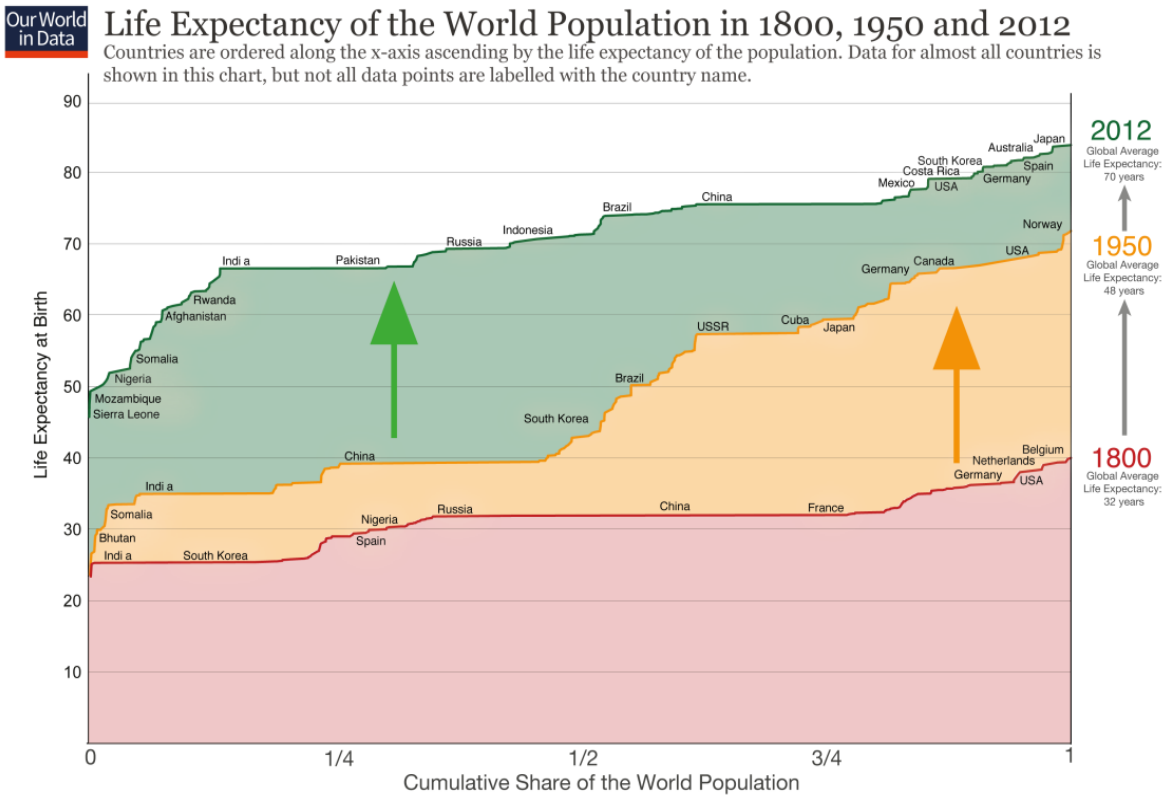

Figure 1. Progress of worldwide life expectancy. Reproduced from https://ourworldindata.org/life-expectancy

Significant scientific advances in our understanding of health and (chronic) diseases since the nineteenth century [9] have been translated into numerous novel treatments,

medication, drugs, surgical procedures and preventive measures that drastically reduced maternal, infancy and elderly mortality (Figure 2). A non-exhaustive list of these innovations [9] range from Louis Pasteur and Robert Koch germ theory of disease in 1870s, to a host of first vaccines in the second half of the 19th century (for cholera, rabies, plague, etc.), as well as the discovery of insulin for diabetes in 1922, the first pacemaker by Paul Zoll in 1952, the first kidney transplant by Dr Jose E. Murray in 1954, the HIV discovery in 1983, the first released draft of the human genome in 2003, the creation of embryonic stem cells from human skin cells in 2007, and the 2014 first FDA-approved US clinical trial for a wearable artificial kidney (Blood Purification Technologies Inc.). The rate of innovations is incrementing swiftly, as corroborated by the double-exponential increase of the biomedical literature in the last 20 years (Figure 2).

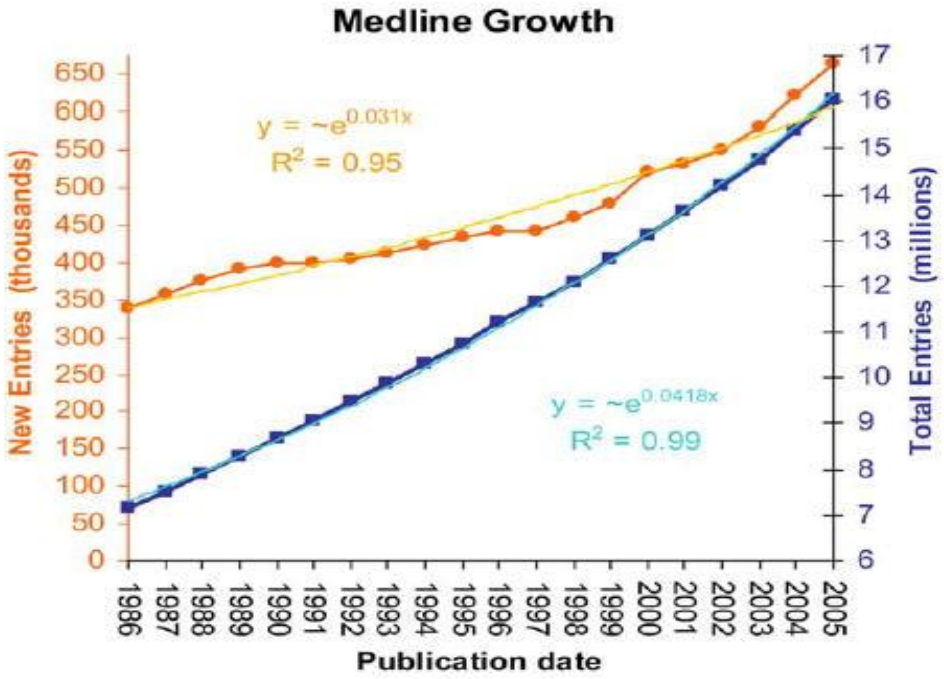

Figure 2. Growth in the Biomedical Literature, 1986–2005, reproduced from ref. [10]

This growth is fueled by increasing worldwide funding for biomedical research (estimated in 2012 at 268 billion of U.S. dollars [11]), and, as mentioned, driven by technological advances and breakthroughs (e.g. internet, which has enabled the fast exchange of information and facilitated scientific collaborations, as well as software and hardware improvements in terms of availability, versatility, power and cost). Nevertheless, all these progressive efforts still remain insufficient as most chronic diseases do not yet have a cure.

2.2 Biotechnological Revolution

Over the last three decades, biomedical research has undergone a fast-paced revolution in methods and scope. Experimental medicine research of NCDs now routinely collects extensive samples data at several biological levels, termed omics, thanks to novel arrays, sequencing and imaging technologies [12], that commonly are genetic (genomics), messenger RNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics). The first international milestone enabled by the interleaving of biology and technology is arguably the Human Genome Project (HGP), which started in 1990 and was declared completed in April 2003 with the release of the first human complete DNA sequence (genome), consisting of 3 billion base pairs, for a total cost of 2.7 billion dollars. Since then, incredible advances in technology and cost reduction have led to the pursuit of the "1000 dollars Genome". It is now a reality in the strict sense if considering only the cost of sequencing. The cost of interpreting the data, however, is still several order of magnitudes higher [13, 14].

Because of the increase in biomedica data size and complexity, many fields of expertise are now necessary to the research efforts on NCDs. The cost of studies is also increasing, partly because the higher the number of biological (omics) determinations characterized, the more samples are required to separate signal from noise and reach statistical significance. Even in the simplified case of a single omic analysis, detecting variants (e.g. genes) that have a different average expression between two conditions (e.g. healthy versus NCD) requires many samples because of the high number of measurements (e.g. up about 10000-50000 genes per sample for a routine transcriptomics array). In statistical terms, because these measured biological variables tend to follow a (normal) distribution of substantial variability, the probability (p-value) that some of them will be significantly differentially expressed by chance (false positives) between any two groups of interest is not negligible. Fortunately, p-values calculations can be corrected for multiple testing [15], e.g. controlling for the relative proportion of false positives to true positives. However, to reach statistical significance, the sample size must be in the order of tens or hundreds of samples for the most complex chronic conditions (or even thousands for exhaustive multi-omics or genome-wide association studies). Collaborations between scientists and research groups have become paramount to cover the scientific expertise and reduce the research costs of these complex studies.

2.3 Systems Medicine, Biostatistics and Bioinformatics

It is plausible to conceptualize human health and NCDs as emergent properties of a complex, non-linear, dynamic multilevel biological system. The existence of heterogeneity as an intrinsic property of a (diseased) biological system implies that the system processes are sufficiently complex for its emergence, and that no isolated part of the system can fully grasp the heterogeneity on its own [16].

As such, the ongoing scientific approach to better understand NCDs like COPD lies in the analysis of the interaction between the many biological components upon which they rest, in an attempt to relate the observed clinical symptoms to their underlying biological (and environmental) systems' parts. These components, or variables, exist as useful abstractions at different conceptual levels, for example organs at the physiological level, proteins at the cellular levels, genes at the (epi-)genetic level, diet/exercise/pollution at the environmental level and so on and so forth [17].

That being said, determining the isolated state of each of these components (e.g. whether an organ is functioning properly or not, how much a single protein is expressed, how healthy the patient's diet is, etc.) fails to capture the disease processes and symptoms, because, as stated, they are emergent [16] properties of the mechanistic interactions between the variables, and not of the isolated variable states by themselves. Systems medicine thus places the dynamic interaction of the parts in a holistic system at the centre of the research approach. Conceptually diseases are understood as abnormal states of a dynamic network of (biological and environmental) interactions.

This NCDs research approach then requires the expression of as many relevant biological components as possible, plus their dynamic interaction, which appears daunting when considering the sheer number of potentially involved genes or genetic variants alone. That is precisely, however, what the exponential progress of (bio)technologies in the last decades has made possible. In parallel, the computational tools required for the task, i.e. bioinformatics and biostatistics algorithms able to process and extract the relevant variability

and processes out of the data, are also the subject of an incredibly fast progress in order to yield powerful mechanistic or predictive models. Network correlation analysis in particular is a novel research approach that is able to unravel the complexity of biological systems [12]. Other useful methods exist, based either on Bayesian statistics, machine learning or matrix factorization [12].

These emerging tools can be divided into biased (also termed supervised) or unbiased (unsupervised) algorithms. Biased algorithms use a priori hypothesis about the data, such as which are the relevant clinical subgroups of a disease and which are the known relationships between variables (e.g. protein-protein interactions), and then identify the variables and mechanisms that best distinguish and describe these subgroups, while unbiased algorithms look for (combinations of) variables that best capture the variance of the data and attempt to cluster patients without leveraging any prior knowledge of their condition. Both analytical strategies have strengths and drawbacks (detailed in table 3 of my systems biology review [12]) that have to be considered when deciding which method is best suited for a particular research question and dataset.

3. COPD: A MAJOR NON-COMMUNICABLE DISEASE

3.1 Epidemiology and Clinical Presentation

Chronic Obstructive Pulmonary Disease (COPD) is currently viewed as a broad diagnostic term that may encompass a continuum of subtypes each characterized by a distinct functional and pathobiological mechanism (endotypes [18]) and is characterized by persistent respiratory symptoms and airflow limitation [19].

COPD global age-standardised prevalence is 9.23% (95% credible interval [CrI]: 8.16%–10.36%) in men and 6.16% (95% CrI: 5.41%–6.95%) in women [20], although it may equalize in the near future, as women are now more exposed to indoor air pollution (from low-income countries biomass fuel used for cooking and heating) [20]. Females appear to be more susceptible to the harmful effects of smoking on lung function [21], and COPD-related deaths in U.S. women have now surpassed those among U.S. men [22].

COPD frequency is increasing worldwide and is projected to be by 2020 the third leading cause of death worldwide. It also represents a major financial burden on countries economies. Direct US healthcare cost was estimated at 29.5 billion dollars in 2010 and is projected to reach 49 billion by 2020 [19, 23], which includes treatment, prevention, detection and rehabilitation. The inability to work cost caused by the disease morbidity and mortality also adds indirect costs to the economy.

Exhaustive and updated diagnostics criteria are established by the Global Initiative for Obstructive Lung Disease (GOLD), which publishes yearly a comprehensive guide for health care professionals [19]. The report also covers treatment recommendations based on severity and disease progression, prevention and management recommendations, medication and therapies review, as well as comorbidities information. The GOLD diagnostic criteria keep updating slightly as clinical research progresses [24]. Additionally, COPD was found to be both regularly misdiagnosed [25] and under-diagnosed [26].

Available treatment options for COPD significantly improve the patient's quality of life, but they arguably mostly operate at the symptoms level, only slow the progression of the disease and are not yet able to restore the lung biological system to a normal healthy and optimal state. Current therapies are not based on biomarkers of specific underlying pathological processes (endotypes) because these are still unknown [12]. In order to provide more effective and personalized therapeutic interventions, as well as to decrease the costs associated to chronic airway diseases, a better understanding of their pathobiology is needed and appropriate patient stratification is required.

3.2 COPD Risk Factors

COPD has been traditionally considered a self-inflicted condition caused by tobacco smoking, that induces an abnormal inflammatory response and accelerates the normal decline of lung function with age [27]. This paradigm is now challenged since recent reports showed that half of patients with spirometrically defined COPD at 60 yrs. of age never had a normal peak lung function in early adulthood [28] (Figure 3), pointing to a dynamic heterogeneity of the natural history of COPD. Furthermore, it is now estimated that 25-45% of COPD patients

never smoked [29]. Of note, however, approximately 75% of individuals with a low peak FEV1 in early adulthood do not develop COPD.

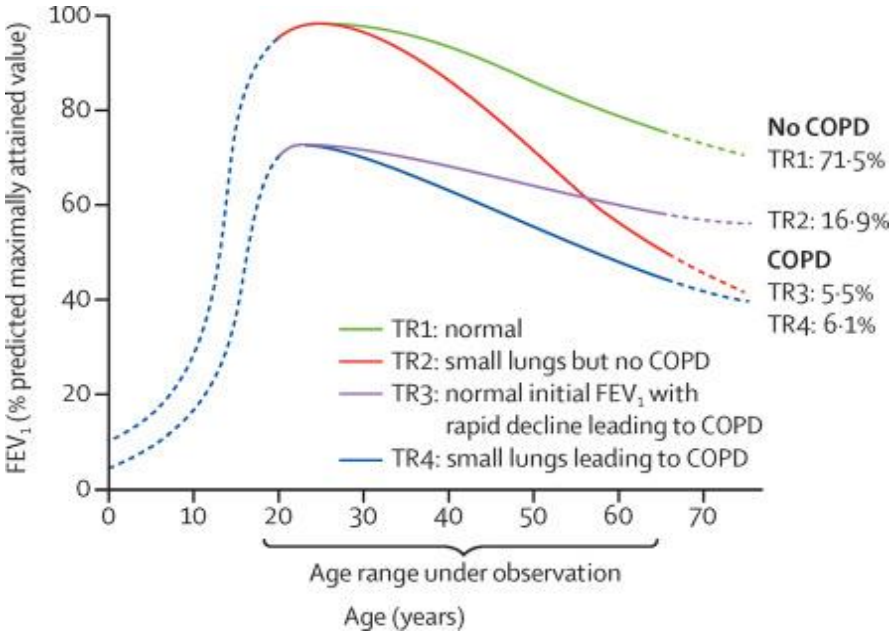

Figure 3. Lung function trajectories leading to COPD reproduced from ref. [30]

More specifically, the following COPD risk factors have been proposed:

Tobacco Smoking and Age

COPD incidence increases with age and is typically diagnosed in individuals older than 40 y.o. (on average at 64 y.o. [31, 32]) who have accumulated years of smoking (commonly measured in pack-years, that is the number of packs of cigarettes smoked per day multiplied by the number of years of smoking). It is estimated that up to 50% of smokers develop COPD [33]. Inversely, currently up to one third of never-smokers meet the COPD criteria [34]. It is worth mentioning that the relative prevalence of COPD never-smokers will increase in developed countries since the proportion of smokers in the general population is decreasing. The effect of smoking is very variable and is tied to the host genetics and immune system [35, 36].

Occupational Exposure to Dusts and Chemicals or to Biomass Fuel

Prolonged exposure to toxic particles for the lung is estimated to be responsible for 15-20% of COPD diagnosis [37]. They are mostly linked to either workplace environments that involve dust, vapours, chemicals, fumes, or household environments that make use of wood smoke, coal or coke open-fires [38]. Both conditions are more prevalent in developing countries due to less stringent protections of employees and less household regulations.

Air Pollution

As mentioned, short-term exposure to air pollution intensifies COPD exacerbations [39], and it generally has adverse effects on COPD symptoms. The influence of chronic exposure to air pollution on COPD is still unclear, although recent cross-sectional studies (on healthy individuals) suggest that it is related to delayed pulmonary function growth in children, and to a faster decline of lung function in adults [40].

Chronic Respiratory Infections

Infections like tuberculosis or HIV are unfortunately still endemic in low and middle-income countries. A meta-analysis evaluates that tuberculosis may double the odds-ratio of chronic airflow obstruction [41], and HIV is a similar risk factor [42]. The inverse is also true as COPD exacerbates the sensitivity to tuberculosis and mycobacterial infections [43, 44]. A history of severe medical illnesses in childhood like respiratory infections and HIV increases COPD risk as well [45].

Genetics

Only about 20% of smokers develop COPD [46], and inversely there is a minority of never-smokers that fit the COPD diagnostic criteria. There is transgenerational association of COPD diagnostic within families [47], so it is likely that genetic (and environmental) factors play a significant role in disease susceptibility. The only endotype of COPD in which the underlying pathobiology is known is due to mutations in the SERPINA1 gene, that cause alpha1–proteinase inhibitor (A1PI) enzyme deficiency [48], and is considered as a different disease entity. Mutations in the SERPINA1 gene account for only 1 to 3% of COPD patients. Since 2009 several genome-wide association studies (GWAS) and meta-analysis have been conducted in several cohorts that include COPD patients [49-52]. Overall these studies have

contributed to the identification of several genomic regions that are associated with COPD at genome-wide significance, including FAM13A, HHIP, CHRNA3/CHRNA5/IREB2, and a region on chromosome 19. Several other genes and gene regions, including ADAM19, FGF7, and SP-D showed evidence for association to develop COPD in smokers. Furthermore, several genes have been associated to the heterogeneity of COPD, for example: i) CHRNA3/5 mutations are associated with cumulative smoking exposure (pack-years), emphysema and airflow limitation [49], ii) HHIP - although not associated with pack-years - is related to FEV_1/FVC ratio, lean body mass and COPD exacerbations in the ECLIPSE cohort [49]; iii) BICD1 SNPs are associated to the presence of emphysema as assessed by radiologist scores [53]. Since variants in BICD1 are correlated with telomere length [53], this observation suggests accelerated aging as a potential mechanism involved in the development of emphysema [54, 55]. It was also found that a significant proportion of emphysema patients have a genetic predisposition for abnormally small telomeres that affects alveolar cells [54], on genes TERT, TR, or NAF1 [56].

Microbiome

Perturbations of the microbiome is an emerging risk factor for both COPD initiation and development [57]. The common characteristic observed in the recent COPD studies is a loss of microbiotic diversity that is correlated to COPD severity, as seen in other non-lung pathological conditions.

Diet

The comparison of dietary elements in terms of preventive and protective effects is generally difficult to investigate due to the lack of relevant longitudinal cohorts data. A 2010 Study of the Hertfordshire Cohort showed by regression analysis that a “prudent” dietary pattern (high consumption of fruit, vegetables, oily fish and wholemeal cereals) is positively associated with FEV1 and FVC in both sexes, and that in males specifically a higher “prudent” pattern score is linked to a higher FEV1/FVC and a lower prevalence of COPD, with associations in males stronger in smokers than non-smokers [58]. A 2016 Spanish cross-sectional study analysis of 207 adult smokers without respiratory disease identified three major dietary patterns from PCA analysis of semi-quantitative food-frequency questionnaire, and then derived from regression analysis that the Mediterranean-like pattern appears to be

associated with preserved lung function, while the Alcohol-consumption pattern and the Westernised pattern are associated with impaired lung function (reduced FEV1, FVC or FEV1/FVC), especially in women [59]. Similarly in 2017, Kaluza J. and colleagues add evidence that high consumption of fruits and vegetables is correlated with reduced COPD incidence in ever-smokers [60], possibly linked to the consumption of antioxidants.

3.3 COPD Heterogeneity

As described above, COPD is currently defined by the presence of chronic airflow limitation [19]. Yet, from the clinical and pathological points of view, we now know that airflow limitation is only one component of COPD [61]. The disease has many other elements that contribute to its clinical presentation, both in the lungs and outside them [62]. As a result, it is often said that COPD is a "complex and heterogeneous disease" [63]. However, in this setting, it is important to define precisely the meaning of words. "*Complex*" means that COPD has several components which display nonlinear interactions between them, whereas "*heterogeneous*" indicates that not all of these components are present in all patients or, in a given patient, at all-time points (i.e., there is *dynamic heterogeneity* [64]). Several examples of this complexity and heterogeneity will be introduced, with special emphasis on exacerbations and comorbidities as they are two aspects that have been investigated in this PhD.

Emphysema and Chronic Bronchitis

The clinical manifestation of COPD can result from a mixture of two pathological processes, emphysema and chronic bronchitis (Figure 4), whose relative proportion vary greatly from patient to patient, evidencing the heterogeneity of the disease.

Emphysema can be broadly defined as impaired alveoli structure (or parenchymal destruction). Alveoli are the tiny air sacs localized in the lungs at the end of the smallest air passages (bronchioles), where the lungs and the bloodstream exchange carbon dioxide and oxygen. Chronic bronchitis refers to inflammation of the bronchial tubes that carry air to and from the alveoli, and is associated with daily cough and mucus production. The presence of emphysema is usually diagnosed by CT scan, and/or impaired diffusing capacity of the lungs

for carbon monoxide (DLCO) [65, 66]. Different pathobiological mechanisms have been postulated for both conditions such as protease/anti-protease imbalance, apoptosis, abnormal immune response and abnormalities in telomeres [67].

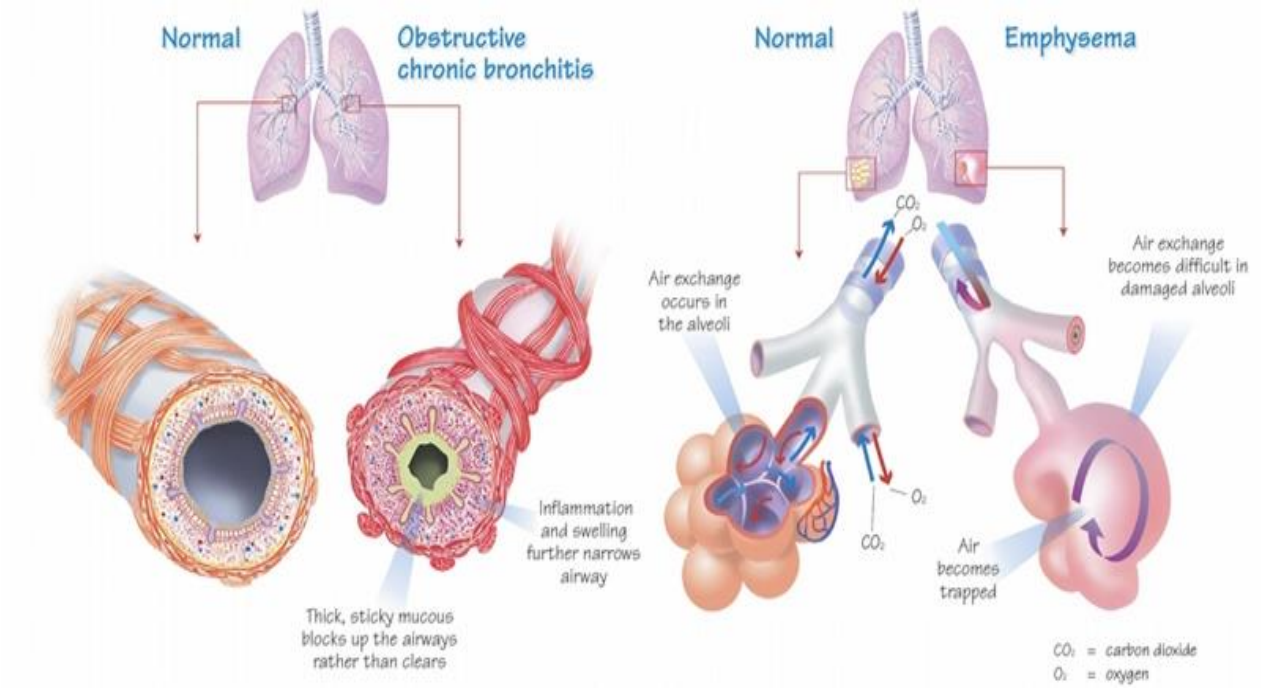

Figure 4. Emphysema versus Chronic Bronchitis. Reproduced from https:// www.livingwellwithcopd.com/en/what-is-copd.html

COPD Exacerbations (ECOPD)

ECOPD are acute episodes of worsening of the symptoms [19], whose frequency is variable and correlates with the disease stage [32]. The episodes are clinically defined by significant lung function alterations, acute dyspnoea or respiratory failure that require special management and hospitalization for the most severe instances. Early signs of exacerbations include feelings of unusual breathlessness, noisy breathing and worse coughing, chest pains, abnormal difficulty in sleeping or eating, changes in skin or nail colour, or fever (in case of infection). Aside from the negative impact in patients regular quality of life, exacerbations also worsen significantly the FEV1 decline and increase the mortality rate [68]. Physiological recovery from an exacerbation do not fully restore patients health, which makes future exacerbations even more frequent. ECOPD are also statistically linked to the incidence of a varied range of comorbidities, such as cardiovascular, cognitive or metabolic chronic disorders, depression, osteoporosis, dysfunctional skeletal myopathy, lung cancer, etc. [69].

Physical fitness, muscle mass, BMI, and diet play a role in the risk of exacerbations. In a 12-months follow-up study of patients hospitalized for acute COPD exacerbation, low initial body mass index (BMI) and weight loss were shown to be risk factors for increased frequency of exacerbations and mortality rate [70]. Peripheral muscle force is also statistically weaker during exacerbations [71]. Finally, daily variations in exposure to outdoor air pollution also intensify the frequency of acute ECOPD [40]. Common biomarkers include plasma or sputum inflammatory mediators (fibrinogen, CRP, tumour necrosis factor-alpha (TNF-α), differentiation factor-15 (GDF-15), interleukins, chemokines) [72] and sub-populations of activated immune cells (decreased CD4+ & CD8+ T cells, increased macrophages and neutrophils) [73]. Neutrophils play not only a role in COPD initiation and inflammatory response but also in exacerbations, in which case their proportion is increased in submucosa and subepithelial tissue [74] and is correlated (r 0.3) with percent FEV1 lost because of the exacerbation [75]. Significant blood eosinophilia (count ≥2%) affects up to 60% of severe exacerbations and airway eosinophilia is increased in 20-40% of exacerbations [76]. These cases respond well to systemic corticosteroid therapy [77, 78].

The pathobiology of exacerbations is an active area of research. It is complex, clearly varies among patients and depends on (epi-)genetic factors, baseline airway inflammation, microbiome, as well as host immunological responses and susceptibility to infections. Most exacerbations are associated to a burst in airway or systemic inflammation that is thought to be caused, for the majority of cases, by respiratory viruses or bacterial species [79], while one third remains of undetermined cause [80]. 58% of viral infections are caused by rhinovirus, while the others comprise human respiratory syncytial virus, coronavirus, influenza virus, parainfluenza virus and adenoviruses [81]. 25% of exacerbations involve coinfection of both viruses and bacteria and recent research suggests that bacterial exacerbation may be precipitated by viruses [82].

A strategy to better manage exacerbations is to distinguish between different clinical subgroups or different pathobiologies so that patients can be treated accordingly. A new "frequent exacerbator" phenotype is now firmly established [83]. These patients are at greater risk of comorbidities and poor health outcomes. They have higher levels of inflammatory

biomarkers (plasma fibrinogen and CRP, sputum interleukin IL-6 and IL-8). In terms of pathophysiology, they are afflicted by increased airway and systemic inflammation, dynamic lung hyperinflation, as well as changes in lower airway bacterial colonization. Arostegui I. et al. identified, from exacerbation variables and past clinical history, four main subgroups of ECOPD patients that have different prognosis, comorbidities, hospitalization and mortality rates [84].

Under-Nutrition and Muscle Mass Wasting

Low Body Mass Index (BMI) and low fat-free mass index (FFMI) are more prevalent in COPD patients (especially in females) than in the general healthy population [85] and are demonstrated to be poor-prognostic factors that can be partly addressed by nutritional supplement therapy [86].

Exercise and Muscle Dysfunction

In relation with low muscle mass, low exercise tolerance affects COPD patients in terms of disease progression; quality of life and mortality rate [86, 87]. Exercise-based pulmonary rehabilitation programme were shown to make a difference in that regard [88].

Comorbidities

More than 80% of COPD patients suffer additional comorbid conditions [89] that are varied and most commonly consist of respiratory, cardiovascular, metabolic and gastrointestinal diseases, as well as lung cancer, osteoporosis, anxiety, depression, skeletal muscle dysfunction, or cachexia. They have significant effects on mortality rate, clinical outcomes and patients quality of life.

Clinically, these conditions share risk factors that explain part of the multimorbidity: smoking and exposure to air pollution in particular are causally associated to many pulmonary and nonpulmonary conditions [90]. Other shared risk factors include early life events (e.g. prematurity [91, 92]), low BMI and physical inactivity [93].

In terms of pathobiology, several conditions share genetic loci for their development (e.g. for COPD and lung cancer [94, 95], or COPD and asthma [96]). A clear biological hallmark of multimorbidities is shared common pathways such as oxidative stress and systemic inflammation. Rubio-Perez C. et al. built networks that combined disease-disease associations, protein-protein interactions as well as gene-disease and variant-disease associations, in order to cluster diseases into related subgroups that internally share genetic alterations and mechanistic (mostly inflammation-based) pathobiological pathways [97]. Similarly, correlation networks analysis by Faner R. et al. added evidence of a shared unspecific molecular diseasome (in particular, mechanisms related to inflammation and vascular tone regulation) to explain the frequent comorbidities occurrence [98].

Research Hypothesis

The **general hypothesis** underlying this PhD Thesis is that the use of multi-level integrated analysis will help us understand holistically highly heterogeneous respiratory diseases such as COPD.

This general hypothesis has been divided in two **specific hypotheses** that correspond to two distinct well defined clinical scenarios.

1) **Exacerbations of COPD**

   ECOPD are highly heterogeneous episodes of worsening of the symptoms with a non-specific diagnosis biomarker, whose pathogenesis and biology is not entirely understood. We hypothesize that the comparison of multi-level (i.e., clinical, physiological, biological, imaging and microbiological) correlation networks determined during ECOPD and clinical recovery can help us identify the key diagnostic biomarkers and features of these highly heterogeneous episodes.

2) **Lung function in early adulthood**

   Low peak lung function in early adulthood, which can result from abnormal lung development, is associated with the diagnosis of COPD later in life. If for any reason the lungs have been poorly developed, it is conceivable that other organs have also done so (e.g. from the cardiovascular or metabolic systems). Accordingly, we hypothesize that abnormal lung development is linked to the impaired development of other organs and systems, and is associated to an increased frequency of subclinical abnormalities and comorbidities in later adulthood.

Objectives

The **general aim** of this PhD Thesis is to apply multi-level integrated analysis to better understand highly heterogeneous respiratory complex diseases such as COPD.

The **specific goals** that have been addressed refer to two specific aspects of COPD heterogeneity:

1) **Exacerbations of COPD (ECOPD**), specific goals:

- To characterize the heterogeneity of ECOPD, using a common set of variables and individuals during the exacerbation phase and at convalescene.
- To integrate and compare the information using Multi-Level Differential Networks.
- To identify ECOPD biomarkers.

2) **Early low lung function and health in later life**, specific goals:

- To determine the prevalence of low peak lung function in early adulthood in the general population.
- To assess the association of early low peak lung function with subclinical abnormalities from the lungs and other organs.
- To evaluate if early low peak lung function is a risk factor for earlier incidence of comorbidities.
- To investigate the relationship between early low peak lung function and later mortality risk.
- To determine the transgenerational reproducibility of early low lung function status.

Results

The core results of this PhD Thesis have been published in the form of two original papers in high impact factor international journals (Eur. Respir. J, IF 2018: 12.2, paper cited 4 times; and the Lancet Respiratory Medicine, IF 2018: 21.5, paper cited 5 times). Besides, the experience gained with this work has also been substantiated in a review paper (Eur. Respir. Rev.) which is presented in the appendix but not discussed directly.

Original Paper 1: Multi-level Differential Network Analysis of COPD Exacerbations

(published in: Noell *et al.* ***Eur. Respir. J.*** **2017**)

Multi-level differential network analysis of COPD exacerbations

Guillaume Noell[1,2,12], Borja G. Cosío[1,3,12], Rosa Faner[1,2,12], Eduard Monsó[1,4], German Peces-Barba[5], Alfredo de Diego[6], Cristobal Esteban[7], Joaquim Gea[1,8], Robert Rodriguez-Roisin[1,2,9], Marian Garcia-Nuñez[1,4], Francisco Pozo-Rodriguez[1,10], Susana G. Kalko[2] and Alvar Agustí[1,2,9,11]

Affiliations: [1]CIBER Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, Barcelona, Spain. [2]Institut d'investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [3]Servei Pneumologia, Hospital Universitari Son Espases-IdISBa, Mallorca, Spain. [4]Servei Pneumologia, Hospital Parc Taulí, Sabadell, Spain. [5]Servicio Neumología, Fundación Jimenez Díaz, Madrid, Spain. [6]Servicio Neumología, Hospital Universitario La Fe, Valencia, Spain. [7]Servicio Neumología, Hospital Galdakao, Bizkaia, Spain. [8]Servei de Pneumologia, Hospital del Mar - IMIM. DCEXS, Universitat Pompeu Fabra, Barcelona, Spain. [9]Departament de Medicina, Universitat de Barcelona, Barcelona, Spain. [10]Hospital Universitario 12 de Octubre, Madrid, Spain. [11]Respiratory Institute, Hospital Clinic, University of Barcelona, Barcelona, Spain. [12]Joint lead authors.

Correspondence: Alvar Agustí, Respiratory Institute, Hospital Clinic, Villarroel 170, 08036 Barcelona, Spain. E-mail: AAGUSTI@clinic.cat

@ERSpublications
This is the first study to investigate COPD exacerbations using multi-level differential network analysis http://ow.ly/uYlW30eMpwR

Cite this article as: Noell G, Cosío BG, Faner R, *et al.* Multi-level differential network analysis of COPD exacerbations. *Eur Respir J* 2017; 50: 1700075 [https://doi.org/10.1183/13993003.00075-2017].

ABSTRACT Patients with chronic obstructive pulmonary disease (COPD) often suffer episodes of exacerbation (ECOPD) that impact negatively the course of their disease. ECOPD are heterogeneous events of unclear pathobiology and non-specific diagnosis. Network analysis is a novel research approach that can help unravelling complex biological systems. We hypothesised that the comparison of multi-level (*i.e.*, clinical, physiological, biological, imaging and microbiological) correlation networks determined during ECOPD and convalescence can yield novel patho-biologic information.

In this proof-of-concept study we included 86 patients hospitalised because of ECOPD in a multicentre study in Spain. Patients were extensively characterised both during the first 72 h of hospitalisation and during clinical stability, at least 3 months after hospital discharge.

We found that 1) episodes of ECOPD are characterised by disruption of the network correlation observed during convalescence; and 2) a panel of biomarkers that include increased levels of dyspnoea, circulating neutrophils and C-reactive protein (CRP) has a high predictive value for ECOPD diagnosis (AUC 0.97).

We conclude that ECOPD 1) are characterised by disruption of network homeokinesis that exists during convalescence; and 2) can be identified objectively by using a panel of three biomarkers (dyspnoea, circulating neutrophils and CRP levels) frequently determined in clinical practice.

This article has supplementary material available from erj.ersjournals.com

Received: Jan 13 2017 | Accepted after revision: June 26 2017

This study is registered at www.clinicaltrials.gov as NCT01750658.

Support statement: Supported, in part, by Fondo de Investigación Sanitaria, Instituto Carlos III (PI 05/1463 and PI 15/00799), CIBERES, an unrestricted grant from GlaxoSmithKline, a PhD scholar grant from the Catalan Government (FI-2016) and CERCA Programme/Generalitat de Catalunya. Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: Disclosures can be found alongside this article at erj.ersjournals.com

Introduction

Patients with chronic obstructive pulmonary disease (COPD) often suffer episodes of exacerbation (ECOPD) that impact negatively their health status and prognosis [1]. The pathogenesis of these episodes is not entirely understood, but it is presumed complex and heterogeneous [2–4]. Their diagnosis relies mostly on symptom perception by the patient [5] and their prevention and treatment is, by and large, empiric [1].

Network analysis is an integrative research strategy well suited for the investigation of heterogeneous and complex diseases [6, 7] such as COPD [8–14]. We hypothesised that multi-level differential network analysis (MLDNA), a novel analytical method that involves the comparison of clinical, physiological, biological, imaging and microbiological (*i.e.* multi-level) correlation networks determined during ECOPD and clinical stability, can provide new insights into the pathobiology and diagnosis of ECOPD [15, 16]. Accordingly, in this proof-of-concept study we used MLDNA, for the first time to our knowledge, to 1) compare the multi-level network structure determined during ECOPD and convalescence; and 2) identify a panel of specific ECOPD biomarkers.

Methods

Methods are described in detail in the supplementary material.

Study design and ethics

This observational, prospective proof-of-concept study was carried out in seven tertiary referral hospitals in Spain (www.clinicaltrials.gov: NCT01750658). Patients were recruited and studied during the first 72 h of hospitalisation because of ECOPD, and investigated again during convalescence, at least 3 months after hospital discharge. The Institutional Review Boards of participating institutions approved the study, and participants gave their informed consent.

Patients

All patients were older than 45 years, current or former smokers (>10 pack-years) and had COPD (and ECOPD) according to the Global Initiative for Chronic Obstructive Lung Disease criteria [1]. In order to homogenise the studied population as much as possible, pneumonia on chest radiography, the presence of severe comorbidity driving the clinical presentation of the patient and/or need of (invasive or noninvasive) mechanical ventilation were exclusion criteria. We initially attempted to recruit patients who had not received oral steroids and/or antibiotic treatment in the community before hospitalisation. Yet, this strategy limited recruitment a lot, so we decided to adopt a more pragmatic design and exclude patients who received oral steroids before hospitalisation (with a potential rapid anti-inflammatory effect) but not those who may have received antibiotics (which may take longer to affect microbiological results). A total of 14 patients (16%) were included in the analysis despite they received antibiotic treatment in the community before hospitalisation.

Measurements

Clinical, functional, biological, microbiological and imaging variables were recorded following standard procedures, as detailed in the supplementary material.

Data analysis

Descriptive statistics

Because many variables were non-normally distributed, the results are presented as median (and 95% confidence intervals) or proportions. Likewise, because not all measurements were available in all patients in both visits, to maximise the potential of available information, the results at ECOPD and convalescence were compared using pairwise statistics (paired Wilcoxon or Chi-squared tests for continuous and discrete variables, respectively). Participants with missing data were discarded on a per-variable basis, such that no value imputation was required. We used false discovery rates (FDRs) to account for multiple comparisons [17]. All analyses were performed using R [18].

Multi-level correlation networks

We built multi-level correlation (Spearman) networks that integrate quantitative and qualitative clinical, functional, biological, microbiological and imaging variables (independently for ECOPD and convalescence) using R [18], and we graphed them with Cytoscape [19].

Module finding

We used the fast-greedy community algorithm to identify network modules on the basis of their module modularity (MM) score, so those with more dense internal connections and fewer external links get higher MM scores [20, 21].

Differential network analysis

To compare multilevel correlation networks at ECOPD and convalescence we 1) nominally contrasted the variables and modules identified under both clinical circumstances; 2) estimated the mean "density" of networks determined at ECOPD and convalescence by comparing (Wilcoxon test) the number of nodes, and the average number of edges per node (node degree, *k*) during ECOPD and convalescence [6]; and, 3) used Monte Carlo permutation tests [22] to identify those Spearman correlations that were significantly different between ECOPD and convalescence.

ECOPD biomarkers

We defined as "outliers" at ECOPD those values below or above the 5th or 95th percentiles, respectively, of the same variable at convalescence, and we identified those ECOPD variables with a significant (bootstrapping FDR <0.05) number of outliers. To identify potential ECOPD biomarkers, we calculated receiver operating characteristic (ROC) curves considering all values determined at ECOPD and convalescence and excluding missing data on a per-variable basis.

Results

We studied 86 patients at ECOPD (mean±SD age of 67±9 years). As shown in figure 1, 19 patients were lost for follow-up, so we could study 67 of them at convalescence. Table 1 presents the main characteristics of participants at both time points.

Observations at ECOPD

Besides the expected observations during ECOPD (dyspnoea, tachypnoea, tachycardia, respiratory failure) some other salient findings were (table 1): 1) elevated blood glucose levels, likely to be in relation to the generalised use of systemic steroids in the management of ECOPD [1, 23]; 2) echography identified the presence of pulmonary hypertension in 21.2% of patients and right chamber enlargement in 19.1%, but no patient suffered heart failure with low ejection fraction; 3) computed tomography (CT) emphysema was present in 56.7% of patients, bronchiectasis in 17.5% and, interestingly, alveolar infiltrates (not seen in chest radiography films) in 23.8%. Pulmonary embolism was found in 1.5% of individuals; and 4) in patients producing spontaneous sputum (77.9%), bacterial culture was positive for potential pathogenic microorganisms (PPMs) in 19.4% of them, whereas viruses were detected by a positive sputum virus in 30.9%. A total of 37.8% of patients were positive for sputum PPMs and/or viruses (table 1). More detailed microbiologic information can be found in the supplementary material.

Changes at convalescence

The main changes from ECOPD to convalescence (highlighted in bold type in table 1) included 1) improved dyspnoea; 2) reduced heart and respiratory rate; 3) reduced serum levels of glucose and urea; 4) improved pulmonary gas exchange without significant changes in spirometric variables; 5) reduced total leukocyte count, with lower circulating neutrophils and higher lymphocyte and eosinophil proportions; 6) reduced concentration of acute phase reactants (C-reactive protein (CRP) and fibrinogen) with increased levels of serum amyloid A (SAA). Other systemic inflammatory markers did not change significantly or changed marginally; and, finally, 7) neither bacterial load, viral load nor inflammatory markers changed significantly.

Multi-level differential network analysis

Figure 2 shows the correlation networks determined at ECOPD and convalescence, and table 2 their quantitative comparison. The main observations were 1) the number of nodes at ECOPD and

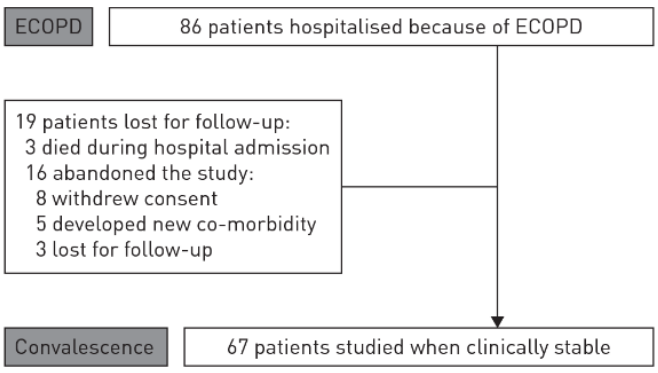

FIGURE 1 Consort diagram of the study.

TABLE 1 Clinical, physiologic, imaging, biological and microbiological data determined during exacerbation of chronic obstructive pulmonary disease (ECOPD) and convalescence

	ECOPD		Convalescence		Pairwise n	FDR p-value
	n	Median (95% CI) or n (%)	n	Median (95% CI) or n (%)		
Vital constants						
Heart rate min^{-1}	**83**	**89 (72–112.5)**	**64**	**83 (69–100.5)**	**64**	**0.004**
Respiratory rate min^{-1}	**82**	**22 (15–30)**	**64**	**20 (16–24.5)**	**64**	**0.000**
Dyspnoea; MMRC scale	**79**	**5 (2–5)**	**63**	**3 (1.5–5)**	**59**	**0.010**
Dyspnoea (1–10); visual scale	**80**	**7 (3.5–8.75)**	**64**	**3 (0.5–6)**	**62**	**0.000**
Body temperature °C	**81**	**36.5 (35.85–37.25)**	**63**	**36 (35.55–36.6)**	**63**	**0.000**
Biochemistry						
Urea $mg{\cdot}dL^{-1}$	**81**	**42 (27–77)**	**65**	**33 (19–53)**	**63**	**0.000**
Creatinine $mg{\cdot}dL^{-1}$	82	0.9 (0.64–1.3)	65	0.9 (0.59–1.13)	64	0.076
Glucose $mg{\cdot}dL^{-1}$	**73**	**169 (101–315)**	**63**	**98 (81–195.5)**	**59**	**0.000**
Haemoglobin $g{\cdot}dL^{-1}$	82	14.8 (12.4–17.1)	65	14.7 (12.25–16.8)	64	0.724
Erythrocyte sedimentation rate $mm{\cdot}h^{-1}$	**64**	**21 (3–67.51)**	**55**	**12 (2–46.5)**	**52**	**0.002**
Lung physiology						
FVC % reference	86	71.5 (51.5–106.5)	63	75 (56–109)	63	1.000
FEV_1 % reference	86	44.2 (25.5–77)	63	46 (28–81.79)	63	0.687
FEV_1/FVC %	86	47.9 (31.05–64.5)	63	49.4 (31–65.5)	63	0.882
IC % reference	69	64 (38.6–94.5)	57	64 (43.7–91.6)	55	1.000
RV % reference	68	169.3 (107.2–245.45)	56	169.1 (108.5–247.35)	53	0.946
TLC % reference	71	115 (82.85–140.5)	59	109.4 (72.18–133.8)	56	0.164
RV/TLC %	57	146 (102.5–192)	40	138.7 (87.3–192)	33	0.914
D_{LCO} % reference	**72**	**56 (31.5–88)**	**52**	**54 (30–84)**	**50**	**0.035**
K_{CO} % reference	71	73 (37.99–111)	57	79 (51–103)	53	0.474
P_{aO_2} mmHg	**82**	**55.4 (41.05–74.1)**	**64**	**66.2 (52.5–91.1)**	**63**	**0.000**
PAFI	**76**	**254.3 (173.55–322.38)**	**64**	**315.2 (247.17–438.57)**	**57**	**0.000**
P_{aCO_2} mmHg	82	43.5 (32.95–69.76)	64	42.7 (34.9–53.35)	63	0.605
Arterial pH	82	7.4 (7.33–7.47)	63	7.4 (7.37–7.46)	62	0.946
6MWD m	67	435 (238.5–536.5)	44	443.5 (274–580)	42	0.280
Cardiovascular physiology						
Creatinine phosphokinase $U{\cdot}L^{-1}$	70	68 (30.5–238.51)	58	69.5 (39–150)	56	0.914
Fibrinogen $mg{\cdot}dL^{-1}$	**74**	**497 (307.5–760)**	**58**	**405 (307.5–574.51)**	**57**	**0.000**
Pro-BNP $pg{\cdot}mL^{-1}$	85	0.4 (0.1–0.78)	67	0.4 (0.04–1.06)	67	0.882
Troponine I % detected above 0.05 $\mu g{\cdot}L^{-1}$	73	6 (8.2%)	60	3 (5.0%)		0.872
Echography						
Left ventricle ejection fraction %	37	68 (42–80.5)	ND	ND		
Right atrial enlargement	68	13 (19.1%)	18	ND		
Pulmonary hypertension	33	7 (21.2%)	ND	ND		
CT imaging						
Emphysema	60	34 (56.7%)	ND	ND		
Bronchiectasis	63	11 (17.5%)	ND	ND		
Alveolar infiltrates	63	15 (23.8%)	ND	ND		
Pulmonary embolism	66	1 (1.5%)	ND	ND		
Lung inflammation (sputum)						
TAS mM	84	0.3 (0.05–1.29)	67	0.2 (0–1.26)	67	0.977
IL-8 $pg{\cdot}mL^{-1}$	84	2146.7 (791.68–2575.22)	67	2128.3 (25.91–2580.18)	67	0.458
IL-1β $pg{\cdot}mL^{-1}$	84	621.5 (73.33–3194.87)	67	504.9 (1.39–2861.79)	67	0.490
IL-6 $pg{\cdot}mL^{-1}$	84	43.2 (3.05–572.87)	67	39.8 (3.05–758.15)	67	0.392
TNF-α $pg{\cdot}mL^{-1}$	83	14.3 (0.52–552.37)	67	6.2 (0.52–345.42)	67	0.450
TGF-β $pg{\cdot}mL^{-1}$	84	0.2 (0.02–2.89)	67	0.3 (0–3.79)	67	0.392
TNF RS $pg{\cdot}mL^{-1}$	84	1.3 (0.04–11.78)	67	2 (0.01–21.61)	67	0.621
SAA $pg{\cdot}mL^{-1}$	84	3.3 (0.38–16.72)	67	2.4 (0.09–14.37)	67	0.392
Systemic inflammation						
Leukocytes $\times 10^3\ \mu L^{-1}$	**82**	**10.9 (6.35–22.59)**	**65**	**8.1 (5.97–12.78)**	**64**	**0.000**
Neutrophils %	**72**	**88.2 (52.35–94.15)**	**58**	**64.8 (38.55–75.7)**	**58**	**0.000**
Lymphocytes %	**82**	**7.4 (3.75–20.25)**	**65**	**22.1 (14–38.4)**	**64**	**0.000**
Eosinophils %	**53**	**0.3 (0–2.45)**	**64**	**2.4 (1.05–7.8)**	**39**	**0.000**
% of patients with eosinophils >2%	**53**	**3 (5.7%)**	**64**	**35 (54.7%)**	**53**	**0.000**
C-reactive protein $mg{\cdot}L^{-1}$	**86**	**3.6 (0.43–16.82)**	**66**	**0.5 (0.09–6.1)**	**66**	**0.000**
Total antioxidant status mM	86	1.5 (0.98–2.41)	67	1.6 (0.79–2.48)	67	0.724

Continued

TABLE 1 Continued

	ECOPD		Convalescence		Pairwise n	FDR p-value
	n	Median (95% CI) or n (%)	n	Median (95% CI) or n (%)		
IL-8 pg·mL^{-1}	84	1 (0.26–4.37)	67	1.3 (0.35–4.53)	67	0.128
IL-1β pg·mL^{-1}	80	0.2 (0.16–0.83)	67	0.2 (0.16–0.75)	67	0.914
IL-6 pg·mL^{-1}	80	0.3 (0.3–7.41)	67	0.3 (0.3–8.81)	67	0.290
TNF-α pg·mL^{-1}	**79**	**0.5 (0.51–1.89)**	**67**	**1.3 (0.51–2.49)**	**67**	**0.000**
Procalcitonin mg·L^{-1}	85	0.4 (0.09–0.93)	67	0.4 (0.03–0.86)	67	0.605
TGF-β pg·mL^{-1}	**86**	**0.8 (0.21–3.55)**	**67**	**1.1 (0.17–6.05)**	**67**	**0.015**
TNF-RS pg·mL^{-1}	86	12.7 (1.36–50.19)	67	19.4 (1.45–71.49)	67	0.256
SAA pg·mL^{-1}	**84**	**0.8 (0.13–4.11)**	**67**	**1.4 (0.13–7.64)**	**67**	**0.002**
Microbiology						
Spontaneous sputum production	86	67 (77.9%)	67	45 (67.2%)		0.513
Positive sputum bacteria (culture)	67	13 (19.4%)	45	12 (26.7%)		0.848
Positive sputum virus (PCR)	55	17 (30.9%)	27	3 (11.1%)		0.394
Positive bacteria (culture) and/or virus (PCR)	74	28 (37.8%)	59	14 (23.7%)		0.753
Adenovirus seroconversion	39	1 (2.6%)	ND	ND		
Chlamydia seroconversion	45	2 (4.4%)	ND	ND		
Influenza seroconversion	43	4 (9.3%)	ND	ND		
Mycoplasma seroconversion	59	1 (1.7%)	ND	ND		
Parainfluenza seroconversion	43	7 (16.3%)	ND	ND		
RSV seroconversion	40	5 (12.5%)	ND	ND		

Values in bold type identify those variables with a statistically significant change from ECOPD to convalescence. (Wilcoxon or Fisher exact tests, corrected for multiple comparison (false discovery rate (FDR), for continuous and categorical variables, respectively). MMRC: modified Medical Research Council; FVC: forced vital capacity; FEV_1: forced expiratory volume in 1 s; IC: inspiratory capacity; RV: residual volume; TLC: total lung capacity; D_{LCO}: carbon monoxide diffusing capacity of the lung; K_{CO}: D_{LCO}/alveolar volume (transfer factor); P_{aO_2}: arterial partial pressure of oxygen; PAFI: P_{aO_2} (mmHg)/inspired fraction of oxygen ratio (%); P_{aCO_2}: arterial partial pressure of carbon dioxide; 6MWD: 6-min walking distance; BNP: brain natriuretic peptide; CT: computed tomography; TAS: total antioxidant status; IL: interleukin; TNF: tumour necrosis factor; TGF: transforming growth factor; SAA: serum amyloid A; TNF-RS: tumour necrosis factor soluble receptor; RSV: respiratory syncytial virus.

convalescence was similar (51 *versus* 47), but the convalescence network was significantly denser, as shown by the higher total number of edges, a significantly higher node degree (*k*), and lower modularity; 2) there were six hubs with a Kleinberg score >0.8 in the ECOPD network and four in the convalescence one. All of the former correspond to sputum inflammatory markers whereas all of the latter correspond to lung function variables; 3) there were five modules at ECOPD and six at convalescence (figure 2, blue areas). All of them appear relatively homogeneous in terms of their biological content, since the majority contained nodes of similar functional category (see colour codes in figure 2). A detailed description of each of these modules is provided in the supplementary material; and 4) the comparison of both networks showed a higher density of significantly different Spearman correlations at convalescence than during ECOPD (table 2 and figure 3) and that more than half of these differential correlations linked different modules (figure 3): at ECOPD, TNF-α was the node with more differential links (n=4) whereas at convalescence these were TGF-β (n=6), K_{CO} (n=5), PAFI (n=5), P_{aO_2} (n=5) and heart rate (n=4). All in all, these observations suggest that the network "perturbation" induced by ECOPD involves a reduction in module co-regulation (*i.e.* co-occurrence).

ECOPD biomarkers: outlier analysis

To investigate potential ECOPD biomarkers, we 1) identified 16 variables (12% of the total number of variables analysed in the study (table 1)) with a significant proportion of ECOPD "outliers", this is a significant (FDR p-value<0.05) proportion of variable values outside the 5th to 95th percentile range of the same variable determined at convalescence (by Monte-Carlo ECOPD/convalescence permutation test on the statistic (% outliers at ECOPD – % outliers at convalescence)); 2) assessed the extent to what these outliers co-occur in the same patients. To this end, we built a co-occurrence network (figure 4) where each node correspond to one of these 16 variables, node size to the proportion of outliers at ECOPD (as indicated by the percentage for each of them), node shape (up or down triangle) indicates if a given variable is higher (up) or lower (down) at ECOPD, and edge colours represent the proportion of co-occurrence between two given nodes (see keys). Circulating lymphocytes and neutrophils were co-altered (albeit in opposite directions) in more than 75% of the exacerbated patients (blue edge), and eosinophils, dyspnoea and glucose levels in 50–75% of patients (green edges; note also the different

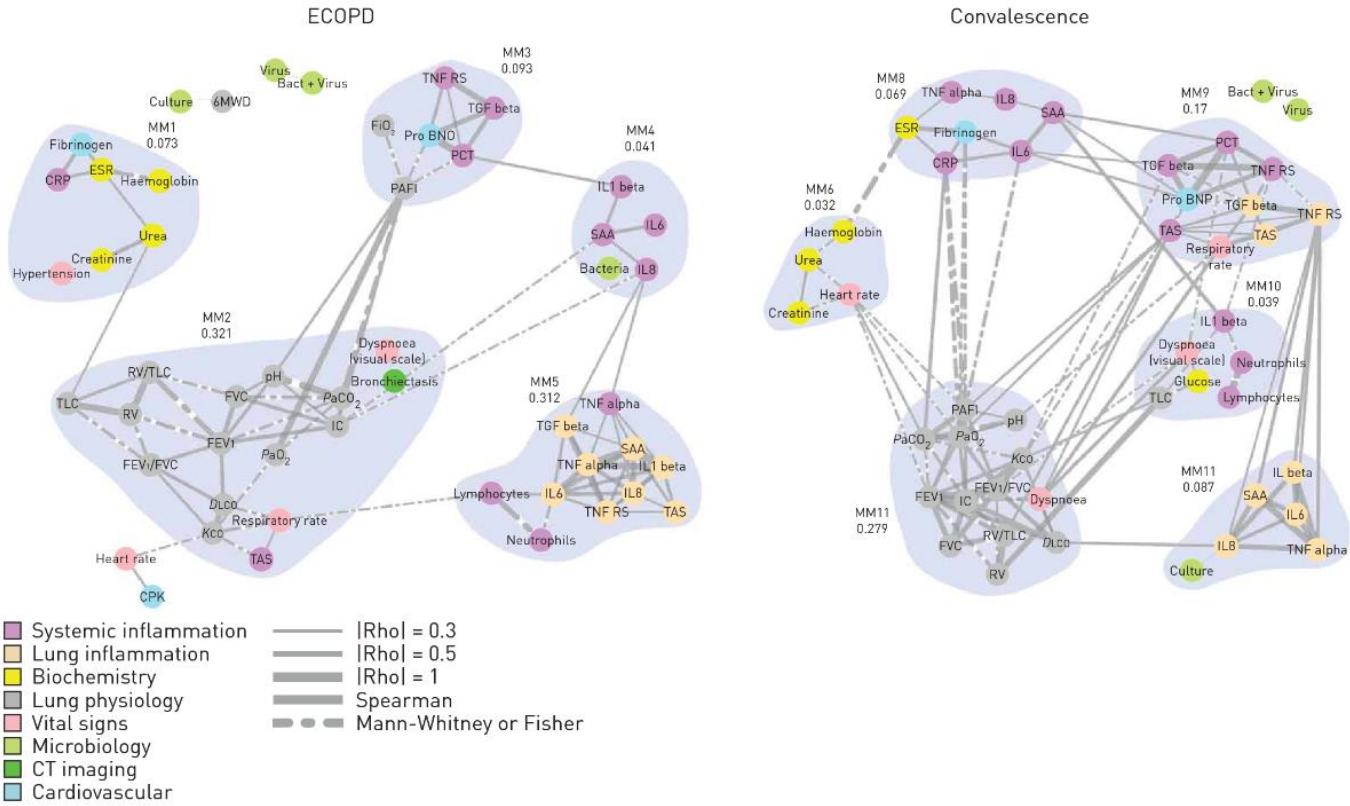

FIGURE 2 Correlation networks at exacerbation of chronic obstructive pulmonary disease (ECOPD) and convalescence. Node colours identify their category (see colour codes). Links between nodes indicate the existence of a statistically significant correlation between them (positive, continuous lines; negative, dashed lines), whereas their width is proportional to the strength of such correlation. Blue shaded areas indicate the different modules (MM) identified by the fast-greedy community algorithm used, so modules with a higher proportion of internal connections get higher MM scores. For further explanations, see text.

triangle shapes). The remaining nodes co-occurred in 25–50% of patients (orange edges); 3) explored the capacity of these 16 variables to predict ECOPD by ROC analysis, and identified a subset of seven of them with an area under the curve (AUC) >0.8. Figure 5a presents the scatter distribution of these seven variables and their individualised ROC profile and AUC (figure 5b); of note, although all of them had a large number of outliers at ECOPD (red symbols), a proportion of values at ECOPD still remained within the 5th to 95th range (horizontal lines) determined at convalescence (figure 5a), likely reflecting the heterogeneity of ECOPD episodes; and, finally, 4) included these seven variables in a general linear mixed model to identify the best diagnostic biomarker panel of ECOPD. We found that the combination of dyspnoea severity, raised circulating neutrophils and elevated CRP levels had an AUC of 0.97 (95% CI 0.95–1) to diagnose ECOPD (figure 5c). Finally we calculated what different combinations of abnormal values of these three variables gave the better specificity, sensitivity, positive and negative prediction values for the diagnosis of an ECOPD (table 3). We observed that dyspnoea levels ⩾5 (on an analogue visual score that ranges from 0 to 10), CRP ⩾3 mg·L^{-1} and ⩾70% circulating neutrophils had a specificity of 0.96, a sensitivity of 0.901, negative predictive value of 0.88 and positive predictive value of 0.97 for the identification of ECOPD.

Discussion

This proof-of-concept study develops and applies for the first time MLDNA to a relevant, complex and heterogeneous clinical problem (ECOPD). By doing so it shows that 1) ECOPD episodes are characterised by fragmentation of the correlation network observed during clinical stability, suggesting loss of system control and reduced resilience during ECOPD [24, 25]; and 2) a panel of biomarkers that includes dyspnoea (⩾5 on an analogue visual score from 0 to 10), CRP level (⩾3 mg·L^{-1}) and ⩾70% circulating neutrophils had an extremely high value (AUC 0.97) for the diagnosis of ECOPD.

Previous studies

Many studies have previously described the clinical, physiological, biological and microbiological characteristics of ECOPD [26]. By and large, our clinical observations are in keeping with them, but some

TABLE 2 Comparison of correlation networks determined at exacerbation of chronic obstructive pulmonary disease (ECOPD) and convalescence

	ECOPD	Convalescence	p-value
Number of nodes	51	47	
Number of edges	96	125	
Within-module edges/between-module edges	12/84	37/88	
Node degree (*k*)	3.8-±2.6	5.3±2.8	<0.01
Hubs with (Kleinberg score) >0.8	SAA (1.00) TNF-α (0.92) IL-1b (0.92) IL-8 (0.92) TNF-RS (0.92) IL-6 (0.88) (all sputum variables)	FEV_1 (1.00) PAFI (0.97) P_{aO_2} (0.97) IC (0.90)	
Number of modules of at least 3 nodes	5	6	
Modularity score (fast-greedy algorithm)	0.871	0.685	
Total number of significantly different Spearman correlations	11	43	
Differential correlations with 0.3<|Rho(V1)–Rho(V2)|<0.5	11 (100%)	25 (58%)	
Differential correlations with |Rho(ECOPD)–Rho(convalescence)| ⩾0.5	0 (0%)	18 (42%)	<0.01
Within-module differential correlations/between-module differential correlations	7/4	24/19	

SAA: serum amyloid A; TNF: tumour necrosis factor; IL: interleukin; TNF-RS: tumour necrosis factor soluble receptor; FEV_1: forced expiratory volume in 1 s; P_{aO_2}: arterial partial pressure of oxygen; PAFI: P_{aO_2} (mmHg)/inspired fraction of oxygen ratio (%); IC: inspiratory capacity.

deserve specific comment. During ECOPD 1) a substantial number of patients had pulmonary hypertension and right chamber enlargement, in keeping with recent reports [27], but we did not identify patients with low ejection fraction heart failure [28]; and 2) CT found evidence of pulmonary embolism in only 1.5% of patients [29, 30] but, in contrast, alveolar infiltrates (not seen in chest radiographs) were identified in about a quarter of patients, as reported recently too [31]. These alveolar infiltrates can correspond to pneumonic condensations not apparent in plain chest radiographs and/or areas of local inflammation/oedema. At convalescence many (but not all) abnormalities observed during ECOPD improved. Of note, 3) even though dyspnoea and pulmonary gas exchange improved, spirometric changes only showed a statistically nonsignificant trait to improvement, which is at variance with other previous, smaller studies [32–34]; 4) as expected, several markers of systemic inflammation (total leukocyte count and levels of circulating neutrophils, CRP and fibrinogen) were reduced at convalescence. Of note, however, only 5.7% of patients showed >2% circulating eosinophils during ECOPD, and this proportion increased up to 54.7% at convalescence. This is at variance with reports from other centres, where between 25% and 50% of the patients have >2% circulating eosinophils during ECOPD [3, 35, 36]. We do not have a clear explanation for these discrepancies but regional differences may play a role [37]; and, finally, 5) in patients producing spontaneous sputum, the prevalence of PPM and/or viruses did not change at convalescence. Given that bronchial colonisation in clinically stable COPD patients that produce spontaneous sputum is common [38], this may have contributed to explain this lack of statistically significant changes.

Interpretation of novel results

Homeokinetic disruptive effects of ECOPD

Homeokinesis has been defined as "the ability of an organism to maintain a highly organised internal environment fluctuating within acceptable limits in a far from equilibrium state" [24, 25]. ECOPD episodes appear to be characterised by disrupted homeokinesis since, during clinical stability we observed a dense and well-connected correlation network with physiologically meaningful modules whereas, during ECOPD, although these modules mostly remain their connections become disrupted to a large extent (figure 2, table 2). Specifically, during clinical stability a central module (MM7), which basically includes all lung function parameters, was closely co-regulated with other modules that include pulmonary and systemic inflammatory markers (MM8, MM9, MM10) as well as a general biochemical module (MM6). By contrast, during ECOPD, the system becomes more fragmented, the sputum inflammation module (MM5) appears isolated, and systemic inflammatory markers are also less well coordinated and distributed across two different modules (MM3 and MM4). That microbiological nodes appear isolated from the main network during ECOPD probably reflects the heterogeneity of these ECOPD. Finally, the Monte Carlo

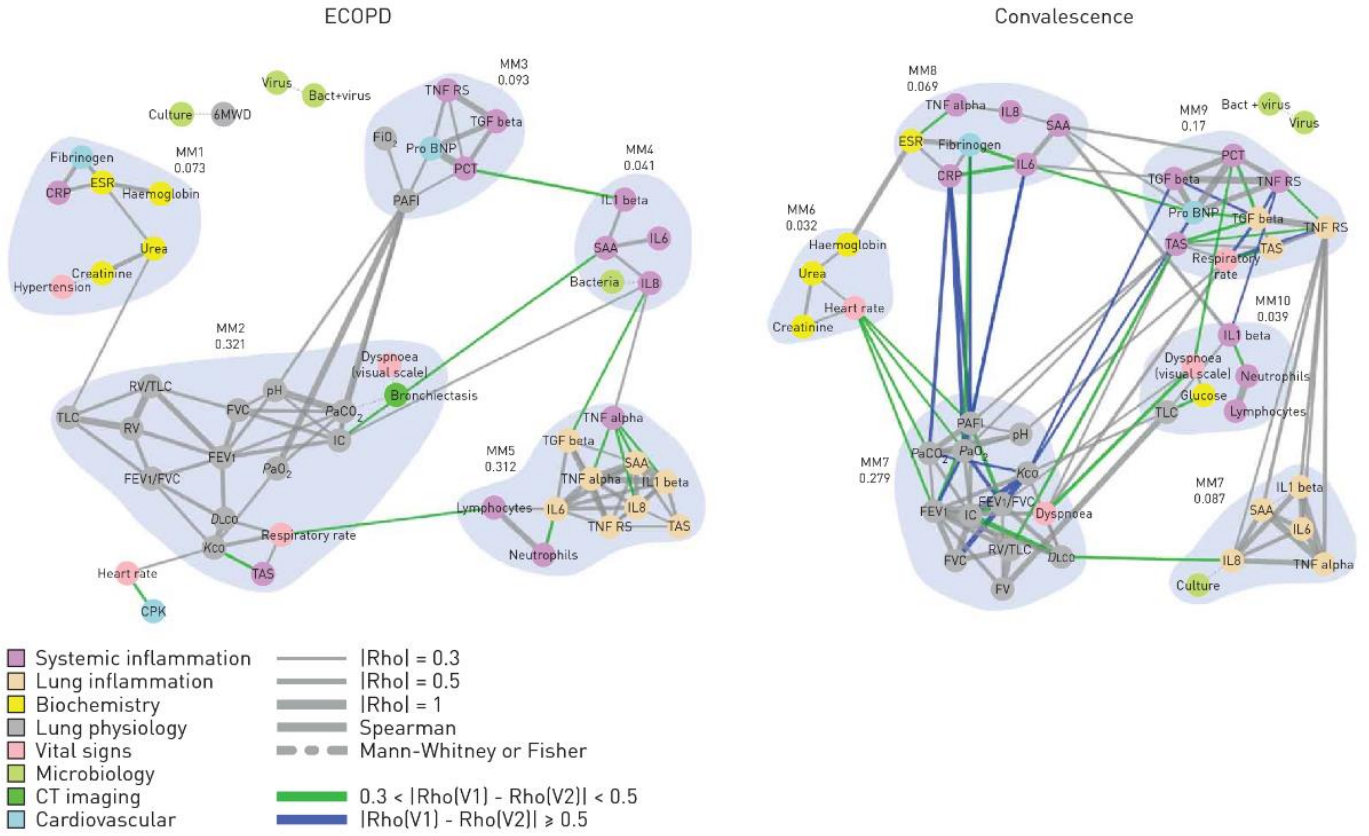

FIGURE 3 Same correlation networks presented in figure 2 now highlighting those that were significantly different at exacerbation of chronic obstructive pulmonary disease (ECOPD) and convalescence (*i.e.* Rho ECOPD≠Rho convalescence) by green or blue lines (see legend for Rho correspondence). For further explanations, see text.

permutation test [22] also identified more significantly different Spearman correlations at convalescence than during ECOPD (figure 3). All in all, these observations suggest that episodes of ECOPD are characterised by breakdown of the normal homeokinetic characteristics of the system with presumably less system control and resilience [24, 25].

A panel of biomarkers for the diagnosis of ECOPD

The diagnosis of ECOPD currently relies on the patient's perception of increased symptoms (mostly dyspnoea) [1, 5]. Yet, recent research has shown that dyspnoea perception vary between patients with frequent and infrequent exacerbations [39]. Thus, having an objective way to diagnose ECOPD is of great clinical relevance [2]. Our results indicate that the combination of increased dyspnoea (⩾5) and raised levels of circulating neutrophils (⩾70%) and CRP (⩾3 mg·L^{-1}) has an excellent value for the diagnosis of ECOPD (AUC 0.97) (figure 5c). Although the methodology we used is different, results are similar to those reported by Hurst *et al.* [40], who showed that elevated CRP levels were the best diagnostic biomarker for ECOPD, although their diagnosis accuracy was suboptimal (AUC 0.73); however, their combination with a major exacerbation symptom (dyspnoea, sputum volume or sputum purulence) significantly increased the AUC to 0.88 ($p<0.0001$) [40]. Our results extend these observations further by showing that this can be further improved (AUC 0.97) by considering too the number of circulating neutrophils. The potential diagnostic utility of this biomarker panel (as well as its specific cut-off values) will have to be validated prospectively in other cohorts, but it may greatly help to advance clinical research in this area by offering for the first time an objective diagnostic tool of ECOPD. Needless to say that increased dyspnoea, elevated CRP and leukocytosis can also occur in other clinical circumstances that may not even arise from the lungs (*e.g.* cholecystitis, pneumonia or sickle cell crisis, among others). Therefore, the clinical context in which these three biomarkers can contribute to the diagnosis of ECOPD is of paramount importance. Finally, using unbiased cluster analysis of 182 ECOPD episodes, Bafadhel *et al.* [3] recently provided convincing evidence of the heterogeneity of such episodes. Unfortunately, the relatively small sample size of our cohort (n=86) limits this type of analysis in our cohort.

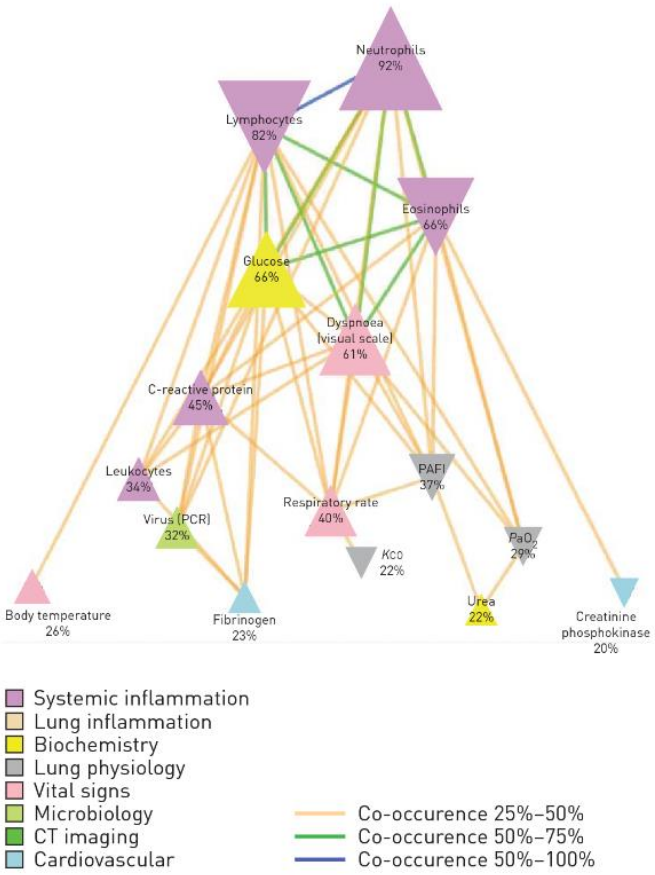

FIGURE 4 Outlier correlation network. Nodes represent the 16 identified variables with a significant proportion of outliers at exacerbation of chronic obstructive pulmonary disease (ECOPD). Node size is proportional to the percentage of outlier values (see digits inside nodes). Edges represent shared proportions (see colour codes). The shape of the triangle (upwards or downwards) indicates that the proportion of ECOPD outliers is lower or higher than the 5th or 95th convalescence percentiles, respectively.

Strengths and limitations

The development and application, for the first time to our knowledge, of a new analytical approach (*e.g.* MLDNA) to get further insight into the complexity of a relevant clinical problem like ECOPD is a clear strength of our study since it provides novel, integrated, dynamic and holistic information on this frequent condition. Importantly, it also paves the way for MLDNA to be applied to other complex biological conditions in respiratory medicine and elsewhere [6, 8, 16, 41, 42].

On the other hand, several potential limitations deserve comment. First, we included in the study a slightly lower number of patients (n=86) than anticipated (n=100; www.clinicaltrials.gov: NCT01750658), and not all measurements were available in all patients for comparison between ECOPD and convalescence. This is why we consider our study as proof-of-concept and we acknowledge that it requires validation in larger cohorts. Second, we studied severe (hospitalised) ECOPD, so our results are not directly generalisable to other milder forms of ECOPD. Third, some clinical variables, such as cough and sputum colour, were not registered. Fourth, it is not clear how much the initiation of systemic corticosteroids, before the collection of biological samples (within 72 h after admission) might have modified the inflammatory profile of ECOPD. Yet, it is of note that we excluded patients who received oral corticosteroid treatment before hospitalisation. Finally, patients present to hospital at various time points in the evolution of an ECOPD. All in all, we acknowledge that the results of this study will have to be confirmed in future studies, since the exclusion of severe co-morbidity, pneumonia, relatively small sample size and study of hospitalised patients (not ambulatory ECOPD) may restricts the generalisability of our observations.

Conclusions

By using a novel analytical strategy (MLDNA), this study shows that ECOPD 1) are characterised by disruption of network homeokinesis observed during clinical stability; and 2) in the appropriate clinical

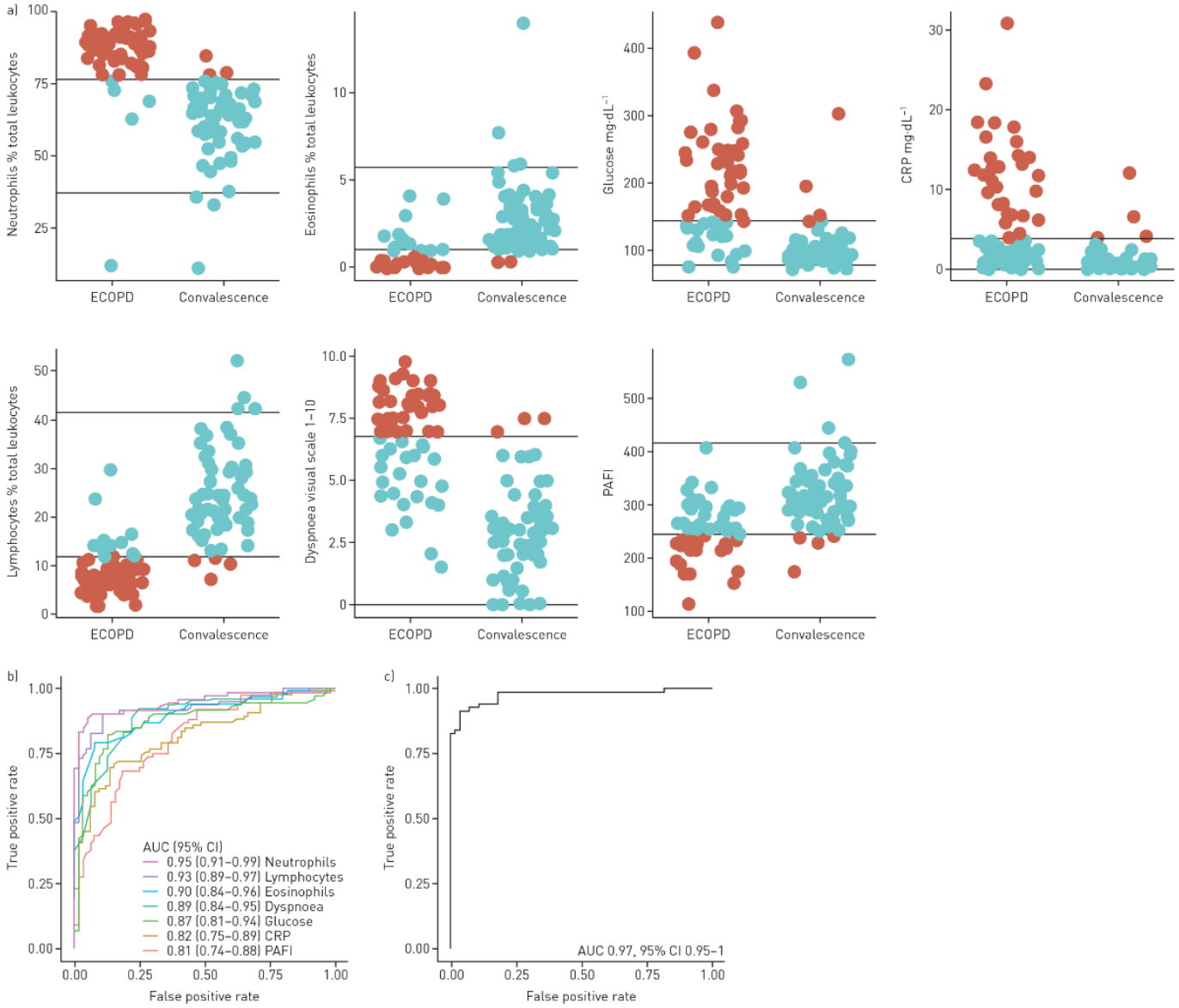

FIGURE 5 a) Scatter plot of seven continuous variables with a significant (bootstrapping FDR p-value <0.05) proportion of exacerbation of chronic obstructive pulmonary disease (ECOPD) outliers (<5TH or >95th percentiles (horizontal lines) at convalescence. Red symbols represent outlier values; blue symbols represent values within the convalescence 5th to 95th percentiles. b) Receiver operating characteristic curves and corresponding area under the curve (AUC) values for each of these 7 potential diagnostic biomarkers of ECOPD identified in a) with an AUC >0.8; c) When the seven variables identified in b) were combined in a general linear mixed model, the best panel of biomarkers to predict ECOPD (AUC 0.97) included circulating neutrophils, C-reactive protein levels and dyspnoea. For further explanations, see text.

TABLE 3 Specificity, sensitivity, negative (NPV) and positive predictive value (PPV) of a logistic regression model that includes different cut-off values of dyspnoea, C-reactive protein (CRP) and circulating neutrophil for the diagnosis of chronic obstructive pulmonary disease exacerbation (ECOPD)

Dyspnoea (visual analogue scale 1–10)	Neutrophils (%)	CRP ($mg \cdot L^{-1}$)	Specificity	Sensitivity	NPV	PPV
⩾5	⩾60	⩾3	0.89	0.94	0.92	0.92
⩾5	⩾65	⩾3	0.95	0.91	0.90	0.95
⩾5	⩾70	⩾3	0.96	0.90	0.88	0.97

context, ECOPD can be objectively identified by a panel of three biomarkers (dyspnoea, circulating neutrophils and CRP) commonly measured in clinical practice.

Acknowledgements

The authors thank participating patients for their willingness to contribute to medical research, and field attending physicians and nurses for their dedication and excellent care. The authors also thank Joan Soriano and Dolores Guerrero for their help in the start-up of the project, and Jørgen Vestbo, Judith Garcia-Aymerich, Josep Maria Antó and Peter J. Barnes for helpful suggestions.

References

1 Vogelmeier C, Agusti A, Anzueto A, *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2017 Report). 2017 Date last accessed: November 16, 2016.
2 Celli BR, Decramer M, Wedzicha JA, *et al.* An official American Thoracic Society/European Respiratory Society statement: research questions in COPD. *Eur Respir J* 2015; 45: 879–905.
3 Bafadhel M, McKenna S, Terry S, *et al.* Acute exacerbations of COPD: identification of biological clusters and their biomarkers. *Am J Respir Crit Care Med* 2011; 184: 662–671.
4 Beghe B, Verduri A, Roca M, *et al.* Exacerbation of respiratory symptoms in COPD patients may not be exacerbations of COPD. *Eur Respir J* 2013; 41: 993–995.
5 Rodriguez-Roisin R. Toward a consensus definition for COPD exacerbations. *Chest* 2000; 117: Suppl. 2, 398S–401S.
6 Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011; 12: 56–68.
7 Diez D, Agusti A, Wheelock CE. Network analysis in the investigation of chronic respiratory diseases: from basics to application. *Am J Respir Crit Care Med* 2014; 190: 981–988.
8 Menche J, Sharma A, Cho M, *et al.* A diVIsive Shuffling Approach (VIStA) for gene expression analysis to identify subtypes in chronic obstructive pulmonary disease. *BMC Syst Biol* 2014; 8: Suppl. 2, S8.
9 Rennard SI, Locantore N, Delafont B, *et al.* Identification of five COPD subgroups with different prognoses in the ECLIPSE cohort using cluster analysis. *Ann Am Thorac Soc* 2015.
10 Faner R, Gutierrez-Sacristan A, Castro-Acosta A, *et al.* Molecular and clinical diseasome of comorbidities in exacerbated COPD patients. *Eur Respir J* 2015; 46: 1001–1010.
11 Faner R, Agustí A. Network analysis: a way forward for understanding COPD multimorbidity. *Eur Respir J* 2015; 46: 591–592.
12 Faner R, Cruz T, López-Giraldo A, *et al.* Network medicine, multimorbidity and the lung in the elderly. *Eur Respir J* 2014; 44: 775–788.
13 Grosdidier S, Ferrer A, Faner R, *et al.* Network medicine analysis of COPD multimorbidities. *Respir Res* 2014; 15: 111.
14 Faner R, Cruz T, Casserras T, *et al.* Network analysis of lung transcriptomics reveals a distinct B cell signature in emphysema. *Am J Respir Crit Care Med* 2016; 193: 1242–1253.
15 Gustafsson M, Edstrom M, Gawel D, *et al.* Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment. *Genome Med* 2014; 6: 17.
16 Gustafsson M, Nestor C, Zhang H, *et al.* Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med* 2014; 6: 82.
17 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodological)* 1995; 57: 289–300.
18 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2015.
19 Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; 13: 2498–2504.
20 Newman ME, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004; 69: 026113.
21 Su G, Kuchinsky A, Morris JH, *et al.* GLay: community structure analysis of biological networks. *Bioinformatics* 2010; 26: 3135–3137.
22 Fishman GS. Monte Carlo: concepts, algorithms and applications. New York, Springer, 1995.
23 Davies L, Angus RM, Calverley PM. Oral corticosteroids in patients admitted to hospital with exacerbations of chronic obstructive pulmonary disease: a prospective randomised controlled trial. *Lancet* 1999; 354: 456–460.
24 Macklem PT. Emergent phenomena and the secrets of life. *J Appl Physiol* 2008; 104: 1844–1846.
25 Macklem PT, Seely A. Towards a definition of life. *Perspect Biol Med* 2010; 53: 330–340.
26 Wedzicha JA, Singh R, Mackay AJ. Acute COPD exacerbations. *Clin Chest Med* 2014; 35: 157–163.
27 Freixa X, Portillo K, Pare C, *et al.* Echocardiographic abnormalities in patients with COPD at their first hospital admission. *Eur Respir J* 2013; 41: 784–791.
28 Singanayagam A, Schembri S, Chalmers JD. Predictors of mortality in hospitalized adults with acute exacerbation of chronic obstructive pulmonary disease. *Ann Am Thorac Soc* 2013; 10: 81–89.
29 Roca M, Verduri A, Corbetta L, *et al.* Mechanisms of acute exacerbation of respiratory symptoms in chronic obstructive pulmonary disease. *Eur J Clin Invest* 2013; 43: 510–521.
30 Tillie-Leblond I, Marquette CH, Perez T, *et al.* Pulmonary embolism in patients with unexplained exacerbation of chronic obstructive pulmonary disease: prevalence and risk factors. *Ann Intern Med* 2006; 144: 390–396.
31 Cheng T, Wan H, Cheng Q, *et al.* Computed tomography manifestation of acute exacerbation of chronic obstructive pulmonary disease: A pilot study. *Exp Ther Med* 2016; 11: 519–529.
32 Stevenson NJ, Walker PP, Costello RW, *et al.* Lung mechanics and dyspnea during exacerbations of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2005; 172: 1510–1516.
33 Parker CM, Voduc N, Aaron SD, *et al.* Physiological changes during symptom recovery from moderate exacerbations of COPD. *Eur Respir J* 2005; 26: 420–428.

34 Cote CG, Dordelly LJ, Celli BR. Impact of COPD exacerbations on patient-centered outcomes. *Chest* 2007; 131: 696–704.
35 Bafadhel M, McKenna S, Terry S, *et al.* Blood eosinophils to direct corticosteroid treatment of exacerbations of chronic obstructive pulmonary disease: a randomized placebo-controlled trial. *Am J Respir Crit Care Med* 2012; 186: 48–55.
36 Bafadhel M, Davies L, Calverley PMA, *et al.* Blood eosinophil guided prednisolone therapy for exacerbations of COPD: a further analysis. *Eur Respir J* 2014; 44: 789–791.
37 Saha S, Brightling CE. Eosinophilic airway inflammation in COPD. *Int J Chron Obstruct Pulmon Dis* 2006; 1: 39–47.
38 Bandi V, Apicella MA, Mason E, *et al.* Nontypeable *Haemophilus influenzae* in the lower respiratory tract of patients with chronic bronchitis. *Am J Respir Crit Care Med* 2001; 164: 2114–2119.
39 Scioscia G, Blanco I, Burgos F, *et al.* Different dyspnea perception in COPD patients with frequent and infrequent exacerbations. *Thorax* 2017; 72: 117–121.
40 Hurst JR, Donaldson GC, Perera WR, *et al.* Use of plasma biomarkers at exacerbation of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2006; 174: 867–874.
41 Menche J, Sharma A, Kitsak M, *et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 2015; 347: 1257601.
42 Ghiassian SD, Menche J, Barabasi AL. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 2015; 11: e1004120.

Original Paper 2: Low Lung Function in Early Adulthood and Health in Later Life: a Transgenerational Cohort Analysis

(Published in: Agusti, **Noell**, Brugada, Faner. *Lancet Respir. Med.* 2017)

Lung function in early adulthood and health in later life: a transgenerational cohort analysis

Alvar Agustí, Guillaume Noell*, Josep Brugada, Rosa Faner*

Summary

Background Early life events can affect health in later life. We hypothesised that low lung function (FEV_1<80% predicted) in early adulthood (25–40 years) is associated with higher prevalence and earlier incidence of respiratory, cardiovascular, and metabolic abnormalities, and premature death.

Methods In this cohort analysis, we tested this hypothesis using data from the Framingham Offspring Cohort (FOC) and validated our observations in CARDIA (an independent cohort) and GenIII (which includes the direct descendants of FOC participants). These were three general population cohorts that included men and women, who were regularly and prospectively followed up to collect extensive clinical, physiological, biological, and imaging information. Main outcomes were prevalence (in early adulthood) and incidence (during follow-up) of comorbidity, and all-cause mortality. χ^2 test, unpaired *t* test, Fisher's exact test, and Cox proportional hazards models were used for data analysis. Differential dropout rates during follow-up were regarded as a potential source of bias.

Findings We found that 111 (10%) of 1161 participants in FOC, 338 (13%) of 2648 participants in CARDIA, and 71 (4%) of 1912 participants in GenIII had FEV_1 of less than 80% predicted at the age of 25–40 years. These individuals also had higher prevalence of respiratory, cardiovascular, and metabolic abnormalities in early adulthood; higher and earlier (about a decade) incidence of comorbidities during follow-up (39 years *vs* 47 years in FOC; 30 years *vs* 37 years in CARDIA, p<0·0001); and higher all-cause mortality than individuals with normal lung function in early adulthood (in FOC, hazard ratio 2·3 [95% CI 1·4–3·7], p=0·001), which was independent of, but additive with, cumulative smoking exposure. In GenIII, we observed that individuals with at least one parent stratified as having low lung function in early adulthood in FOC (n=115) had lower FEV_1 in early adulthood (10% had FEV_1 of less than 80% predicted; this proportion was 3% in those with both parents classified as normal in FOC [n=248]; p<0·0001); and early adulthood FEV_1 of GenIII participants was related (R^2=0·28, p<0·0001) to FOC parents' average FEV_1 in early adulthood.

Interpretation Low peak lung function in early adulthood is common in the general population and could identify a group of individuals at risk of early comorbidities and premature death.

Funding Fondo de Investigacion Sanitaria, Sociedad Española de Neumologia y Cirurgia Torácica, Formació Personal Investigador, Agencia de Gestió d'Ajuts de Recerca 2016, and AstraZeneca Foundation Young Researcher Award.

Lancet Respir Med 2017

Published Online November 14, 2017 http://dx.doi.org/10.1016/S2213-2600(17)30434-4

See Online/Comment http://dx.doi.org/10.1016/S2213-2600(17)30437-X

*Contributed equally

Hospital Clínic, University of Barcelona, Barcelona, Spain (A Agustí MD, Prof J Brugada MD); **Institut d'Investigacions Biomédiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain** (A Agustí, G Noell MS, Prof J Brugada, R Faner PhD); **and Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Barcelona, Spain** (A Agustí, G Noell, R Faner)

Correspondence to: Dr Rosa Faner, Centro de Investigación Biomédica en Red Enfermedades Respiratorias Institut d'Investigacions Biomédiques August Pi i Sunyer (CIBERES-IDIBAPS), Barcelona 08036, Spain
RFANER@clinic.cat

Introduction

Chronic obstructive pulmonary disease (COPD) is a major cause of disability and death around the globe.[1] COPD is generally considered to be a self-inflicted disease caused by tobacco smoking and characterised by an accelerated decline of lung function with age.[2] Yet, other COPD risk factors, including occupational exposures to organic and inorganic dusts; chemical agents and fumes; indoor pollution from biomass cooking and heating in poorly ventilated dwellings; and a history of severe childhood respiratory infection, HIV, or tuberculosis, have also been identified.[1] Furthermore, low peak lung function in early adulthood has been shown to increase the risk of COPD later in life, independently of the rate of lung function decline.[3] A previous study showed that about half of patients diagnosed with COPD in late adulthood had evidence of low peak lung function in early adulthood.[3] These observations suggest that abnormal lung development (in utero, after birth, or both) could be a novel risk factor for COPD.[1,3]

Lung development is a complex process that can be altered by various genetic or environmental factors,[4] including passive smoking, poor nutrition, and repeated infections.[5–7] These factors (acting alone or in combination) might also compromise the development of other organ systems (eg, the cardiovascular and metabolic systems).[8–10] We hypothesised that individuals with low lung function in early adulthood would also present a higher prevalence of respiratory, cardiovascular, and metabolic abnormalities, as well as a higher and earlier incidence of comorbid diseases and premature mortality during follow-up compared with individuals with normal lung function. Given that there is familial COPD aggregation,[11] and that lung function has been related to several environmental exposures and gene polymorphisms,[12] we also aimed to explore the trans-generational reproducibility of these traits.

Methods

Study design and participants

For this cohort analysis, we obtained permission to access two large independent cohorts (the Framingham

Research in context

Evidence before this study

We searched for articles published in English up to June 30, 2017, in PubMed with the search terms "loci associated lung function (GWAS)", "lung function trajectories", and "longitudinal lung function patterns". We also searched for relevant references in major review articles from noted experts. We identified evidence supporting that lung function is heritable (with up to 95 associated genetic variants described so far), and that lung function early in life tracks with lung function later in life and is a novel risk factor for chronic obstructive pulmonary disease (COPD).

Added value of this study

To our knowledge, this is the first study to test the hypothesis that the genetic or environmental factors that affect lung development might also affect other organ systems, such as the cardiovascular and metabolic systems, and that this might increase the likelihood of having a higher prevalence and earlier incidence of comorbidities during follow-up, which could lead to premature death. We tested this hypothesis in the Framingham Offspring Cohort (FOC) and validated the reproducibility of observations in CARDIA (an independent cohort) and GenIII (which includes the direct descendants of participants in FOC). Our results showed that low peak lung function in early adulthood is associated with a higher prevalence, and about a decade earlier incidence, of respiratory, cardiovascular, and metabolic abnormalities, as well as with premature death.

Implications of all the available evidence

Our results confirm previous observations that indicate that smoking is not the only risk factor for COPD and extend them by showing that low peak lung function in early adulthood identifies a group of individuals at risk of poor health outcomes later in life (higher incidence of comorbidities and premature death). Thus, the possibility that spirometry measured during infancy or early adulthood identifies these individuals and facilitates the implementation of effective preventive or therapeutic measures merits further research.

Offspring Cohort [FOC][13] and the Coronary Artery Risk Development in Young Adults Study [CARDIA] Cohort[14]) and the Framingham Generation III cohort (GenIII), which includes the direct descendants of FOC participants.[15]

The FOC started between August, 1971, and September, 1975, and includes 5124 participants aged between 5 and 93 years.[13] The offspring cohort consists of children of individuals in the original Framingham cohort, who were respondents of a random sample of two-thirds of the adult population of Framingham, MA, USA.

CARDIA is a community-based cohort that recruited black and white individuals (aged 18–30 years) from March 26, 1985, to June 7, 1986, from four US centres.[14] The GenIII cohort includes the offspring of FOC participants (aged 19–78 years). Most participants (98%) were white. Thus, these two cohorts are not fully independent and might share some genetic background and might have been exposed to similar environmental factors. GenIII started in April, 2002, and is ongoing.[15]

We obtained ethics approval from the institutional review board of Hospital Clinic (Barcelona, Spain) for the analysis (DbGaP project 7202).

Procedures

In these cohorts, we investigated cross-sectional differences between participants with normal versus low lung function (FEV_1 ≥80% [normal] or <80% [low] predicted), both in early adulthood (25–40 years; FOC, CARDIA, and GenIII) and late adulthood (50–65 years; FOC, CARDIA); and the incidence of comorbidities (appendix) and death during follow-up (FOC, CARDIA).

See Online for appendix

For our analysis, we extracted data from eight clinical visits for the FOC (exam 1, 1971–75; exam 2, 1979–83; exam 3, 1983–87; exam 4, 1987–91; exam 5, 1991–95; exam 6, 1995–98; exam 7, 1998–2001; exam 8, 2005–08), which spanned almost 40 years of follow-up. Following the same criteria used in our previous analysis of the FOC cohort,[3] participants were stratified in two groups (normal or low) according to their FEV_1 value (≥80% or <80% predicted[16]) in early adulthood (25–40 years). To reduce classification errors due to spirometry variability, we restricted our analysis to FOC participants with two or more concordant (normal or low) FEV_1 values in early adulthood (n=1161). These individuals were followed up until they dropped out of the study, death, or late adulthood (50–65 years), when clinical and functional measurements were repeated in survivors.

For the CARDIA cohort, we included in the analysis only those participants with two or more concordant FEV_1 values in early adulthood (n=2648). We extracted data from six clinical visits for these participants (recruitment and visits at 2, 5, 7, 10, and 15 years), which spanned 20 years of follow-up.[14]

For the GenIII cohort, we extracted data from two visits (2002–05 [n=4095], and 2008–11 [n=3411]) and finally included 1912 individuals with available spirometric measurements at the age of 25–40 years. In these individuals, only one spirometry was available (and used) for analysis.

Outcomes

Main outcomes were the prevalence and incidence of comorbid diseases and all-cause mortality in normal versus low individuals. Differential dropout rates during follow-up were considered as a potential source of bias.

	Normal		Low		p value
	N	n (%) or mean (SD)	N	n (%) or mean (SD)	
Demographics					
Age (years)	1050	33·7 (1·5)	111	34·0 (1·5)	0·102
Sex					
Men	1050	488 (47%)	111	57 (51%)	0·380
Women	1050	562 (53%)	111	54 (49%)	0·380
BMI (kg/m^2)	1050	25·2 (4·1)	111	25·3 (5)	0·766
Morbid obesity (BMI>40 kg/m^2)	1050	14 (1%)	111	2 (2%)	0·999
Pregnancy, delivery, and infancy data					
Low birthweight (<2·5 kg)	472	41 (9%)	50	10 (20%)	0·021
Caesarean delivery	745	62 (8%)	76	9 (12%)	0·409
Maternal obesity, diabetes, or hypertension	241	21 (9%)	30	6 (20%)	0·105
Overweight children	268	63 (24%)	32	9 (28%)	0·720
Smoking exposure					
Ever smoker	1050	676 (64%)	111	84 (76%)	0·023
Age started smoking (years)	619	17·5 (3·5)	75	16·8 (2·4)	0·020
Number of cigarettes smoked per day	576	20·3 (11·9)	74	28·8 (14·1)	<0·0001
Current smoker	676	486 (72%)	84	71 (85%)	0·012
Respiratory measures					
FEV_1 (% predicted)	1024	97·3 (8·6)	109	69·5 (8·1)	<0·0001
FEV_1/FVC (%)	1050	84·4% (5·7)	111	76·8% (8·9)	<0·0001
Inhaled medication for respiratory diseases	771	5 (1%)	91	6 (7%)	<0·0001
Asthma	1050	86 (8%)	111	28 (25%)	<0·0001
Chronic bronchitis, emphysema, COPD	1049	58 (6%)	110	24 (22%)	<0·0001
Pneumonia	571	6 (1%)	45	1 (2%)	0·999
Other pulmonary disease	910	10 (1%)	69	2 (3%)	0·458
Cardiovascular measures					
Electrocardiogram abnormality	1046	287 (27%)	111	46 (41%)	0·003
Arterial hypertension treatment	1050	81 (8%)	111	7 (6%)	0·731
Cardiovascular disease medication	585	17 (3%)	47	1 (2%)	0·999
Arrhythmia	1048	41 (4%)	111	6 (5%)	0·613
Myocardial infarction	1049	1 (<1%)	111	0 (0)	0·999
Peripheral vascular disease	1050	53 (5%)	111	9 (8%)	0·254
Valvular disease (aortic, mitral, rheumatic)	1050	30 (3%)	111	4 (4%)	0·883
Other heart disease (excluding above)	928	13 (1%)	92	2 (2%)	0·894
Echocardiographic abnormalities	902	140 (16%)	96	19 (20%)	0·347
Left ventricular ejection fraction	905	74·2% (4·2)	98	73·8% (5·4)	0·551
Metabolic measures					
Diabetes	1046	7 (1%)	109	4 (4%)	0·011
Hypercholesterolaemia treatment	1034	3 (<1%)	110	1 (1%)	0·845
Circulating blood measures					
White blood count (1000 cells per μL)	909	6·3 (1·7)	98	7·1 (2·1)	0·001
C-reactive protein (mg/L)	916	2·0 (4·1)	93	2·2 (3)	0·576
Plasma fibrinogen (mg/dL)	172	284·9 (49·6)	7	312·3 (53·6)	0·229

FOC=Framingham Offspring Cohort. N=number of individuals available for each specific comparison. BMI=body-mass index. FVC=forced vital capacity. COPD=chronic obstructive pulmonary disease.

Table 1: **Characteristics of FOC participants with normal or low lung function in early adulthood**

Statistical analysis

When several continuous variable measurements from different clinical visits during a given study period (early or late adulthood) were available for the same individual, they were averaged to get a unique estimate per participant and time period. The key categorising variable (FEV_1) was not averaged when it was used to stratify participants into the low or normal group, as we required the

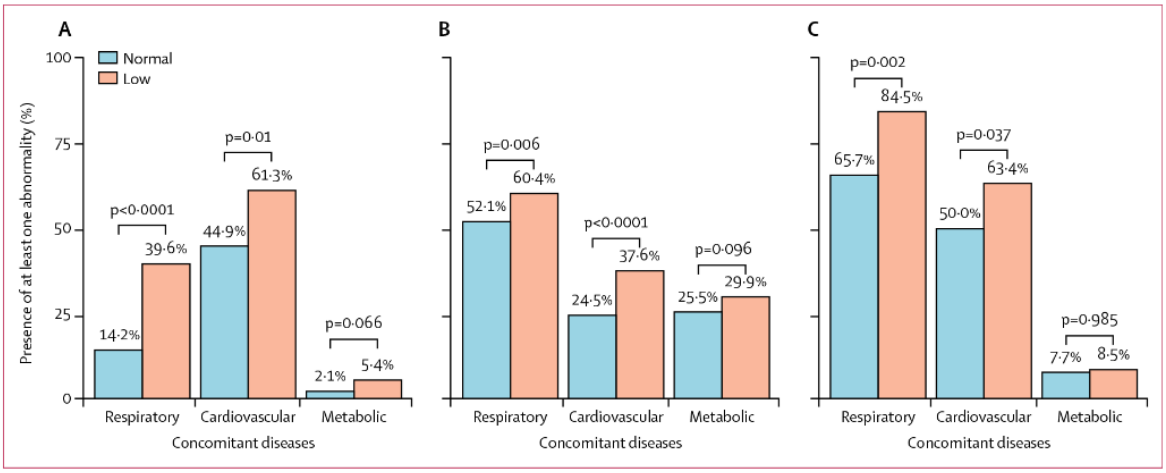

Figure 1: **Proportion of participants with at least one respiratory, cardiovascular, or metabolic abnormality by lung function level in early adulthood**
(A) FOC. (B) CARDIA. (C) GenIII. CARDIA=Coronary Artery Risk Development in Young Adults Study. FOC=Framingham Offspring Cohort. GenIII=Framingham Generation III cohort.

	Normal		Low		p value
	N	n (%) or mean (SD)	N	n (%) or mean (SD)	
Demographics					
Age (years)	2310	31·9 (1·6)	338	31·7 (1·7)	0·009
Sex					
Men	2310	1015 (44%)	338	176 (51%)	0·006
Women	2310	1295 (56%)	338	162 (49%)	0·006
BMI (kg/m²)	2304	25·6 (5·1)	337	28·0 (7·2)	<0·0001
Morbid obesity (BMI>40 kg/m²)	2304	77 (3%)	337	30 (9%)	<0·0001
Pregnancy, delivery, and infancy data					
Paternal or maternal diabetes	2258	491 (22%)	324	111 (34%)	<0·0001
Maternal diabetes	2239	269 (12%)	318	76 (24%)	<0·0001
Paternal diabetes	2103	273 (13%)	278	53 (19%)	0·015
Maternal high blood pressure	2160	928 (43%)	313	185 (59%)	<0·0001
Paternal high blood pressure	1970	887 (45%)	268	131 (49%)	0·263
Smoking exposure					
Ever smoker	2051	1089 (53%)	277	170 (61%)	0·011
Age started smoking (years)	1002	17·6 (4·3)	158	17·9 (4·9)	0·406
Number of cigarettes smoked per day	1843	6·2 (8·3)	245	7·8 (8·8)	0·005
Years smoked regularly	781	8·7 (5·6)	107	10·5 (5·3)	0·001
Current smoker	1085	710 (64%)	169	66 (38%)	<0·0001
Respiratory measures					
FEV_1 (% predicted)	2310	97·7 (9·3)	338	70·5 (6·3)	<0·0001
FEV_1/FVC (%)	2310	82·0% (5·3)	338	78·0% (8·1)	<0·0001
Lung problems before age of 16 years	1557	98 (6%)	191	20 (11%)	0·044
Respiratory symptoms	2310	670 (29%)	338	115 (34%)	0·040
Asthma	2310	347 (15%)	338	98 (29%)	<0·0001
Chronic bronchitis, emphysema, COPD	2307	349 (15%)	337	61 (18%)	0·184
Pneumonia	2303	413 (18%)	337	67 (20%)	0·429
Tuberculosis or lung cancer	2153	65 (3%)	313	22 (7%)	0·001
Cardiovascular measures					
High blood pressure or treatment	2310	467 (20%)	338	109 (32%)	<0·0001

(Table 2 continues on next page)

FEV_1 % predicted values of each individual to be concordant in the age range of 25–40 years, both in FOC and CARDIA (ie, either all values higher than 80% predicted or all below 80% reference). Likewise, main outcome variables (comorbidity, death) are categorical and were not averaged either. We selected variables to be included in the analysis from those available in each of the three cohort datasets by clinical judgment—ie, by considering those which could eventually be more helpful to test our hypothesis and to interpret the results clinically. We used the χ^2 test to compare categorical variables and the unpaired *t* test to compare continuous variables, in normal versus low participants cross-sectionally. We used a Cox proportional hazards model[17] adjusted for potential baseline confounders (sex and body-mass index [BMI]) to estimate the time to first reported comorbidity or death in individuals with normal or low peak lung function in early adulthood both in FOC and CARDIA. We compared differential dropout proportions during follow-up (excluding deaths) in 5-year bins, from 20 to 65 years, with Fisher's exact tests. p values less than 0·05 were considered statistically significant. All statistical analyses were performed with custom R scripts and relevant Bioconductor Packages.[18]

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. GN and RF had access to the raw data. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

In participants from the FOC cohort, recruited between August, 1971, and September, 1975, FEV_1 in early adulthood (25–40 years) was consistently 80% or higher than

predicted in 1050 (90%) of 1161 participants, who were therefore classified as having normal lung function, whereas FEV_1 was less than 80% predicted in 111 (10%), who were classified as having low lung function. Demographics were similar between the two groups (table 1). The proportion of individuals with low birthweight was two-times higher in participants with low lung function that in those with normal lung function. There were no differences in reported maternal illnesses, caesarean deliveries, or overweight children. The prevalence of ever smokers was higher in individuals with low lung function, who also started smoking almost a year earlier, and had higher cumulative smoking exposure, and included a higher proportion of current smokers than in the normal lung function group. By design, FEV_1 was less than 80% predicted in individuals with low lung function. Notably, the FEV_1-to-forced vital capacity ratio (FEV_1/FVC) was also significantly lower in these participants than in those with normal lung function. In keeping with these functional abnormalities, individuals with low lung function used inhaled medications for respiratory diseases more often and were diagnosed with respiratory diseases such as asthma, chronic bronchitis, emphysema, or COPD more frequently than those with normal lung function. Accordingly, the proportion of cumulative (at least one) respiratory abnormalities was significantly increased in participants with low lung function (figure 1A).

Individuals with low lung function also had a significantly higher prevalence of electrocardiogram abnormalities, and other clinical cardiological diagnoses were numerically more prevalent, but not statistically different (table 1). However, there was a significantly higher prevalence of cumulative cardiovascular abnormalities in participants with low lung function (figure 1A). Likewise, the prevalence of diabetes was four times higher in individuals with low lung function and, these individuals also had higher circulating leucocyte counts (table 1).

338 (13%) of 2648 participants in CARDIA were classified as individuals with low lung function in early adulthood. Observations in CARDIA were largely the same as in FOC participants (table 2, figure 1B), but some differences should be noted. Although statistically different, probably due to the large sample size, there were clinically insignificant differences in age (31·7 *vs* 31·9 years). There were more men and participants who were overweight in those with low lung function, and their mothers and fathers reported having diabetes and arterial hypertension more often. Similar to FOC, smoking exposure was higher in individuals with low lung function, but the proportion of current smokers was lower than in participants with normal lung function. CARDIA participants with low lung function in early adulthood also reported more frequent respiratory symptoms, before the age of 16 years, and reported having been diagnosed with other respiratory diseases, such as asthma, tuberculosis, and lung cancer, more often than in the normal lung function group. These individuals had numerically more cardiovascular and metabolic abnormalities, including higher concentrations of some systemic inflammatory markers, although differences did not reach statistical significance in some instances (table 2). The proportion of participants with at least one respiratory, cardiovascular, and metabolic abnormality in CARDIA was higher in participants with low lung function than in those with normal lung function (figure 1B).

The GenIII cohort included 1912 individuals, 1841 (96%) of whom had normal, and 71 (4%) of whom had low lung function in early adulthood (table 3). Because participants in GenIII are direct descendants of FOC participants, they are genetically related and, therefore, observations need to be considered with caution when considering the reproducibility of observations. With this caveat in mind, participants in GenIII with low lung function in early adulthood had similar proportions of abnormalities measured to those in FOC (related cohort) and CARDIA (independent cohort). Participants with low lung function in GenIII were again most often men, were more often overweight, with a higher prevalence of parental asthma, with higher smoking exposure, who had evidence of airflow limitation and reported more respiratory

	Normal		Low		p value
	N	n (%) or mean (SD)	N	n (%) or mean (SD)	
(Continued from previous page)					
Cardiovascular disease medication	2193	63 (3%)	314	21 (7%)	0·001
Coronary artery disease	377	14 (4%)	62	6 (10%)	0·079
Arrhythmia	938	2 (<1%)	91	1 (1%)	0·633
Heart failure	2229	5 (<1%)	318	2 (1%)	0·474
Valvular heart disease (including rheumatic heart disease)	365	51 (14%)	61	10 (16%)	0·763
Congenital heart diseases	114	7 (6%)	16	2 (13%)	0·680
Maximum heart rate during exercise	2016	179·8 (13·9)	277	172·3 (18·0)	<0·0001
Recovery time to heart rate 130 bpm (s)	2007	268·3 (115·3)	268	243·1 125·5)	0·002
Metabolic measures					
Diabetes	2310	144 (6%)	338	35 (10%)	0·007
Glucose (mg/dL)	1533	83·2 (12·8)	188	84·0 (27·4)	0·669
Insulin (µU/mL)	1541	9·5 (6·4)	187	12·7 (9·4)	<0·0001
Leptin (ng/mL)	186	12·2 (14·4)	29	20·5 (21·0)	0·048
Plasma lipoprotein A (mg/dL)	2132	18·6 (21)	306	26·2 (22·6)	<0·0001
Hypercholesterolaemia or treatment	2310	434 (19%)	338	58 (17%)	0·520
Circulating blood measures					
White cell count (1000 cells per µL)	1545	6·0 (1·7)	187	6·2 (2.2)	0·183
C-reactive protein (µg/mL)	876	1·9 (2·4)	155	2·5 (3.0)	0·009
Fibrinogen (mg/dL)	2247	253·5 (52·8)	321	272·2 (61.8)	<0·0001
Interleukin 6 (pg/mL)	199	1·6 (1·3)	38	2·1 (1.8)	0·115

CARDIA=Coronary Artery Risk Development in Young Adults Study. N=number of individuals available for each specific comparison. BMI=body-mass index. FVC=forced vital capacity. COPD=chronic obstructive pulmonary disease. bpm=beats per min.

Table 2: **Characteristics of CARDIA participants with normal or low lung function in early adulthood**

	Normal		Low		p value
	N	n (%) or mean (SD)	N	n (%) or mean (SD)	
Demographics					
Age (years)	1841	34·9 (4·1)	71	35·7 (3·7)	0·134
Sex					
Men	1692	793 (47%)	65	36 (55%)	0·221
Women	1692	899 (53%)	65	29 (45%)	0·221
BMI (kg/m^2)	1841	26·5 (5·1)	71	29·3 (7·9)	0·004
Morbid obesity (BMI>40 kg/m^2)	1841	41 (2%)	71	5 (7%)	0·028
Pregnancy, delivery, and infancy data					
Paternal asthma	1457	217 (15%)	54	15 (28%)	0·017
Maternal asthma	1549	245 (16%)	58	12 (21%)	0·417
Smoking exposure					
Ever smoker	1841	654 (36%)	71	29 (41%)	0·428
Age started smoking regularly	121	17·5 (2·1)	7	17·7 (4·5)	0·795
Number of cigarettes smoked per day	259	13·1 (7·1)	15	17·4 (8·2)	0·065
Current smoker	653	287 (44%)	29	18 (62%)	0·084
Respiratory measures					
FEV_1 (% predicted)	1692	102·0 (8·2)	65	72·0 (10·5)	<0·0001
FEV_1/FVC (%)	1692	79·0% (10·3)	65	67·0% (5·6)	<0·0001
Presence of respiratory symptoms (dyspnoea, wheezing, chest discomfort)	1841	711 (39%)	71	39 (55%)	0·008
Asthma	1840	259 (14%)	71	21 (30%)	0·001
Chronic bronchitis, emphysema, COPD	1841	676 (37%)	71	31 (44%)	0·287
Pneumonia	1841	378 (21%)	71	25 (35%)	0·005
Sleep apnoea	1840	75 (4%)	71	4 (6%)	0·732
Pulmonary fibrosis	1833	2 (<1%)	70	0 (0)	0·999
Other pulmonary disease	917	106 (12%)	22	5 (23%)	0·204
DLCO (% reference)	1564	98·6 (13·1)	53	88·3 (16·2)	<0·0001
CT-diagnosed emphysema (measured at 40–50 years)	397	17 (4%)	18	4 (22%)	0·004
Cardiovascular measures					
Electrocardiogram abnormality	1841	849 (46%)	71	37 (52%)	0·383
Mean arterial pressure (mm Hg)	1686	86·7 (9·8)	64	92·3 (12·8)	0·001
Arterial hypertension treatment	1840	89 (5%)	71	9 (13%)	0·008
Cardiovascular disease treatment	1838	11 (1%)	71	0 (0)	0·999
Arrhythmia	1838	22 (1%)	71	1 (1%)	0·999
Myocardial infarction	1692	1 (<1%)	65	0 (0)	0·999
Valvular disease (aortic, mitral, rheumatic)	1839	22 (1%)	70	0 (0)	0·726
Peripheral vascular disease	1841	19 (1%)	71	0 (0)	0·802
Other heart disease	1778	15 (1%)	68	1 (2%)	0·999
Echocardiography: left ventricular percentage fractional shortening	1663	34·8 (3·4)	60	35·2 (4)	0·412
NT-proBNP (pg/mL)	1690	36·9 (37·7)	64	41·5 (48·6)	0·452
Metabolic measures					
Glucose (mg/dL)	1692	92·8 (16·5)	64	94·6 (9·6)	0·167
Glucose levels 2 h post tolerance test beverage (mg/dL)	719	98·5 (22·7)	19	114·6 (27·2)	0·019
HbA_{1c} (%)	765	5·3% (0·4)	22	5·6% (0·5)	0·028
Diabetes medication	1840	23 (1%)	71	2 (3%)	0·543

(Table 3 continues on next page)

symptoms, and who were more likely to have been diagnosed with asthma or pneumonia. Unlike FOC or CARDIA, which did not have information on emphysema, GenIII used two specific methods to assess the disorder (CT scan and lung diffusing capacity [DLCO]) and both methods resulted in the report of higher prevalence of emphysema in individuals with low lung function (table 3). As in FOC and CARDIA, individuals in GenIII with FEV_1 less than 80% predicted had a higher prevalence of cardiovascular and metabolic abnormalities (figure 1C), as well as higher concentrations of circulating inflammatory markers (table 3).

Longitudinal observations during follow-up in FOC showed that dropout rates were higher in participants with low lung function in early adulthood (appendix). The incidence of reported comorbid diagnoses during follow-up was higher in individuals with low lung function at any age (appendix). The mean age at which 50% of individuals reported the presence of one comorbid diagnosis was around a decade earlier in those with low lung function in early adulthood than in those with normal lung function (39 years *vs* 47 years, p<0·0001; figure 2A). Cox analysis showed that low lung function in early adulthood significantly increased first disease occurrence during follow-up, whereas never smoking (in or before early adulthood) and lower baseline BMI decreased it significantly; sex had no significant effect on age at first disease occurrence (appendix).

All-cause mortality during follow-up in FOC was higher in ever smokers (Cox model hazard ratio [HR] 1·8 [95% CI 1·1–2·8], p=0·028) and in individuals with low lung function in early adulthood (2·3 [1·4–3·7], p=0·001; figure 3). These two effects were statistically additive and independent (non-significant Fisher association and non-significant interaction in Cox models), and BMI did not significantly influence mortality. We did not find statistically significant differences in cause-specific mortality between high and low lung function groups, but there was a numerically higher, but non-significant, cardiovascular mortality in participants with low lung function (appendix).

FOC participants who were alive and not lost to follow-up were reassessed in late adulthood (50–65 years). Most of the differences observed between participants with low and normal lung function in early adulthood remained (appendix). The prevalence of emphysema (not assessed in early adulthood) was much higher in individuals with low lung function in early adulthood than in those with normal lung function (appendix).

Available follow-up data in CARDIA is shorter (20 years) than in FOC (40 years). However, observations during follow-up in CARDIA were similar to those of FOC discussed above. Similar to FOC, dropout rates and the incidence of comorbid diagnoses during follow-up were higher in those with low lung function in early adulthood (appendix), and the mean age at which 50% of individuals reported the presence of one comorbid diagnosis was

7 years earlier in CARDIA participants with low lung function in early adulthood (30 years *vs* 37 years, p<0·0001; figure 2B). Similarly, using a Cox model to investigate which factors affect the incidence of comorbid diagnoses during follow-up in CARDIA, we found that low lung function in early adulthood significantly increases risk; whereas, unlike in FOC, smoking status and baseline BMI did not have a significant effect on comorbidity incidence in CARDIA (appendix).

Because the precise date of death is not registered in the CARDIA database, we could not generate Kaplan-Meier survival curves for this cohort. However, all-cause mortality before the age of 50 years in CARDIA participants with low lung function in early adulthood was three times higher than that of people with normal lung function (3% *vs* 0·7%, odd ratio [OR] 4·1 [95% CI 1·7–9·6], p=0·001).

We could not investigate the reproducibility of FOC and CARDIA longitudinal observations in GenIII because it is an ongoing cohort and extended follow-up data are not yet available.

Finally, to assess the transgenerational reproducibility of low lung function in early adulthood, we compared the characteristics of Gen III participants for whom both parents were classified as normal in FOC (n=248) with those with at least one parent classified as having low lung function in FOC (n=115; figure 4). More individuals with at least one parent who had low lung function in early adulthood were women, reported more parental history of asthma, and, albeit within the normal range, had lower FEV_1 (% predicted) and FEV_1/FVC than those who had two parents with normal lung function. In GenIII participants with at least one parent stratified as having low lung function in FOC (n=115), 10% had FEV_1 of less than 80% predicted; this proportion was 3% in those with both parents classified as normal in FOC (n=248; p<0·0001; figure 4A, appendix). Further, we found a positive correlation between the FEV_1 of GenIII participants and FOC parents' average FEV_1 (R^2=0·28, p<0·0001; figure 4B). We did not find significant cardiovascular, metabolic, or systemic inflammatory marker differences between these two groups, although the prevalence of arterial hypertension was higher in descendants of FOC participants with low lung function in early adulthood (8·2% *vs* 2·9%, p=0·051).

	Normal		Low		p value
	N	n (%) or mean (SD)	N	n (%) or mean (SD)	
(Continued from previous page)					
Total cholesterol (mg/dL)	1692	184·3 (35·3)	64	184·6 (35·7)	0·961
HDL cholesterol (plasma, mg/dL)	1840	54·7 (15·0)	70	51·0 (15·6)	0·050
Triglycerides (mg/dL)	1692	109·3 (86·8)	64	126·0 (76·5)	0·094
Cholesterol medication	1840	90 (5%)	71	1 (1%)	0·285
Circulating blood measures					
White blood count (10^6 cells per µL)	480	6·1 (1·4)	14	7·3 (1·8)	0·027
C-reactive protein (mg/L)	1838	2·5 (4·4)	70	3·8 (4·9)	0·029
Fibrinogen (mg/dL)	1680	330·5 (67)	60	330·6 (74·9)	0·991
Interleukin 6 (pg/mL)	1673	1·6 (2·0)	63	2·3 (2·3)	0·020

GenIII=Framingham Generation III cohort. N=number of individuals available for each specific comparison. BMI=body-mass index. FVC=forced vital capacity. COPD=chronic obstructive pulmonary disease. DLCO=diffusion lung capacity for carbon monoxide. NT-proBNP=N-terminal pro-B-type natriuretic peptide. HbA_{1c}=glycated haemoglobin.

Table 3: **Characteristics of GenIII participants with normal or low lung function in early adulthood**

Discussion

In this study, we analysed three large cohorts (FOC, GenIII, and CARDIA) and found that 4–13% of the general population has low lung function (FEV_1 <80% predicted) in early adulthood (25–40 years of age); that this is not a bystander effect because these individuals also have a higher prevalence of respiratory, cardiovascular, and metabolic abnormalities and a higher and earlier incidence of comorbidities during follow-up than those with normal lung function in early adulthood, and these individuals also die prematurely; and some of these abnormalities are also found in direct descendants (GenIII).

It is well established that low birthweight is associated with dysfunction of several organs later in life[6,8,9,19] and that low lung function in infancy tracks into adulthood.[5,20–22] To our knowledge, however, this is the first study to test the hypothesis that the genetic or environmental factors that govern lung development[5,6,8,9] might also affect the development of other organ systems, such as the cardiovascular and metabolic systems, and that this might increase the likelihood of having a higher prevalence and earlier incidence of comorbidities during follow-up, eventually causing premature death. Our results support this hypothesis because individuals with low lung function in early adulthood consistently reported more frequent symptoms, were more often diagnosed with (and received treatment for) various clinical conditions, had higher and earlier incidence of comorbidities, and died earlier. Further, individuals with low lung function surviving into late adulthood continued to have a higher proportion of abnormal cardiopulmonary and metabolic disorders than those who had normal lung function, as well as evidence of low-grade systemic inflammation. An analysis of the Tucson Epidemiological Study of Airway Obstructive Disease recently confirmed that low FEV_1 by the age of 21–35 years predicts risk of early cardiopulmonary mortality.[23] Overall, these observations suggest that some of the comorbidities frequently reported in patients with COPD might originate earlier in life than previously thought and might not always be associated with ageing. This finding indicates that potential opportunities exist for prevention and early intervention.[7,24] Finally, the high reported prevalence of a previous diagnosis of several respiratory diseases, such as asthma, was notable. We propose that, if the lungs develop suboptimally, resulting symptoms or airflow limitation can be easily misdiagnosed as asthma. This potential misclassification should be considered in future studies.

The precise biological mechanisms underlying these observations cannot be disentangled from our results.

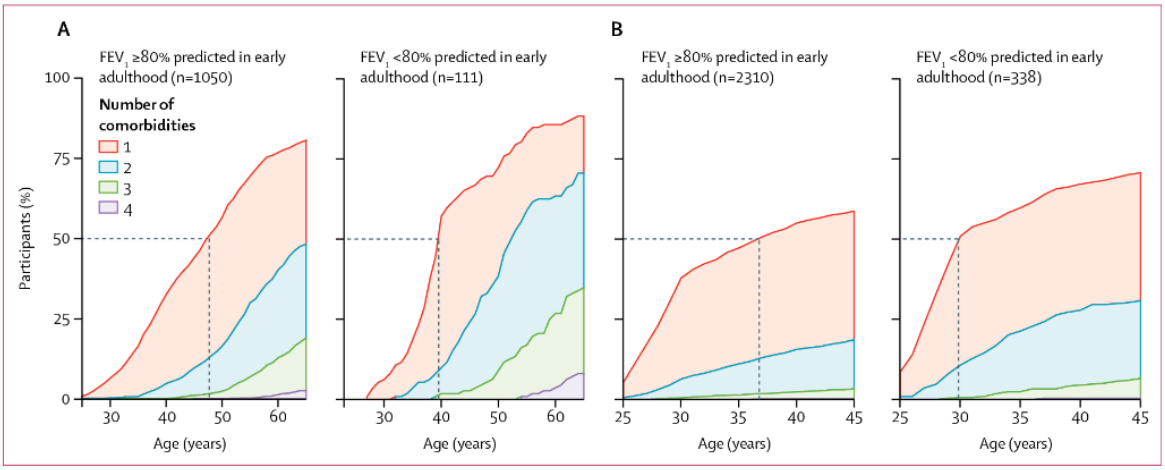

Figure 2: **Cumulative incidence of respiratory, cardiovascular, and metabolic abnormalities during follow-up in FOC (A) and CARDIA (B) participants by lung function level in early adulthood**

Age range for FOC was 40 years; age range for CARDIA was 20 years. Dotted lines indicate the age at which half of the population reports the first comorbidity. CARDIA=Coronary Artery Risk Development in Young Adults Study. FOC=Framingham Offspring Cohort.

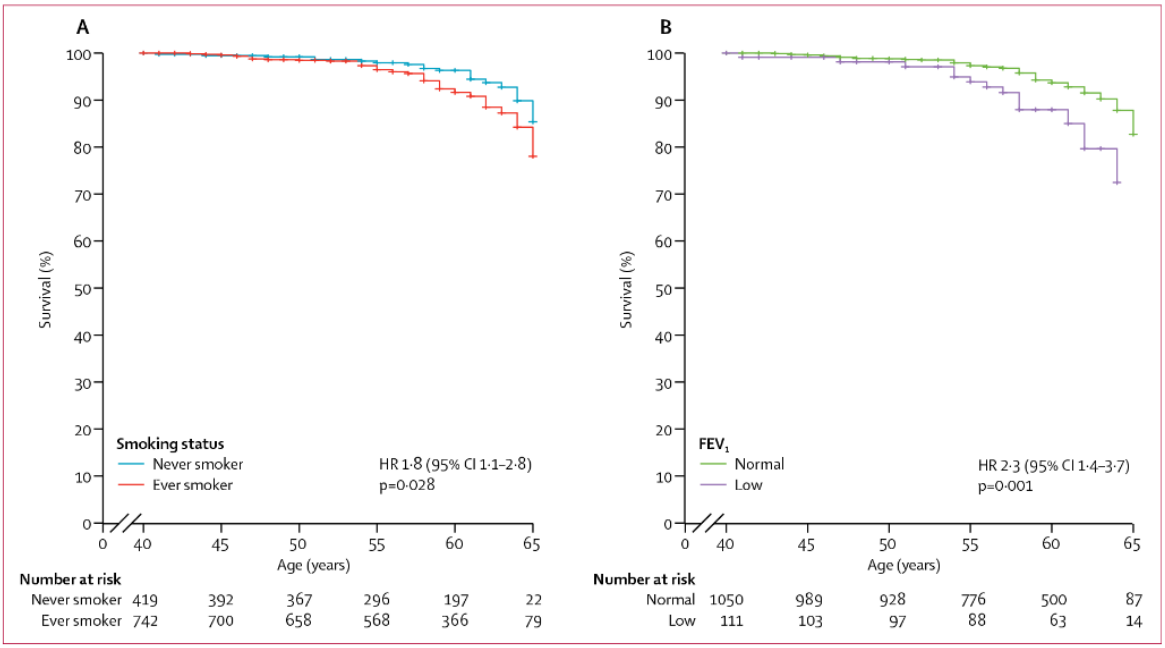

Figure 3: **Kaplan-Meier survival curves and Cox model HRs in FOC participants**

(A) Smoking (ever *vs* never smoker by 40 years of age). (B) Lung function in early adulthood. FOC=Framingham Offspring Cohort. HR=hazard ratio.

Theoretically, the observed associations between low peak lung function and the earlier incidence of comorbidities and premature death can result from shared genetic or environmental factors (eg, cumulative smoking exposure) or both;[7,25,26] in fact, smokers with low lung function in early adulthood consumed significantly more cigarettes than smokers with normal lung function. This exposure in early life might be a risk factor for having low lung function in the first place, and is certainly associated, as shown in many previous studies, with a higher incidence of comorbidity and premature death during follow-up. Further mechanistic studies are required to investigate the interaction between these factors. That observations in FOC and CARDIA were largely reproduced in GenIII participants (who were direct descendants of FOC participants) support genome-wide association studies in the general population, which identified specific gene variants associated with lung function levels,[27] but might also be the result of a shared exposome (eg, similar smoking habits or diet in families)

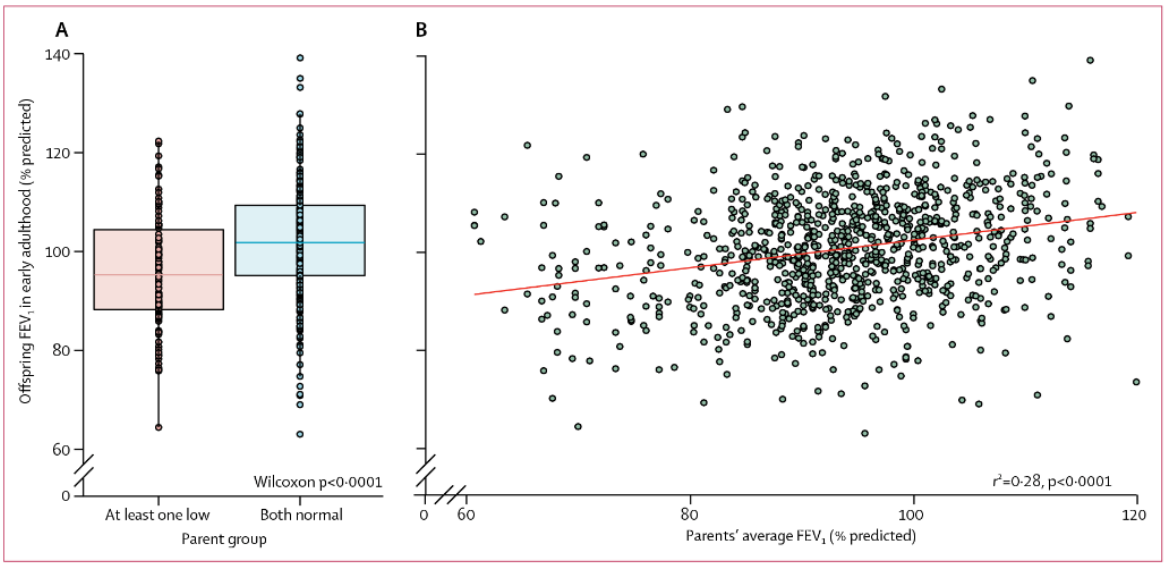

Figure 4: **Transgenerational FEV_1 observations**
(A) Box plot showing median FEV_1 (% predicted) of GenIII participants with at least one parent in FOC classified as low in early adulthood (n=115) and participants with both FOC parents classified as normal (n=248). (B) Scatter plot showing the relationship between early adulthood FEV_1 (% predicted) of GenIII participants and parents' average early adulthood FEV_1 (% predicted). FOC=Framingham Offspring Cohort.

or, again, the interaction of both.[26] However, it is interesting to note that individuals with low peak lung function also reported a higher prevalence of emphysema, which might represent poor lung development, enhanced lung destruction, or deficient lung maintenance capacity.[28,29] Finally, individuals with low lung function were often men, overweight, and had a high prevalence of diabetes and who often reported family history of asthma, hypertension, and diabetes, potentially illustrating the complex interactions between the genome and the exposome.[30]

The main strength of our study is that it tests a novel hypothesis in three large and well characterised, independent (FOC and CARDIA) and family related (FOC and GenIII) cohorts, whose participants were followed up for long periods of time (FOC and CARDIA). Potential limitations include higher dropout rates during follow-up in participants with low lung function, both in FOC and CARDIA. Yet, if anything, this should contribute to underestimate the effect size of the observed differences. The prevalence of individuals' low lung function in early adulthood in GenIII was lower than that in FOC or CARDIA. One potential explanation is that, because only one spirometry was available for analysis in GenIII (compared with two or more in FOC and CARDIA), a significant number of participants might have been misclassified in GenIII. We do not think that this is the case because this might have overestimated (not underestimated) the prevalence of participants with low lung function in GenIII. Further, we calculated that the proportion of individuals misclassified as having low lung function if one rather than two or more measures were used for stratification would have been 10% in FOC and 5% in CARDIA. It is therefore, unlikely, that a similar number in GenIII would have altered the main results. Additionally, the GenIII cohort started about 30 years later than FOC and about 17 years later than CARDIA. During this long period of time, many environmental factors (eg, smoking prevalence, air pollution, or diet) have changed significantly. For instance, findings from one study[31] showed that air pollution levels have chronic, adverse effects on lung development in children from the age of 10–18 years, so the proportion of individuals with FEV_1 of less than 80% predicted was 1·6% at the lowest level of exposure compared with 7·9% at the highest (p=0·002). 10 years later, this same group of investigators also showed that the progressive decrease in air pollution levels that occurred after the implementation of air quality control policies in southern California was associated with statistical and clinical improvements in respiratory health in children.[32] In any case, although we cannot ascertain with certainty the causes of a reduced prevalence of participants with low lung function in early adulthood in GenIII, observations basically reproduced those of FOC and CARDIA.

Because variables included in the analysis vary between cohorts, their pairwise comparison is limited. However, within each cohort, observations comparing individuals with low and normal lung function were consistent, showing that individuals with low lung function always had a higher proportion of cardiorespiratory and metabolic abnormalities. We also cannot exclude a potential detection bias because individuals with low lung function and poorer health might be seen more often in the health-care system, presenting increased

opportunities for tests and diagnoses, and hence treatments for comorbid conditions. This potential bias does not detract from the basic observation of this study that individuals with low peak lung function in early adulthood have increased prevalence and incidence of comorbidities, as well as premature death. The cross-sectional nature of comparisons in early and late adulthood between individuals with low and normal lung function in early adulthood makes inference about causation challenging, although the reproducibility of observations in the three cohorts studied make them more robust. Finally, because many comparisons were explored in this study, some of them might have achieved statistical significance by chance. However, the reproducibility of observations in the three cohorts studied argues against this possibility, as do the findings of a 2017 analysis of the Tucson Epidemiological Study of Airway Obstructive Disease, which also showed that low levels of FEV_1 achieved by the age of 21–35 years predict risk of early mortality.[23]

The results of this study can have clinic and public health implications because they show that abnormal spirometry results in early adulthood, a cheap and reproducible test, has the potential to identify a group of individuals at high risk of having earlier comorbidities and premature death. Although we did not do a predictive risk modelling study that determines the sensitivity, specificity, positive and negative predictive values of spirometry in this particular setting, these are well established in practice where spirometry is routinely used for diagnosis and treatment of many respiratory diseases.[33] Spirometric evaluation of the general population at an early age (at school or, when applying for a driving licence) can potentially help in the identification of this high-risk group of individuals in whom to establish the appropriate preventive measures, monitor health status regularly and closely, and implement therapeutic measures as early as possible when needed.[7,24] Alignment of spirometric testing to a highly focused and effective educational campaign on the dangers of smoking might likely have a bigger impact than either in isolation.[34]

In conclusion, low lung function (FEV_1 <80% predicted) in early adulthood (aged 25–40 years) occurs in 4–13% of the general population and is associated with increased prevalence and earlier incidence of respiratory, cardiovascular, and metabolic comorbidities, and premature death.

Contributors

AA conceived the hypothesis. GN and RF had access to raw data and did the statistical analysis. AA and RF wrote the first draft of the report with input from GN and JB. All authors contributed intellectually to the content of the paper and approved the final published version.

Declaration of interests

AA reports grants and personal fees from AstraZeneca, GlaxoSmithKline, and Menarini; grants from MSD; and personal fees from Novartis, TEVA, and Chiesi, outside of the submitted work. JB reports personal fees from Livanova, Saint Jude, Medtronic, and Boston Scientific, outside of the submitted work. GN and RF declare no competing interests.

Acknowledgments

This study was supported by FIS (PI15/00799-CP16/00039), SEPAR (PI065/2013-PI068/2015), FI-DGR 2016-CERCA, and AstraZeneca Foundation Young Researcher Award. We thank Ferran Torres (Medical Statistics Core Facility, Institut d'Investigacions Biomédiques August Pi i Sunyer [IDIBAPS] Hospital Clínic, Barcelona, Spain) for providing statistical support.

References

1 Vogelmeier C, Agusti A, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2017 report). 2017. http://goldcopd.org/gold-2017-global-strategy-diagnosis-management-prevention-copd/ (accessed Nov 16, 2016).

2 Fletcher C, Peto R. The natural history of chronic airflow obstruction. *Br Med J* 1977; **1:** 1645–48.

3 Lange P, Celli B, Agusti A, et al. Lung-function trajectories leading to chronic obstructive pulmonary disease. *N Engl J Med* 2015; **373:** 111–22.

4 Jobe AH, Whitsset JA, Abman SH. Fetal and neonatal lung development. Clinical correlates and technologies for the future. New York, NY: Cambridge University Press, 2016.

5 Martinez FD. Early-life origins of chronic obstructive pulmonary disease. *N Engl J Med* 2016; **375:** 871–78.

6 Bush A. Lung development and aging. *Ann Am Thorac Soc* 2016; **13** (suppl 5)**:** S438–46.

7 Bolton CE, Bush A, Hurst JR, Kotecha S, McGarvey L. Lung consequences in adults born prematurely. *Thorax* 2015; **70:** 574–80.

8 Barker DJ, Godfrey KM, Fall C, Osmond C, Winter PD, Shaheen SO. Relation of birth weight and childhood respiratory infection to adult lung function and death from chronic obstructive airways disease. *BMJ* 1991; **303:** 671–75.

9 Paneth N, Susser M. Early origin of coronary heart disease (the "Barker hypothesis"). *BMJ* 1995; **310:** 411–12.

10 Cameron N, Demerath EW. Critical periods in human growth and their relationship to diseases of aging. *Am J Phys Anthropol* 2002; **119:** 159–84.

11 Silverman EK, Speizer FE, Weiss ST, et al. Familial aggregation of severe, early-onset COPD: candidate gene approaches. *Chest* 2000; **117** (suppl 1)**:** S273–74.

12 Obeidat M, Hao K, Bosse Y, et al. Molecular mechanisms underlying variations in lung function: a systems genetics analysis. *Lancet Respir Med* 2015; **3:** 782–95.

13 Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. *Am J Epidemiol* 1979; **110:** 281–90.

14 Thyagarajan B, Jacobs DR, Apostol GG, Smith LJ, Lewis CE, Williams OD. Plasma fibrinogen and lung function: the CARDIA Study. *Int J Epidemiol* 2006; **35:** 1001–08.

15 Tsao CW, Vasan RS. The Framingham Heart Study: past, present and future. *Int J Epidemiol* 2015; **44:** 1763–66.

16 Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general US population. *Am J Respir Crit Care Med* 1999; **159:** 179–87.

17 Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat* 1982; **10:** 1100–20.

18 Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* 2015; **12:** 115–21.

19 Henderson AJ. The child is father of the man: the importance of early life influences on lung development. *Thorax* 2014; **69:** 976–77.

20 McGeachie MJ, Yates KP, Zhou X, et al. Patterns of growth and decline in lung function in persistent childhood asthma. *N Engl J Med* 2016; **374:** 1842–52.

21 Berry CE, Billheimer D, Jenkins IC, et al. A distinct low lung function trajectory from childhood to the fourth decade of life. *Am J Respir Crit Care Med* 2016; **194:** 607–12.

22 Bui DS, Burgess JA, Lowe AJ, et al. Childhood lung function predicts adult chronic obstructive pulmonary disease and asthma-chronic obstructive pulmonary disease overlap syndrome. *Am J Respir Crit Care Med* 2017; **196:** 39–46.

23 Vasquez MM, Zhou M, Hu C, Martinez FD, Guerra S. Low lung function in young adult life is associated with early mortality. *Am J Respir Crit Care Med* 2017; **195:** 1399–401.

24 Bush A. Low lung function in young adult life is associated with early mortality. *Am J Respir Crit Care Med* 2017; published online Sept 19. DOI:10.1164/rccm.201707-1416LE.

25 Faner R, Gutierrez-Sacristan A, Castro-Acosta A, et al. Molecular and clinical diseasome of comorbidities in exacerbated COPD patients. *Eur Respir J* 2015; **46:** 1001–10.

26 Allinson JP, Hardy R, Donaldson GC, Shaheen SO, Kuh D, Wedzicha JA. Combined impact of smoking and early life exposures on adult lung function trajectories. *Am J Respir Crit Care Med* 2017; published online May 21. DOI:10.1164/rccm.201703-0506OC.

27 Soler AM, Wain LV, Repapi E, et al. Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function. *Am J Respir Crit Care Med* 2011; **184:** 786–95.

28 Coxson HO, Chan IHT, Mayo JR, Hlynsky J, Nakano Y, Birmingham CL. Early emphysema in patients with anorexia nervosa. *Am J Respir Crit Care Med* 2004; **170:** 748–52.

29 Barbera JA, Peinado VI. Disruption of the lung structure maintenance programme: a comprehensive view of emphysema development. *Eur Respir J* 2011; **37:** 752–54.

30 Barabasi AL. Network medicine—from obesity to the "diseasome". *N Engl J Med* 2007; **357:** 404–07.

31 Gauderman WJ, Avol E, Gilliland F, et al. The effect of air pollution on lung development from 10 to 18 years of age. *N Engl J Med* 2004; **351:** 1057–67.

32 Gauderman WJ, Urman R, Avol E, et al. Association of improved air quality with lung development in children. *N Engl J Med* 2015; **372:** 905–13.

33 Quanjer PH, Stanojevic S, Cole TJ, et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J* 2012; **40:** 1324–43.

34 Pavord ID, Beasley R, Agusti A, et al. After asthma: redefining airways diseases. *Lancet* 2017; published online Sept 11. http://dx.doi.org/10.1016/S0140-6736(17)30879-6.

General Discussion

The main findings of the two original papers that form the core of this PhD Thesis are that:

1) Episodes of ECOPD are characterized by (1) a wide range of bioclinical variables significantly altered at exacerbation (including lung physiology, vital signs, microbiology, lung inflammation, CT imaging as well as biochemistry, systemic inflammation and cardiovascular variables), (2) a panel of biomarkers - comprised of increased levels of dyspnoea ($\geqslant$5 on an analogue visual score from 0 to 10), C-reactive protein level ($\geqslant$3 mg/L -1) and $\geqslant$70% circulating neutrophils - that has a high predictive value for ECOPD diagnosis (AUC = 0.97) and by, (3) a disruption of the biological correlation network associated with clinical stability.

2) Early adulthood low peak lung functions is clearly associated with global increased health risks later in life. These observations were first made in the Framingham Offspring Cohort (FOC) and then reproduced in CARDIA (independent cohort) and GenIII (which includes the direct descendants of FOC participants). In all cohorts a sizeable proportion of individuals (in the range 4-12%) had FEV1<80% ref. at 25-40 years of age and were therefore classified as having Early adulthood Low peak Lung Function (ELLF). Analysis of the data revealed that: (1) they have, vs. those with Early adulthood Normal peak Lung Function (ENLF), a higher prevalence of respiratory, cardiovascular and metabolic abnormalities in early adulthood; and, (2) they also have a higher and earlier (about a decade) incidence of co-morbidities during follow-up as well as an increased all-cause mortality (Hazard Rate (HR) 2.3 [95% CI 1.4-3.7], p=0.001). Finally, in GenIII we observed that: (3) individuals with at least one parent stratified as ELLF in FOC had lower FEV1 in early adulthood ($p<0.0001$); and early adulthood FEV1 of GenIII participants was related ($p<0.0001$) to their FOC parents average FEV1 also in early adulthood.

All in all, these observations indicate that COPD heterogeneity refers to both cross-sectional differences between patients as well as longitudinal variations at several time scales, months in the case of exacerbations and throughout life for the lung function trajectories that

lead to COPD. These disparities identify subgroups of patients that have on average significant differences in aetiology, pathobiological patterns, as well as different clinical implications such as differential prognosis, comorbidity susceptibility and therapeutic targets.

The challenge of investigating the heterogeneity of complex conditions lies in the integration of clinical and biological variables that have distinct properties, in terms of proportion of variance explained, distribution shape, effect size that can be considered clinically relevant, normalization requirements, redundancy, etc. Clinical data is often lacking in details, reported as discrete (e.g. categorical) information, may contain missing values whose imputation is not trivial, or be of imprecise quantification (e.g. smoking exposure, one of the most central risk factors, is rarely accurately reported in longitudinal studies since at best it consists of self-reported periodic estimations of average consumption). In contrast, biological data usually consists of a large amount of continuous determinations that individually show poor correlations to clinical outcomes and phenotypes (small effect size and high dimensionality). Furthermore, since omics technologies are still in infancy and regularly upgraded, the normalization procedures that need to be applied to the data, as well the statistical methods for differential expression and multi-level integration, are not yet well established and standardized.

In this context, given the research questions and cohorts data of this PhD Thesis, standard statistics methodologies as well as a networks approach were applied to the study of ECOPD heterogeneity, while standard statistical tools only were used for the study of low peak lung function in early adulthood in relation to health in later adulthood. The networks approach to the study of correlation structures presents several specific advantages and drawbacks:

- As mentioned, networks are an accurate representation of biological systems since these consist of interacting parts in a global dynamic system. Networks allow for the visualization of mechanistic pathways and subsystems (modules) that are perturbed in disease states (e.g. ECOPD versus clinical stability).
- In contrast to standard (mixed effects or logistic regression) models, correlation networks explicitly lay out all the relationships (collinearity) between covariates.

- Due to the high number of associations estimated (between every pair of variables), networks approaches need to comply with several requirements in order to filter out false positives (correlations significant by chance only): a high sample size, the use of multiple p-value adjustment methods [99], and optionally the application of custom bootstrapping methods.
- A correlation (calculated with Pearson/Spearman method or similar) only reveals an association and as such do not necessarily imply causality, nor give the direction of causality if any (which or the two variables is the cause or consequence of the other).
- The comparison of correlation networks and their interpretation are challenging, in part because of the lack of causality information, but also because a cut-off based on p-value or correlation strength (such as R^2 or odds-ratio) has to be arbitrarily chosen below which weaker correlations are not shown. The density of the networks is thus arbitrary (as well as depends heavily on the sample size) and must be interpreted with caution.

The specific research results of the two core original papers that form this PhD Thesis are discussed below:

1. SPECIFIC DISCUSSION OF THE FIRST AIM: MULTILEVEL CHARACTERIZATION OF COPD EXACERBATIONS

This paper constitutes a proof-of-concept study in which Multi-Level Differential Network Analysis (MLDNA) was applied for the first time to a relevant, complex and heterogeneous clinical problem (ECOPD). Below we further discuss the specific findings and the main limitations of that work.

1.1 Characterization of ECOPD

The reported biological and clinical characteristics determined in core paper 1 both during exacerbation and at convalescence are largely in agreement with previous studies on ECOPD, in terms of individual measures (physiological, biological and microbiological) [100]. Importantly the heterogeneity of exacerbations is made clear in the analysis of the 16 continuous variables that had a significant (bootstrapping False Discovery Rate (FDR) p-

value <0.05) proportion of ECOPD outliers (<5th or >95th percentiles established at convalescence). More than 50% of values of all the study variables at ECOPD still remained within the 5th to 95th range determined at convalescence, indicating a major overlap between the two clinical states. Furthermore, an analysis of the correlations between alterations for all possible pairs of variables reveals that only neutrophils and lymphocytes had their outliers co-occurring in more than 75% of patients. In contrast, all other pairs of variables had at least 25% of their outliers associated with a "normal" value (non-outlier value in the 5%-95% percentiles range of convalescence) in the other pair's variable, pointing out that patients do not all share the same subset of altered variables at exacerbations. Likewise, considering possible causes of the acute event, pathogenic virus or bacteria were not detected from spontaneous sputum at ECOPD in more than 60% of the patients, likely due to the fact that positive cultures were used instead of the qPCR-based techniques (16sRNA sequencing) that tend to provide better detection sensitivity. Of note, the large proportion of outliers in glucose levels is probably caused by systemic steroids taken at or before hospitalization [19, 101]). Additionally, only 5.7% of patients at ECOPD had >2% circulating eosinophils, which increased to 54.7% at convalescence. This differs significantly from other studies where 25% to 50% of patients have >2% circulating eosinophils during ECOPD [78, 102, 103], suggesting that in this cohort the population did not capture eosinophilic-associated exacerbations.

1.2 Biomarkers Diagnostic of Exacerbations

Biomarkers analysis derived from the patients' data showed that ECOPD episodes can be accurately identified (with an AUC of 0.97) by combining the levels of dyspnoea (⩾5 on 1-10 visual analog scale), blood neutrophils (⩾70%) and plasma CRP (⩾3 mg/L) into an optimized general linear mixed model, providing a simple yet reliable diagnostic tool of hospitalized exacerbations. The patients were all recruited at the hospital because of the episode so that the validity of the model to detect non-hospitalized (thus milder) forms of exacerbations could not be investigated. Several other studies [104-108] have defined a higher threshold of CRP for hospitalized acute ECOPD, mostly >10 mg/L, possibly due to a higher severity of the included exacerbations [109]. Needless to say that increased dyspnoea, elevated CRP and leucocytosis can also occur in other clinical circumstances that may not even arise from the lungs (e.g. cholecystitis, pneumonia, sickle cell crisis, pulmonary embolism (PE) or congestive heart failure). The dyspnoea levels reported consist of patients

self-evaluation and the diagnostic would be less subjective if dyspnoea were replaced by a biological biomarker, for example from a serum or sputum sample. For research and clinical use, the computational model requires cross-examination and validation in external cohorts [72]. Maria Montes de Oca et al., in a 2018 review citing our study and others, present different schemes to precisely define exacerbations and propose to use the same three parameters and associated cut-off values of the core paper 1 in addition to three new parameters in order to improve the definition of acute ECOPD. These new parameters are: Procalcitonin >0.25 μg/L (suggestive of bacterial aetiology and encourage the use of antibiotic therapy), N-terminal pro b-type natriuretic peptide (NT-proBNP) >300 pg/mL (suggestive cardiac dysfunction [108]) as well as X-Rays (to evaluate the presence of pneumonia) [110].

1.3 Multi-Level Differential Network Analysis of Exacerbations (MLDNA)

With the application of MLDNA, the core paper 1 aimed to draw attention to the usefulness of this new analytical approach to add novel, integrated, dynamic and holistic information to the heterogeneity of a complex biomedical condition. Networks medicine premise is that complex multi-level states like exacerbations or COPD can be viewed as derailed biological systems, or perturbed networks, in which the normal dynamic interactions at the subclinical levels (for example in terms of lung gas exchange pathways and cellular processes) are going through an abnormal state far from ideal homeokinetic operating conditions [16, 111]. In that paradigm, the phenotypic abnormalities observed in patients (i.e. clinical symptoms of exacerbations) are emergent properties of a dysfunctional physiological system that are associated to subclinical alterations and improper biological interactions. These can be represented as perturbed pathways or perturbed correlation networks. In the worst-case scenario where the disease progressively gets worse, the system departs too far from functional equilibrium towards a partial system collapse (i.e. lung respiratory failure, heart attack, etc.) or complete collapse (death). In core paper 1, the perturbed Spearman correlation network of ECOPD with respect to clinical stability indicates a loss of system control and reduced resilience during ECOPD. Most of the correlations that significantly differ between the two states are present at clinical stability while absent at ECOPD. Furthermore, supporting the idea that network modules represent biological subsystems that consist of variables highly connected internally but little to outsiders, it can be noted that network nodes in both clinical states did cluster into sparsely connected modules that appear

biological homogeneous: five modules at ECOPD and six at clinical stability were identified by a common unsupervised unbiased algorithm, and the resulting modules were mostly comprised of node of one or two of the following categories: systemic inflammation, lung inflammation, biochemistry, lung physiology, vital signs, microbiology, CT imaging or cardiovascular. In addition, more than half of the differential correlations between the two states linked different modules at ECOPD, which can be interpreted as a reduction in biological subsystems (modules) co-regulation. In addition, dynamic non-linear systems may have some components (biological mediators or variables) that are more critical than others to their proper regulation. In a network framework, these are referred to as "hubs" and "bottleneck" nodes and are defined by being respectively either more central and more connected than the other nodes, or forming a non-redundant thus important link between modules [112, 113]. Specifically in core paper 1 networks, only one node had more differential links at ECOPD (TNF-α, with n=4 links to other nodes) while there were many nodes with more differential links at convalescence: TGF-β (n=6), KCO (n=5), PAFI (n=5), PaO2 (n=5) and heart rate (n=4), suggesting that these markers are central to the regulatory processes of lung function, and it can be hypothesized that their alteration are more likely to lead to health complications. For pharmaceutical research purposes, the central hubs and bottleneck variables are a priori the ones most susceptible to be relevant, as targeting them with a specific pharmaceutical agent might help returning the network to the clinically stable equilibrium state, or inversely to prevent a healthy network topology from turning into one susceptible to lead to future episodes of exacerbations.

The heterogeneity of exacerbations can then be conceptualized by considering that there is always more than one way for a system to dysfunction and display the same subset of observable symptoms. In the physiological and biological network operating in COPD patients, nominal alterations (such as increased CRP levels) and pressure points (prolonged submission to tobacco toxic particles) have consequences that spread throughout the network via the connections of the different parts and nodes (e.g. via systemic inflammation mediators). Thus exacerbations are not only associated to lung abnormalities but may also be correlated to cardiovascular, metabolic and systemic complications in COPD patients.

Systems biology approaches that involve networks are now widely used in respiratory medicine studies that involve omics data [12], for example to investigate lung transcriptomics

of emphysema [114] or to characterise COPD comorbidities co-occurrence [115]. To my knowledge, it is the first time that networks medicine is used to get further insight into the complexity of a relevant clinical problem like ECOPD. Hopefully, core paper 1 paves the way for network analytical strategies to be applied to other complex biological conditions in respiratory medicine and elsewhere.

Heterogeneity can be investigated with other systems biology approaches than differ from networks medicine and complements it. Another common and useful scheme to look into complex conditions is to perform unbiased clustering of the patients and their data, then look for significant differences between the clusters with the end goal to (in)validate the findings in other cohorts. Such an unbiased cluster approach was successfully used on ECOPD by Bafadhel et al. on a dataset of 182 ECOPD episodes [102]. Unfortunately, the sample size of the core paper 1 cohort (n=86) was too small for that approach.

1.4 Core Paper 1 Limitations

This paper's research has several limitations worth pointing out. It is not clear how much the initiation of systemic corticosteroids, before the collection of biological samples (within 72h after admission) might have modified the inflammatory profile of ECOPD. It must also be noted that only hospitalised ECOPD were included, so that the results are not directly generalisable to other milder (or more severe) forms of ECOPD. Furthermore, patients presented to hospital at various time points in the evolution of an ECOPD and it is unclear where the exacerbations were sampled along that continuum. Additionally, the recovery phase data at 3 months was used as a proxy for COPD clinical stability in our (networks and biomarkers) differential analysis. However, the underlying assumption that the patients' bioclinical status is the same before and after the exacerbations is partially incorrect. The cohort further lacks controls without COPD and COPD patients who do not suffer exacerbations. Finally, the core paper 1 networks analysis must be considered a proof-of-concept study because of the relatively low number of patients included (n=86) so that the findings require validation in larger cohorts. All in all, the exclusion of severe co-morbidity, pneumonia, relatively small sample size and study of hospitalised patients (not ambulatory ECOPD) restrict the generalisability of the results.

2. SPECIFIC DISCUSSION OF THE SECOND AIM: RELEVANCE OF LOW LUNG FUNCTION IN EARLY ADULTHOOD

The analysis of the prevalence, associated biomarkers and clinical relevance of the low lung function in early adulthood, in three large cohorts (FOC, GenIII, and CARDIA) constitutes the follow-up of the vital lung function trajectories described by my group in 2015 [28]. Specifically, the novel information provided is that 4–12% of the general population has low lung function (FEV1 <80% predicted) in early adulthood (25–40 years of age); that this is not a bystander effect because these individuals also have a higher prevalence of respiratory, cardiovascular, and metabolic abnormalities and a higher and earlier incidence of comorbidities during follow-up than those with normal lung function in early adulthood, as well as a higher rate of premature death; and that low lung function status in early adulthood is correlated between parents and direct descendants (GenIII), suggesting genetic heritability.

These results were not derived from a novel networks analysis as in core paper 1 analysis of COPD exacerbations, but instead from standard statistical tools. However, the core paper 2 also analyzed the heterogeneity globally from a multi-level perspective since, on one hand, the analysis processed longitudinal data at multiple time points within adulthood and across population generations, and on the other hand, they integrated multiple biomedical variables (e.g. FEV1 or biomarkers) in correlation with clinical outcomes such as comorbidities and death. These specific aspects and the main limitations are further discussed hereafter:

2.1 Early Life and Pre-Birth Factors of Abnormal Low Lung Function

The traditional hypothesis of COPD aetiology is an accelerated decline of lung function with age mainly caused by prolonged tobacco smoking [27]. That paradigm is now challenged as reports showed that up to half of COPD patients never had a normal peak lung function in early adulthood [30], pointing to a dynamic heterogeneity of the natural history of COPD. The observations of core paper 2 also support the idea that COPD might arise from failure to attain the normal early adulthood spirometric plateau since low peak lung function in early adulthood (25-40 years old) in FOC and CARDIA was significantly correlated to also having an abnormally low peak lung function later in life (50-65 years) and an increased COPD prevalence.

The aetiology can be pushed even further back in time, as low lung function and respiratory symptoms in infancy appear to track from birth into adulthood, and that parental respiratory factors (such as maternal or paternal asthma or maternal smoking) are correlated to children lung health [116]: In 1991 Barker D. J. and colleagues showed that childhood respiratory symptoms such as bronchitis, pneumonia, or whooping cough are associated to a reduced adult lung function [117]. Spirometry measurements in more than a hundred infants in the Tucson Children's Respiratory Study demonstrated that those who had a low (1st quartile) maximal expiratory flows at functional residual capacity (Vmax(FRC)) at birth ended up with statistical lower FEV1, FVC, and forced expiratory flow between 25% and 75% of FVC (FEF25-75) throughout childhood up to early adulthood [118]. A 2016 study of the same cohort, analysing FEV1/FVC trajectories in 599 subjects from 11 to 32 years old, identified with latent class analysis a significant proportion of individuals (9.3%) who had a persistently low trajectory throughout. The latter was associated to more paternal asthma than in the normal trajectory group, as well as more early life lower respiratory illnesses caused by respiratory syncytial virus, physician-diagnosed active asthma at age 32 years and lower VmaxFRC at age 6 years [119]. Similarly, Owens L. et al. derived, from a 2018 longitudinal study that tested lung function from 1 year old to 24 year old in 253 individuals, that the airway framework is laid down in the antenatal period and tracks into adulthood [120] and that childhood low lung function is associated to increased respiratory symptoms later on. They further uncovered two pre-birth factors associated with a lower FEV1 between 6 and 24 years old: maternal smoking and maternal asthma. In CARDIA, paternal asthma had a significant different prevalence between individuals with early adulthood low peak lung Function (ELLF) and individuals with normal peak lung function (ENLF). In 2009, Svanes C. et al. uncovered several "childhood disadvantage factors" significantly associated with lower FEV1 in adulthood, faster lung function decline and higher COPD incidence: maternal asthma, paternal asthma, childhood, asthma, maternal smoking, and childhood respiratory infections [121]. In 2016, a longitudinal study that followed asthmatic children also established childhood asthma as well as specific patterns of abnormal (reduced FEV1) lung function growth as predictors of early adulthood low long function [122]. Multinomial regression on spirometry measurements of 1389 individuals from a Tasmanian cohort at 7 and 45 years old also exposed that the lowest quartile of FEV1 at 7 years was associated with the co-occurrence of asthma and COPD (ACOS) at 45 years old, but not COPD or asthma alone,

wile the lowest quartile of FEV1/FVC ratio at 7 years was associated with ACOS, COPD but not asthma alone [123, 124]. However, it must be noted that the diagnostic of asthma in children is of limited utility since it is very unspecific: it relies only on clinically observable phenotypes that do not use biological objective tests and may include distinct pathobiological endotypes, such as abnormal lung development.

Aside from paternal or maternal asthma and maternal smoking, there are other (pre-) birth factors correlated to abnormal lung development [125], such as low birth weight or premature birth [126]. In FOC, low birth weight (<2.5 kg) was twice as high (20% vs 9%, p=0.021) in participants with low lung function that in those with normal lung function, which corroborates several studies. Barker D. J. hypothesised in 1991 that intrauterine influences that retard foetal weight gain may also irrecoverably impair the growth of the airways [117]. In that study he noted that lower birth weight was associated with worse adult lung function, and that COPD death in adult life was associated with lower birth weight and weight at 1 year. The idea was developed in what is known as the "Barker Hypothesis" [127]: suboptimal foetal development caused by undernutrion lead to permanent changes in structure and metabolism that may be the origin of a number of diseases later in life, including coronary heart disease, stroke, diabetes and hypertension. A 2015 study confirmed with a logistic regression the negative influence of low birth weight on COPD incidence [128], and likewise a 2016 study used a linear mixed model to negatively correlate birth weight, gestational age and gestational maternal smoking to lung function in children [129].

2.2 Tobacco as a Potential Cause of Early Adulthood Low Lung Function

Tobacco (accumulative) smoking is a central adverse factor for COPD and lung function in general, and, as mentioned, maternal smoking is also known to have a negative influence on children's lung function [130-132].

Given the core paper 2 data, it was impossible to unequivocally disentangle the effect of smoking (or maternal smoking, or chronic exposure to smoking) from the effect of early adulthood low lung function on the occurrence of later abnormalities, mainly because maternal smoking, passive exposure during childhood and smoking in adolescence are themselves three known causal risk factors for early adulthood low lung function, reduced

FEV1/FVC and increased airway resistance [133, 134]. In fact, FOC and CARDIA smoking exposure data in the age range 20-40 y.o. revealed that ELLF had significantly more (compared with ENLF), % of ever-smokers (76% vs versus 64% in FOC, 61% vs 53% in CARDIA), number of cigarettes smoked per day (28.8 vs 20.3 in FOC, 7.8 vs 6.2 in CARDIA), as well as possibly an earlier age of smoking onset (16.8 vs 17.5 in FOC, not different in CARDIA: 17.7 vs 17.5). The same tendency was also observed in GenIII, although not statistically significant (possibly because of the lower sample size, or because later generations smoke less on average). Furthermore, a recent paper by Mathew A. R. and colleagues [135] correlated (low rate) tobacco smoking trajectories to severe increase of emphysema risk, which may indicate that ELLF smoked more than ENLF since in GenIII the prevalence of emphysema in early adulthood and during follow-up was significantly higher in ELLF than ENLF. In the former, similarly, the FOC prevalence was also higher during late adulthood (50-65 years old, not measured in early adulthood). All in all, deconvoluting accumulated smoking exposure from early adulthood low peak lung function would require large cohorts spanning from childhood until late adulthood that contain extensive smoking records, including maternal exposure to smoking and infancy data (e.g. birth weight and spirometry).

2.3 Novel Hypothesis: The factors that Cause an Abnormal Lung Development Might Also Compromise the Cardiovascular and Metabolic Systems

Going further that limiting the possible putative effects of early-life factors to the pulmonary system, it can be hypothesized that the factors that lead to a sub-optimal lung development might not be specific to the lungs but instead also cause abnormalities in the cardiovascular and metabolic systems, as well as more premature death [117]. This new hypothesis is consistent with the core paper 2 analysis since early low lung function individuals had higher cumulative and earlier incidence of abnormalities in all these biological systems when contrasted with normal lung function participants, at 25-40 and during follow-up until 65 years of age.

A recent analysis of the Tucson Epidemiological Study of Airway Obstructive Disease confirmed that low FEV1 (and to a lesser extent FVC) in early adulthood is a risk factor for early cardiopulmonary (heart disease or COPD) mortality [136]. A 2018 analysis of the Hertfordshire Cohort by Humphreys J. et al. with more than 2000 participants, for which

perinatal and infant health records were available, as well as medication and chronic diseases data until about 66 years old, reported that early-life factors such as childhood illnesses significantly increased the risk of multimorbidities in later life [137].

Furthermore, low birth weight (LBW) is associated with dysfunction of several organs later in life. It is a risk factor not only for COPD onset [128], but also for metabolic conditions (e.g. diabetes mellitus), circulatory and heart conditions [138], obesity incidence [139], severe steatosis and non-alcoholic steatohepatitis [140]. Boeri L. and colleagues assessed in 2016, in adult males, that LBW predicted higher Charlson Comorbidity Index (CCI) values, a measure of Health-significant comorbidities, as well as more pathologic progressive motility and pathologic sperm morphology [141]. Two recent reviews of the accumulated evidence, combined with other birth and pre-birth adverse factors (maternal/foetal disease states, nutritional deficits/excess, stress, exposure to environmental chemicals, medical interventions) suggest that insults occurring during the perinatal period alter the developmental trajectory of the offspring's cardiopulmonary system [142] and other systems [143] leading to long-term detrimental outcomes that often culminate in adult pathologies.

2.4 (Epi-)Genetic Factors that Lead to Abnormal Lung Development

The core paper 2 research revealed a trans-generational reproducibility of low lung function in early adulthood since it showed a significant correlation or R^2=0.28 between the FEV1 of GenIII participants and their FOC parents's average FEV1. Accordingly, 10% of GenIII participants that had at least one parent stratified as having low lung function in FOC had a FEV1 below 80% predicted, while in contrast the proportion was only 3% in those whom both FOC parents were classified as normal. Furthermore, those GenIII participants with at least one ELLF parent also had a significantly lower FEV1/FVC ratio, a higher proportion of women and more reported parental asthma. These associations suggest that there may be a genetic component to abnormal lung development and resulting early low lung function.

Much of the lung function development and COPD heritability remains unexplained, although several GWAS studies have established gene variants significantly associated with either lung function parameters (e.g. to FEV1, FVC, and FEV1/FVC ratio, longitudinal

variations), tobacco effects on lung decline, or COPD parameters (e.g. onset, or airflow obstruction severity) [144-147]. Recently, McGeachie J., Yates P. and colleagues uncovered a specific genetic polymorphism (rs4445257) associated to early decline in lung function after normal growth that may also protect against early decline in reduced growth groups [148].

Recent research uncovered gene variants that affect lung development as early as the embryonic stage [149, 150], and now that early adulthood peak lung function and trajectories importance for COPD are established, more research is needed into their genetic basis. In terms of epigenetic determinants in early life, a 2018 Epigenome-wide association study of cord blood and mid-childhood peripheral blood total serum Immunoglobulin E (IgE) levels identified several cord blood methylation signals that were correlated to mid-childhood IgE, thus providing evidence that IgE-mediated hyper-sensitivity may be epigenetically programmed in utero and during early childhood [151]. Several of these methylation sites were already associated to asthma (ADAM19, EPX, IL4, IL5RA, and PRG2) [152-154].

These studies lead to new interpretations of COPD pathobiology unrelated to tobacco smoking for a subset of patients, via abnormal early life lung development supported by genetic susceptibility and/or early life adverse programming of epigenetic sites.

2.5 Potential Opportunities for Treatment, Prevention and Early Intervention

The findings of this PhD Thesis suggest that some of the comorbidities frequently reported in COPD patients might originate earlier in life than previously thought, especially for the high proportion (up to 50%) of COPD patients who had a low peak lung function in early adulthood. For clinical practice it means that these individuals might benefit greatly from early detection (potentially via systematic population-wise spirometry tests in childhood and early adulthood), early intervention and targeted preventive measures.

2.6 Core Paper 2 Limitations

In both FOC and CARDIA cohorts, the drop-out rate during follow-up was higher in participants with early low adulthood peak lung function (ELLF) than in the normal peak group individuals (ENLF). It is a potential results bias, although it may underestimate the

observed higher proportion of bioclinical abnormalities in ELLF, as adverse medical conditions and poor health may be the reason of drop-out for a number of cases in these decade-long observational cohorts of the general population.

Additionally, the comparison of observations between the three cohorts (FOC, CARDIA and GENIII) was limited by the fact that most often the nominal biological variables, clinical variables and questionnaires were not the same across cohorts (or even in-between visits of a given cohort). This potential bias was mitigated by summarizing the alterations into their respective clinical category (e.g. respiratory, cardiovascular or metabolic) and then calculating for each category the proportion of individuals that have at least one clinical alterations in any of the category's variables. These accumulative proportions are more robust than nominal variable prevalences and more readily compared between ELLF vs ENLF across cohorts as well as longitudinally during follow-up.

Finally, the associations reported in this Thesis do no establish causation and the observations are prospective. Therefore, the findings require validation and confirmatory analysis in other cohorts, as well as a more detailed analysis of the clinical factors discussed (effects of early life factors on early lung function, causal interactions with tobacco smoking exposure, etc.).

2.7 Futures Challenges of Systems Medicine in this Field

If the lungs develop suboptimally, resulting symptoms of airflow limitation may be diagnosed as asthma [155], which would represent an important misdiagnosis since the underlying pathobiological mechanisms of individuals born prematurely are different from those of common asthmatics [126]. This potential misclassification should be further considered in future studies.

Longitudinally the core paper 2 reports important statistical associations between clinical and biological factors across time, but it does not uncover the biological mechanisms and endotypes that underlie these relationships. Extensive omics data collection and

genotyping in large cohorts of the appropriate clinical setting will provide the necessary basis for the analysis of such mechanisms.

One of the hardest (but major) aspects of NCDs to quantify is the environmental impact on disease initation and development [156]. The environmental variables (termed exposome) range from prenatal events to lifelong exposure variables (e.g. air pollution, low physical activity and adverse diet) that could not be properly tracked in past cohorts. Of note, although cumulative tobacco smoking is a crucial COPD risk factor, its current quantification based on patients self-estimation clearly lacks accuracy. Future advancements in technology (e.g. wearables or drones to continuously track air pollution) will improve the quantification in future prospective (ideally trangenerational) cohorts.

Access to extensive electronic medical records is also important to the proper study of NCDs comorbidities and their of confounding factors (sex, age, socioeconomic status, etc.). The centralized collection of that much individual data (omics data, environmental data and medical records) in large cohorts - arguably necessary to fully understand NCDs heterogeneity - poses substantial ethical challenges, as well as confidentiality, security and legal issues [157].

Finally, it is possible that the future implementation of personalized medicine in healthcare will partly rely on probabilistic models that do not use mechanistic pathobiological information, but instead leverage big data with unbiased machine learning algorithms to predict clinical outcomes and best medication strategies in tools tailored to the profile of individual patients [158, 159].

Conclusions

In conclusion, this PhD Thesis has used multi-level integrated analysis to shed light on two specific aspects of COPD heterogeneity:

1) **Exacerbations of COPD (ECOPD)**

- ECOPD are characterized by several alterations (dyspnoea, tachypnoea, tachycardia and respiratory failure, lung and cardiovascular physiology, systemic inflammation markers, biochemistry markers, sputum bacteria or viral infection), although, for practically all variables, significant overlap remain between the two clinical states distributions, thus highlighting the heterogeneity of the events.
- ECOPD are characterised by a fragmentation of the correlation network observed during convalescence, suggesting loss of system control, homeostasis and reduced resilience.
- These acute events can be identified objectively (AUC 0.97) by using a panel of three biomarkers (dyspnoea, circulating neutrophils and CRP levels) frequently determined in clinical practice.

2) **Early low lung function and health in later life**

- Low peak lung function in early adulthood (FEV1 of less than 80% predicted at the age of 25-40 years) is common in the general population, with a prevalence of 4-12%.
- Early low peak lung function individuals have a higher prevalence of respiratory, cardiovascular, and metabolic abnormalities in early adulthood.
- These individuals also have a higher and earlier (about a decade) incidence of cardiovascular, metabolic and systemic comorbidities in later adulthood.
- They are burdened by an increased risk of premature death (hazard ratio 2.3 [95% CI 1.4-3.7]).

- Low peak lung function status in early adulthood is significantly correlated (R^2=0.28) in-between parents and offsprings, indicating a possible genetic heritability.

Appendix

Published review

From systems biology to P4 medicine: applications in respiratory medicine.

Noell G., Faner R.tc, Agustí A.

Eur Respir Rev. 2018 Feb 7;27(147)

From systems biology to P4 medicine: applications in respiratory medicine

Guillaume Noell[1,2], Rosa Faner[1,2] and Alvar Agustí[1,2,3]

Number 4 in the Series "Personalised medicine in respiratory diseases"
Edited by Renaud Louis and Nicolas Roche

Affiliations: [1]Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [2]CIBER Enfermedades Respiratorias (CIBERES), Barcelona, Spain. [3]Respiratory Institute, Hospital Clinic, Universitat de Barcelona, Barcelona, Spain.

Correspondence: Alvar Agustí, Respiratory Institute, Hospital Clínic, Villarroel 170, 08036 Barcelona, Spain. E-mail: AAGUSTI@clinic.cat

@ERSpublications
Systems biology and network medicine have the potential to transform medical research and practice
http://ow.ly/r3jR30hf35x

Cite this article as: Noell G, Faner R, Agustí A. From systems biology to P4 medicine: applications in respiratory medicine. *Eur Respir Rev* 2018; 27: 170110 [https://doi.org/10.1183/16000617.0110-2017].

ABSTRACT Human health and disease are emergent properties of a complex, nonlinear, dynamic multilevel biological system: the human body. Systems biology is a comprehensive research strategy that has the potential to understand these emergent properties holistically. It stems from advancements in medical diagnostics, "omics" data and bioinformatic computing power. It paves the way forward towards "P4 medicine" (predictive, preventive, personalised and participatory), which seeks to better intervene preventively to preserve health or therapeutically to cure diseases. In this review, we: 1) discuss the principles of systems biology; 2) elaborate on how P4 medicine has the potential to shift healthcare from reactive medicine (treatment of illness) to predict and prevent illness, in a revolution that will be personalised in nature, probabilistic in essence and participatory driven; 3) review the current state of the art of network (systems) medicine in three prevalent respiratory diseases (chronic obstructive pulmonary disease, asthma and lung cancer); and 4) outline current challenges and future goals in the field.

Introduction

Human health and disease are emergent properties of a complex, multilevel biological system that spans from the molecular domain to the microbiome, exposome and social levels (figure 1) [1, 2]. Ideally,

Previous articles in this series: No. 1: Chung KF. Personalised medicine in asthma: time for action. *Eur Respir Rev* 2017; 26: 170064. **No. 2:** Bonsignore MR, Suarez Giron MC, Marrone O, *et al.* Personalised medicine in sleep respiratory disorders: focus on obstructive sleep apnoea diagnosis and treatment. *Eur Respir Rev* 2017; 26: 170069. **No. 3:** Mascaux C, Tomasini P, Greillier L, *et al.* Personalised medicine for nonsmall cell lung cancer. *Eur Respir Rev* 2017; 26: 170066.

Received: Sept 28 2017 | Accepted after revision: Nov 30 2017

Support statement: This work was supported, in part, by Instituto de Salud Carlos III (PI12/01117, PI15/00799, CP16/00039), Recercaixa-2012 (AA084096), SEPAR (PI065/2013, PI192/2012, PI68/2015), AGAUR FI-DGR 2016 PhD grant and AstraZeneca foundation COPD Young Researcher Award. CERCA Programme/Generalitat de Catalunya. This work was developed at the building Centre de Recerca Biomèdica Cellex, Barcelona, Spain. Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: None declared.

Provenance: Commissioned article, peer reviewed.

therefore, if we want to intervene prophylactically to preserve health or therapeutically to cure disease, in a safe and effective way, we should understand these dynamic gene–environment interactions in greater detail. Certainly, this will not be an easy task, but the alliance of new high-throughput "omic" methodologies, novel imaging techniques and current (and future) computational power can project us forward in this endeavour and eventually facilitate the development of novel therapeutic strategies (and the repurposing of old ones) [3]. However, as wisely highlighted by one of the anonymous reviewers of this paper, to whom we are grateful: "… full understanding of complex nonlinear systems in physics and biology might not be ever possible and, fortunately, might not be even required because probabilistic decisions are (and will become) more powerful than decisions based on precise mechanistic understanding. This is a real revolution already happening in society (Google and Amazon can predict your behaviour without knowing (less understanding) you). Similarly, Artificial Intelligence (AI) will be able soon to predict the clinical course and responsiveness to intervention based on probabilities rather than on deep understanding of the system …". We think that both concepts are actually synergistic since a more comprehensive and precise understanding of human biology (figure 1) will, no doubt, feed back to any AI platform, which will in turn provide new hypotheses to test iteratively. In any case, embracing a holistic scientific approach (as opposed to the reductionist research strategy used traditionally) for the understanding of human health and disease is a unique (and mandatory) opportunity to really move medical practice forward in the 21st century.

In this review, we: 1) discuss the principles of systems biology, a relatively recent research strategy that leverages from omics and bioinformatics to gain a holistic understanding of complex biological systems; 2) elaborate on how this can pave the way towards the effective deployment of the so-called "P4 medicine" (predictive, preventive, personalised and participatory) [4], which can shift healthcare from treatment of illness to prediction and prevention of illness, in a revolution that will be personalised in nature, probabilistic in essence and participatory driven; 3) review the state of the art of network (systems) medicine in three prevalent respiratory diseases (chronic obstructive pulmonary disease (COPD), asthma and lung cancer); and 4) outline current challenges and future goals in the field.

Systems biology

System approaches and emergent properties

System approaches stem from the premise that separate analysis of information gathered from different elements, compartments or levels of a dynamic system (figure 1) cannot yield appropriate understanding/

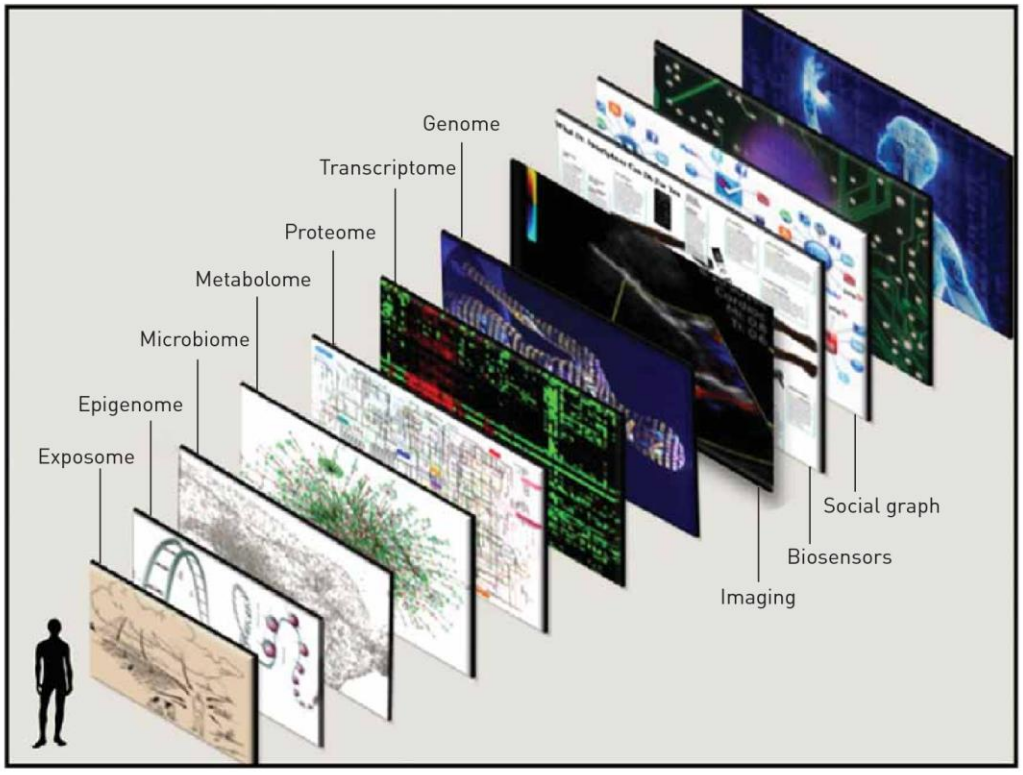

FIGURE 1 Multilevel layers of biological, environmental and social information ideally integrated in systems biomedicine approaches. For further explanations, see text. Reproduced and modified from [2] with permission.

prediction of the global behaviour of the system (so-called emergent properties, which are implicit in nonlinear systems) nor allow to fix it if found globally away from an homeokinetic state (*e.g.* disease *versus* health), with alterations that may spread throughout various levels or compartments of the system [5]. As MACKLEM [5] pointed out, emergent properties arise spontaneously as self-organised order from the nonlinear interactions of the different biological components and thus the overall emergent behaviour transcends the behaviour from each part in isolation. It follows that a more holistic approach, integrating information of the interacting parts and subentities into a single mathematical representation or model, can potentially offer better clues as to the causal chain of events that leads to the apparent phenotypic manifestations and how to remedy the situation [6]. Therefore, systems biology departs from the reductionist approach followed by traditional biomedical research by integrating (rather than taking apart) different biological levels (genes, molecules, cells, organs and the environment) and mechanisms, and shares a very similar goal with integrative physiology: to better understand holistically the systemic dynamic state of individuals [7, 8]. In this context, systems biology (and systems or network medicine) is nothing more than physiology, which has always meant to be multiscale and integrative [7, 8]. The difference is that today's availability of new tools, high-throughput technologies and computing power allows, for the first time, real physiology to be performed. In essence, it is all about perspective [9]. Before "perspective" (*i.e.* three-dimensional) painting was "invented", classical painting considered only two dimensions. Systems biology includes many different biological levels (dimensions) as well as the element of time dynamics. Hence, it has the potential to provide a much better definition for "the eye of the beholder" [9].

Biology as an informational science

In recent decades, faced with the biological complexity of human diseases, biomedical scientists have increasingly turned their efforts to apply high-throughput methodologies that embrace the Cartesian view that the human body is a system of formally interacting parts and that biology is an informational science. A nonexhaustive list of information sources (table 1) includes "omics" data ((epi-)genomic, transcriptomic, proteomic, metabolomic and microbiomic), single-cell analyses, phenotypic assays, extensive medical records and an endless list of environmental factors ("exposome"), such as smoking, exercise, diet and pollution, among others (figure 1). Common respiratory-specific levels of information are lung function and imaging.

System representation: networks

A network (or graph) is a practical graphical representation of complex data in the context of systems approaches (figure 2), where nodes are the elements of the system under study (*e.g.* genes, proteins, biochemical or physiological measures, individuals or patients, among many others) and edges (or links) connect nodes that interact somehow (causality, correlation). The network(s) constructions are hypothesis driven, *i.e.* there is not a single, fixed, network "template"; on the contrary, they can be "custom-made". Networks are used to make inferences regarding the emergent dynamic (spatial and temporal) behaviour of the system in response to perturbations of putative critical network elements (nodes and/or edges).

Diseases as network perturbations

Any disease can be viewed as a system in an abnormal state (a perturbed network) far from homeokinetic operating conditions [5], either with: 1) associated nonemergent (*i.e.* subclinical) alterations, or 2) observable phenotypic abnormalities (*i.e.* clinical symptoms) progressively departing from functional equilibrium towards partial system collapse (*i.e.* organ failure, *etc.*) or complete collapse (death). In opposition, perfect health, or wellness, can be viewed as the optimal and quantifiable state of a system in dynamic equilibrium (*i.e.* homeokinesis [5]).

Biological network properties

Several aspects of biomedical networks are due to their particular biological nature and must always be considered in a research setting [16]. In terms of "topology" (*i.e.* their spatial distribution) they are generally scale-free (as opposed to random networks). In this setting, "scale-free" means that this type of network contains many nodes with few connections and a few nodes with many links (hubs) (figure 2). This topology makes networks more robust against random perturbations [17] because of their higher modularity [18]. They are composed of loosely connected subparts (modules), which are groups of nodes highly connected internally but little to outsiders. Modules are usually coupled with specialised biological subtasks. Additionally, not all nodes are equal relative to the network structure. Central elements that are much more connected than the average are denominated "hub" nodes [19], while linkers between modules are termed "bottleneck" nodes (figure 2) [20]. Perturbations of these elements (hubs and bottlenecks) often alter the system behaviour drastically, whereas the impact of more peripheral nodes on systems behaviour (emergent properties) is often marginal. Other influential network properties with regard to the

TABLE 1 Common omics data types

	Assay	Platform	Main advantages and disadvantages	Standard bioinformatics pipelines
Genomics	Identify nucleotide variants (SNPs) in the whole genome associated with clinical traits (GWAS)	Genotyping arrays, whole-exome sequencing	SNP variability is stable during life; provides limited information in complex diseases due to several loci implicated	GWAS protocol review [10]
Transcriptomics	Quantify expression levels of cellular transcripts (*e.g.* mRNA)	Expression arrays, RNA sequencing	Widely used due to its high information content on cell status; differences in mRNA expression do not imply differences in proteins; does not take into account post-transcriptional modifications	RNA sequencing pipelines review [11]
Proteomics	Characterise protein expression levels of cells/samples	MS-based approaches	Expected to be closer to the phenotype; not widely used, expensive and more cumbersome analysis	Next-generation proteomics review [12]
Metabolomics	Characterise abundance profile of metabolites and their relative ratios	MS-based approaches	Representative of the cellular status; applicable to many biological fluids (*i.e.* breath, blood, urine, *etc.*); not widely used	Review of analytical methods for metabolomics [13]
Epigenomics	Determine modifications in DNA and small RNA that interfere with gene expression	DNA methylation analysis with arrays (Infinium MethylationEPIC 850K; Illumina, San Diego, CA, USA), next-generation sequencing, small RNA sequencing, arrays, *etc.*	Provides additional information to transcriptomics; related to exposures; more expensive than transcriptomics; sequencing-based approaches have computational tools in active development	Bioinformatics aspect of DNA methylation studies [14]
Microbiomics	Characterise bacterial (and viral) composition of a sample	Targeted sequencing of 16S rRNA gene, shotgun metagenomics sequencing	Provides information of external factors likely to be associated with disease; 16S sequencing does not differentiate between the presence of live/dead bacteria	Bioinformatics analysis for the characterisation of the human microbiome [15]

SNP: single nucleotide polymorphism; GWAS: genome-wide association study; MS: mass spectrometry.

robustness of the system include "redundancy" and "degeneracy" [21]. Finally, nodes and edges may be characterised qualitatively (*e.g.* fold-change sign for nodes that represent gene products) or quantitatively (*e.g.* chemical binding constant for edges that connect drug ligands to their target molecules) (figure 2).

Medical uses

Although systems biology is best suited for experimental models of disease, it can also provide actionable and useful insights in clinical medicine [22–24]. Systems (network) medicine can lead to the identification of disease biomarkers or drug targets, both defined as key nodes whose perturbation transits the state of the biological system from health to disease or *vice versa*. A paradigmatic example comes from the field of cancer and the observation that the sequential use of anticancer drugs enhances cell death by rewiring apoptotic signalling networks [25].

P4 medicine

The holistic approach of systems biology discussed earlier has enabled the emergence of a new comprehensive paradigm in medicine, called P4 medicine, for predictive, preventive, personalised and participatory [4, 26–28].

From treatment to prediction and prevention

Current western medicine mostly focuses on treating diseases and symptoms when they appear. Thus, the current healthcare system organisation (and its major stakeholders, *i.e.* hospitals and primary care centres, pharmaceutical industry, insurance companies, policy makers, providers (*e.g.* physicians) and patients) is based on the provision of medication and related health products to individuals once they are sick and

biosensors continuously tracking essential variables, such as exhaled breath [39], urine [40], imaging [41] and/or ambient pathogens or allergens [42–44].

Participatory driven

Finally, the benefits of this new P4 medicine will only be possible if patients and healthy subjects become active agents in the continuous assessment and preservation of their health. The role of health providers, both traditional (physicians, nurses, physiotherapists) and novel (genetic counsellors, behavioural coaches), will evolve to facilitate actionable information to individuals, which they can use to maintain their health [45]. Importantly, a new legal framework of rights, obligations and protections for individuals/patients and health professionals alike remains to be established and implemented. The emergence of personalised "big" data repositories raises unprecedented ethical, privacy, confidentiality, security and policy issues related to information ownership, access and management. Of note, the insurance company regulatory framework is markedly unprepared in most countries.

How to do it?

Research strategy

In principle, there are two different approaches to analyse data in this setting: "supervised" analysis based on *a priori* knowledge (*e.g.* clinical characteristics of patients) and "unsupervised" analysis (*i.e.* hypothesis-free). Both strategies have advantages and disadvantages, and in a sense they are complementary; their characteristics are further discussed in the Analytical complexity section.

Input data

Systems biology leverages from several omics data types. The most commonly used data types are genomics, transcriptomics, proteomics, metabolomics, epigenomics and microbiomics. Table 1 summarises their definitions, available experimental platforms, advantages/disadvantages and the bioinformatics tools needed. In each omic, data is curated, normalised and the differences between groups are usually computed using general linear models [46, 47]. We acknowledge that exposomics and imaging are missing in table 1; this is on purpose as both fields are currently developing very actively [48, 49].

Analytical complexity

Single-level analysis

A common research approach is to perform standard (supervised or unsupervised) single-level omic analysis (table 1) and then use further bioinformatics tools to facilitate the translational interpretation (table 2 and figure 3). For instance, from a list of genes/proteins of interest, in order to identify underlying biological mechanisms, functional enrichment can be performed against many databases that host annotated information on functional roles (figure 3d): Gene Ontologies of biological processes, cellular components or molecular functions [62], KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways [63], Reactome pathways [64] and gene set enrichment analysis (GSEA) [50]. Furthermore, the

TABLE 2 Widely used tools to facilitate biomedical interpretation from omics analysis

Analytical tool	Goal	Advantages and disadvantages	Pipelines
Functional enrichment	From lists of identifiers (commonly genes) computes the over-representation in a specific molecular function, Gene Ontology, pathway, biological process, cell localisation, *etc.*	Noise and dimension reduction, helps interpret gene sets; useful to aggregate the individual gene contribution to overall changes; results are dependent on database knowledge and thus may be biased	Gene set enrichment analysis (GSEA): http://software.broadinstitute.org/gsea/index.jsp [50]; gene set variation analysis (GSVA) [51]; Enrichr: http://amp.pharm.mssm.edu/Enrichr [52]; FunRich: http://funrich.org [53]; STRING: https://string-db.org [54]
Data clustering	Classifies samples/variables based on their similarity in order to obtain homogeneous groups	Unsupervised, data driven and probabilistic; requires medium/large data sets	*k*-means [55, 56]; hierarchical bottom-up [57]; hierarchical top-down (divisive analysis clustering (DIANA)) [58]
Coexpression networks	From the dataset builds a correlation network to identify groups of related genes (modules), which can be investigated for biological functions and/or related to clinical traits	Coexpression in order to reflect causative processes must be coupled with functional enrichment and validation; correlations are affected by sample size of the dataset; requires proper data normalisation	Weighted gene coexpression networks analysis (WGCNA) [59]; conventional coexpression measures (Pearson/Spearman/Kendall, mutual information [60]); miRNA (targets)–genes [61]

with any imputed (usually clinical) characteristic. WGCNA can be complemented with functional enrichment analysis.

Multilevel analysis: the true revolution

Although some studies have and will continue to work successfully on a single omic level, recent decades have seen an ever-increasing body of work where several distinct omics datasets, including also other biological or clinical levels, are analysed conjointly using multiscale integrative methods such as SNF (similarity network fusion) [69]. This combination of levels has the potential to provide researchers with simultaneous information from several compartments of the biological system of interest, thus facilitating the modelling of the dynamic nonlinear relationships that characterise emergent properties (phenotypes) and complex diseases. Accordingly, this strategy would be able to provide more power to identify groups of patients affected with the same pathobiological mechanism or more power to probabilistically model (without understanding) the health *versus* disease states. The main multiscale analytical tools described to date are summarised in table 3. The "supervised" methods can be grouped mostly into either network-based, machine learning or multistep approaches [86], while the "unsupervised" can be further classified as based primarily on networks, Bayesian approaches or matrix factorisation (table 3).

Current applications of systems approaches in respiratory medicine

The pathogenesis of most common respiratory diseases is complex and largely undefined from a precise pathobiological point of view. Chronic respiratory conditions, such as asthma or COPD, are still diagnosed (and treated) based on respiratory symptoms and traditional lung function measures, but they are highly heterogeneous and often overlap. In fact, they are the end result of complex genetic and environmental interplays that are yet to be explicitly modelled. This poorly defined characterisation of the basic disease mechanisms results in nonspecific, mostly symptom-driven treatment options, or lack thereof, that may eventually be able to slow the progression of these diseases in fortunate, responsive patients.

Systems biology and network medicine approaches are being put forth in an effort to palliate this painful lack of knowledge and understanding by tackling two fundamental and interrelated matters: 1) as in other biomedical fields such as cancer, a novel classification (*i.e.* "taxonomy") of chronic airway diseases is needed, based not on clinical presentation (*i.e.* "phenotypes") but instead either on the underlying biological mechanisms (*i.e.* "endotypes") when characterised or resulting directly from data-driven probabilistic clustering of patients data; and 2) a more precise patient stratification that can be transferred to distinct and personalised preventive or therapeutic prognosis as well as improved prognosis (*i.e.* P4 medicine) is also needed, as recently highlighted in a review focused on biological therapies for airway diseases [87].

COPD

COPD is a heterogeneous disease with pulmonary and extrapulmonary manifestations [88], and variable response to pharmacological treatment [89], suggesting that the condition affects several distinct biological pathways. To characterise this heterogeneity at the molecular level, several studies have already used a number of different systems approaches. 1) WGCNA and GSEA showed that a molecular signature composed of gene modules related to B-cell activity, NK-cell activity or viral infection cellular markers might be detectable in peripheral blood months following COPD exacerbations [90]. 2) Xue *et al.* [91] used other network-centric procedures to reveal an unexpected loss of inflammatory signature in COPD patients, as well as an activation-independent core signature for human and murine macrophages. 3) Glass *et al.* [92] used the network inference analysis PANDA (Passing Attributes between Networks for Data Assimilation) [93], designed for improved integration of individual with public datasets, and discovered network rewiring of lymphocyte activation signalling circuits in a known gene variant implicated in COPD by genome-wide association studies. 4) Faner *et al.* [94] unravelled differences in the molecular pathogenesis of emphysema and bronchiolitis by performing correlation network analysis of lung transcriptomics on COPD patients. They found that B-cell-related genes were significantly enriched in emphysema (compared with COPD patients without emphysema), paving the way for differential therapeutic research on inflammatory pathways of the adaptive immune response. 5) Two COPD studies demonstrated the utility of unsupervised *k*-means clustering by identifying robust cluster associations with clinical characteristics and known COPD genetic variants [95, 96]. 6) Very recently, Ross *et al.* [97] introduced a new Bayesian method for COPD subtyping. They applied it to the COPDGene cohort and identified nine different patient subgroups with distinct disease progression trajectories. Of note, Ross *et al.* [97] prove that their sophisticated model has a better predictive capacity than multivariate ordinary least squares regression analysis.

clinical and biomarker profiles (from blood, sputum and airway data) [98]. 2) Kuo *et al.* [99] recently reported three novel molecular phenotypes of asthma in the U-BIOPRED cohort by analysing sputum cell transcriptomics in asthmatic and nonasthmatic subjects. They applied hierarchical clustering of differentially expressed genes as well as gene set variation analysis, gene–protein coexpression and pathway enrichment analysis. 3) Sharma *et al.* [100] used network-based tools to analyse the predictive value of the asthma interactome, and characterised high-impact pathways central to the disease heterogeneity and drug response. 4) Qiu *et al.* [101] used PANDA on participants of the Childhood Asthma Management Program cohort to assess the differential connectivity between the gene regulatory network of good responders to inhaled corticosteroids *versus* that of poor responders. The method allowed them to integrate their dataset with public data interactions of genes, transcription factors and proteins, and eventually implicate several network hubs and transcription factors (as well as regulatory rewiring) in the heterogeneity of drug treatment effects. Specifically, the differential network topology of good responders *versus* that of poor responders revealed enriched corticosteroid-induced pro-apoptosis pathways in the former and anti-apoptosis pathways in the latter, as well as key regulatory transcription factors (hubs) that drove differential downstream gene expression in the two groups.

Lung cancer

Lung cancer is the leading cause of cancer death in the world. Lung cancer is highly heterogeneous genetically because of a high mutation rate, as well as extremely complex since it comprises a disparate subset of diseases with distinct and possibly overlapping pathobiologies that share a common phenotypic manifestation. Smoking is a core shared risk factor for COPD and lung cancer; up to 65–70% of lung cancer patients suffer both lung cancer and COPD [102, 103]. So far, no single satisfactory circulating (*i.e.* liquid biopsy) tumour marker has been properly validated, but recently a panel of six tumour markers showed a very high specificity and sensitivity in patients referred to a tertiary hospital because of the clinical suspicion of lung cancer [104, 105]. Given that inherited genetic variants play a significant role in lung cancer development [106], but contribute little to risk estimates of classical predictive statistical models [107–109], it is hoped that systems biology approaches will allow the comparison multilevel high-throughput omics data between tumour and normal tissue, and facilitate the identification of early diagnostic lung cancer biomarkers. WGCNA has already been used successfully in lung cancer research. 1) Tang *et al.* [110] related the gene expression profile of lung squamous cell carcinoma with five differentially expressed long noncoding RNAs that could help in prognosis evaluation. Their gene signature was statistically associated with overall survival in important clinical subsets (stage I, epidermal growth factor (EGFR) wild-type and EGFR mutant). 2) Tian *et al.* [111] analysed coexpression networks and protein–protein interactions of data available in public repositories (The Cancer Genome Atlas, KEGG and Gene Ontology).

What's next? Future challenges

For the successful development and implementation of systems biology and network medicine approaches in respiratory medicine, several challenges need to be faced and eventually solved.

Technical challenges

In any clinical study, only a fraction of the biological variability is captured (and therefore analysed) due to technical limitations (and cost) of the experimental tools available. The development of new experimental tools (*e.g.* high-throughput next-generation sequencing, mass spectrometry-based flow cytometry or real-time molecular imaging) will generate new information but, at the same time, massive amounts of (big) data that will have to be adequately handled, analysed and interpreted [112–114]. In this context, Riekeberg and Powers [115] recently reviewed the methodological advancements and successful applications of metabolomics, one the newest omic fields.

However, research would benefit not only from measuring "more" relevant variables, but also from estimating with better precision those variables already determined in the context of a more complete definition of reference and pathological ranges (that vary in time, across individuals and biological codeterminants) [116]. Of the variability supposedly present in experimental data, these currently unaccounted factors and batch effects should not be underrated since they can partly explain the general difficulty to replicate scientific findings in the biomedical field, of which respiratory biomedicine is not exempt.

Computational challenges

Computational methodologies and programming analytical tools are being constantly refined as they translate advancements from complementary areas such as AI and information science. However, challenges and difficulties remain. For instance, in differential expression (omics) analysis, one of the main

7 Greenhaff PL, Hargreaves M. 'Systems biology' in human exercise physiology: is it something different from integrative physiology? *J Physiol* 2011; 589: 1031–1036.
8 Grocott MP. Integrative physiology and systems biology: reductionism, emergence and causality. *Extrem Physiol Med* 2013; 2: 9.
9 Snyder LJ. Eye of the Beholder: Johannes Vermeer, Antoni van Leeuwenhoek, and the Reinvention of Seeing. New York, Norton, 2015.
10 Hamberg M, Backes C, Fehlmann T, *et al.* MiRTargetLink – miRNAs, genes and interaction networks. *Int J Mol Sci* 2016; 17: 564.
11 Teng M, Love MI, Davis CA, *et al.* A benchmark for RNA-seq quantification pipelines. *Genome Biol* 2016; 17: 74.
12 Altelaar AF, Munoz J, Heck AJ. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet* 2013; 14: 35–48.
13 Nobakht M Gh BF, Aliannejad R, Rezaei-Tavirani M, *et al.* The metabolomics of airway diseases, including COPD, asthma and cystic fibrosis. *Biomarkers* 2015; 20: 5–16.
14 Adusumalli S, Mohd Omar MF, Soong R, *et al.* Methodological aspects of whole-genome bisulfite sequencing analysis. *Brief Bioinform* 2015; 16: 369–379.
15 Noecker C, McNally CP, Eng A, *et al.* High-resolution characterization of the human microbiome. *Transl Res* 2017; 179: 7–23.
16 Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. *Nat Rev Genet* 2016; 17: 615–629.
17 Barabasi AL, Albert R. Emergence of scaling in random networks. *Science* 1999; 286: 509–512.
18 Mitra K, Carvunis A-R, Ramesh SK, *et al.* Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet* 2013; 14: 719–732.
19 Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* 2005; 22: 803–806.
20 Yu H, Kim PM, Sprecher E, *et al.* The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 2007; 3: e59.
21 Baffy G, Loscalzo J. Complexity and network dynamics in physiological adaptation: an integrated view. *Physiol Behav* 2014; 131: 49–56.
22 Schadt EE, Björkegren JLM. NEW: network-enabled wisdom in biology, medicine, and health care. *Sci Transl Med* 2012; 4: 115rv111.
23 Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature* 2009; 461: 218–223.
24 Silverman EK, Loscalzo J. Network medicine approaches to the genetics of complex diseases. *Discov Med* 2012; 14: 143–152.
25 Lee MJ, Ye AS, Gardino AK, *et al.* Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* 2012; 149: 780–794.
26 Hood L, Heath JR, Phelps ME, *et al.* Systems biology and new technologies enable predictive and preventative medicine. *Science* 2004; 306: 640–643.
27 Weston AD, Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* 2004; 3: 179–196.
28 Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol* 2012; 29: 613–624.
29 Leadley RM, Armstrong N, Lee YC, *et al.* Chronic diseases in the European Union: the prevalence and health cost implications of chronic pain. *J Pain Palliat Care Pharmacother* 2012; 26: 310–325.
30 Stallard E. Estimates of the incidence, prevalence, duration, intensity, and cost of chronic disability among the U.S. elderly *N Am Actuar J* 2011; 15: 32–58.
31 Harrington RA, Liu ET. Quantitative biology and clinical trials: a perspective. *In:* Liu ET, Lauffenburger DA, eds. Systems Biomedicine: Concepts and Perspectives. San Diego, Academic Press, 2010; pp. 415–424.
32 Ferguson BS, Hoggarth DA, Maliniak D, *et al.* Real-time, aptamer-based tracking of circulating therapeutic agents in living animals. *Sci Transl Med* 2013; 5: 213ra165.
33 Damani S, Bacconi A, Libiger O, *et al.* Characterization of circulating endothelial cells in acute myocardial infarction. *Sci Transl Med* 2012; 4: 126ra133.
34 Hwang D, Lee IY, Yoo H, *et al.* A systems approach to prion disease. *Mol Syst Biol* 2009; 5: 252.
35 Wang K, Zhang S, Marzolf B, *et al.* Circulating microRNAs, potential biomarkers for drug-induced liver injury. *Proc Natl Acad Sci USA* 2009; 106: 4402–4407.
36 Iinuma H, Watanabe T, Mimori K, *et al.* Clinical significance of circulating tumor cells, including cancer stem-like cells, in peripheral blood for recurrence and prognosis in patients with Dukes' stage B and C colorectal cancer. *J Clin Oncol* 2011; 29: 1547–1555.
37 Sinnott M, Kinsley BT, Jackson AD, *et al.* Fasting plasma glucose as initial screening for diabetes and prediabetes in Irish adults: the Diabetes Mellitus and Vascular health initiative (DMVhi). *PLoS One* 2015; 10: e0122704.
38 Mosca T, Menezes MC, Dionigi PC, *et al.* C3 and C4 complement system components as biomarkers in the intermittent atopic asthma diagnosis. *J Pediatr* 2011; 87: 512–516.
39 Lawal O, Ahmed WM, Nijsen TME, *et al.* Exhaled breath analysis: a review of 'breath-taking' methods for off-line analysis. *Metabolomics* 2017; 13: 110.
40 Gonzalez-Guerrero AB, Maldonado J, Dante S, *et al.* Direct and label-free detection of the human growth hormone in urine by an ultrasensitive bimodal waveguide biosensor. *J Biophotonics* 2017; 10: 61–67.
41 Atukorale PU, Covarrubias G, Bauer L, *et al.* Vascular targeting of nanoparticles for molecular imaging of diseased endothelium. *Adv Drug Deliv Rev* 2017; 113: 141–156.
42 Chen R, Mias GI, Li-Pook-Than J, *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012; 148: 1293–1307.
43 Stanberry L, Mias GI, Haynes W, *et al.* Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. *Metabolites* 2013; 3: 741–760.
44 Sperisen P, Cominetti O, Martin FP. Longitudinal omics modeling and integration in clinical metabonomics research: challenges in childhood metabolic health research. *Front Mol Biosci* 2015; 2: 44.

84 Chari R, Coe BP, Vucic EA, *et al.* An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Syst Biol* 2010; 4: 67.
85 Ovaska K, Laakso M, Haapa-Paananen S, *et al.* Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med* 2010; 2: 65.
86 Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet* 2017; 8: 84.
87 Tan H-TT, Sugita K, Akdis CA. Novel biologicals for the treatment of allergic diseases and asthma. *Curr Allergy Asthma Rep* 2016; 16: 70.
88 Agusti A, Soriano JB. COPD as a systemic disease. *COPD* 2008; 5: 133–138.
89 Hanania NA. The impact of inhaled corticosteroid and long-acting beta-agonist combination therapy on outcomes in COPD. *Pulm Pharmacol Ther* 2008; 21: 540–550.
90 Morrow JD, Qiu W, Chhabra D, *et al.* Identifying a gene expression signature of frequent COPD exacerbations in peripheral blood using network methods. *BMC Med Genomics* 2015; 8: 1.
91 Xue J, Schmidt SV, Sander J, *et al.* Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity* 2014; 40: 274–288.
92 Glass K, Huttenhower C, Quackenbush J, *et al.* Passing messages between biological networks to refine predicted interactions. *PLoS One* 2013; 8: e64832.
93 Lao T, Glass K, Qiu W, *et al.* Haploinsufficiency of Hedgehog interacting protein causes increased emphysema induced by cigarette smoke through network rewiring. *Genome Med* 2015; 7: 12.
94 Faner R, Cruz T, Casserras T, *et al.* Network analysis of lung transcriptomics reveals a distinct B-cell signature in emphysema. *Am J Respir Crit Care Med* 2016; 193: 1242–1253.
95 Cho MH, Washko GR, Hoffmann TJ, *et al.* Cluster analysis in severe emphysema subjects using phenotype and genotype data: an exploratory investigation. *Respir Res* 2010; 11: 30.
96 Castaldi PJ, Dy J, Ross J, *et al.* Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* 2014; 69: 415–422.
97 Ross JC, Castaldi PJ, Cho MH, *et al.* A Bayesian nonparametric model for disease subtyping: application to emphysema phenotypes. *IEEE Trans Med Imaging* 2017; 36: 343–354.
98 Loza MJ, Adcock I, Auffray C, *et al.* Longitudinally stable, clinically defined clusters of patients with asthma independently identified in the ADEPT and U-BIOPRED asthma studies. *Ann Am Thorac Soc* 2016; 13: Suppl. 1, S102–S103.
99 Kuo CS, Pavlidis S, Loza M, *et al.* T-helper cell type 2 (Th2) and non-Th2 molecular phenotypes of asthma using sputum transcriptomics in U-BIOPRED. *Eur Respir J* 2017; 49: 1602135.
100 Sharma A, Menche J, Huang CC, *et al.* A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet* 2015; 24: 3005–3020.
101 Qiu W, Guo F, Glass K, *et al.* Differential connectivity of gene regulatory networks distinguishes corticosteroid response in asthma. *J Allergy Clin Immunol* 2017; in press [http://doi.org/10.1016/j.jaci.2017.05.052].
102 Young RP, Hopkins RJ, Christmas T, *et al.* COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. *Eur Respir J* 2009; 34: 380–386.
103 de Torres JP, Bastarrika G, Wisnivesky JP, *et al.* Assessing the relationship between lung cancer risk and emphysema detected on low-dose CT of the chest. *Chest* 2007; 132: 1932–1938.
104 Molina R, Marrades RM, Auge JM, *et al.* Assessment of a combined panel of six serum tumor markers for lung cancer. *Am J Respir Crit Care Med* 2016; 193: 427–437.
105 Holdenrieder S, Pagliaro L, Morgenstern D, *et al.* Clinically meaningful use of blood tumor markers in oncology. *Biomed Res Int* 2016; 2016: 9795269.
106 Timofeeva MN, Hung RJ, Rafnar T, *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14900 cases and 29485 controls. *Hum Mol Genet* 2012; 21: 4980–4995.
107 Raji OY, Agbaje OF, Duffy SW, *et al.* Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: the Liverpool Lung Project. *Cancer Prev Res* 2010; 3: 664–669.
108 Spitz MR, Etzel CJ, Dong Q, *et al.* An expanded risk prediction model for lung cancer. *Cancer Prev Res* 2008; 1: 250–254.
109 Young RP, Hopkins RJ, Hay BA, *et al.* Lung cancer susceptibility model based on age, family history and genetic variants. *PLoS One* 2009; 4: e5302.
110 Tang R-X, Chen W-J, He R-Q, *et al.* Identification of a RNA-Seq based prognostic signature with five lncRNAs for lung squamous cell carcinoma. *Oncotarget* 2017; 8: 50761–50773.
111 Tian F, Zhao J, Fan X, *et al.* Weighted gene co-expression network analysis in identification of metastasis-related genes of lung squamous cell carcinoma based on the Cancer Genome Atlas database. *J Thorac Dis* 2017; 9: 42–53.
112 Merelli I, Perez-Sanchez H, Gesing S, *et al.* Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. *Biomed Res Int* 2014; 2014: 134023.
113 Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics* 2015; 8: 33.
114 Gligorijevic V, Malod-Dognin N, Przulj N. Integrative methods for analyzing big data in precision medicine. *Proteomics* 2016; 16: 741–758.
115 Riekeberg E, Powers R. New frontiers in metabolomics: from measurement to insight. *F1000Res* 2017; 6: 1148.
116 Ozarda Y. Reference intervals: current status, recent developments and future considerations. *Biochem Med* 2016; 26: 5–16.
117 Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011; 12: 56–68.
118 Haanstra JR, Bakker BM. Drug target identification through systems biology. *Drug Discov Today Technol* 2015; 15: 17–22.
119 Chen S, Jiang H, Cao Y, *et al.* Drug target identification using network analysis: taking active components in *Sini* decoction as an example. *Sci Rep* 2016; 6: 24245.
120 Cesario A, Auffray C, Russo P, *et al.* P4 medicine needs P4 education. *Curr Pharm Des* 2014; 20: 6071–6072.

Informe de publicaciones de la Tesis

La tesis del doctorando se centra en dos artículos originales publicados el año 2017. En ambas el doctorando es el primer co-autor, ambas son del primer décil del ámbito respiratorio, y en ambas el doctorando ha realizado el análisis de datos, interpretación, extracción de conclusiones y readccion de los mismos. Específicamente:

1. **Multi-level differential network analysis of COPD exacerbations.**
   Autores: ***Noell G****, Cosío BG*, Faner R*, Monsó E, Peces-Barba G, de Diego A, Esteban C, Gea J, Rodriguez-Roisin R, Garcia-Nuñez M, Pozo-Rodriguez F, Kalko SG, Agustí A.
   Eur Respir J. 2017 Sep 27;50(3). PMID: 28954781. (*: authors contributed equally)

   - La revista de este articulo (*European Respiratory Journal*) tiene un **impact factor** en 2017 de **12.2 puntos.**
   - El doctorando es **primer co-autor** de este artículo*. Los otros dos coautores ya son doctores y han participado en el diseño del estudio y recolección de los datos. Mientras que el doctorando, específicamente ha trabajado con los datos crudos, realizado todos los análisis estadísticos del articulo y ha participado en la interpretación de los mismos, extracción de conclusiones, elaboración y escritura del articulo y de sus revisiones.

2. **Lung function in early adulthood and health in later life: a transgenerational cohort analysis.**
   Agustí A*, Noell G*, Brugada J, Faner R.
   (*: authors contributed equally). Lancet Respir Med. 2017 Dec;5(12):935-945. doi: 10.1016/S2213-2600(17)30434-4. Epub 2017 Nov 14. PMID: 29150410

   - La revista de este articulo (*Lancet Respiratory Medicine*) tiene un **impact factor** actual de **21.5, siendo la primera de su categoría.**
   - El doctorando es **co-pimer autor** de este artículo con Àlvar Agustí, que es codirector del doctorando y ha contribuido en el diseño del estudio e interpretación de resultandos.
   - En este caso, el doctorando, ha tenido acceso a los datos crudos, ha realizado todos los análisis estadísticos del artículo y ha participado en la interpretación de resultdos, extracción de conclusiones, elaboración y escritura del artículo y de sus revisiones.

Asimismo, declaro formalmente que ***ninguno de les coautores de estos dos artículos ha utilizado, implícitamente o explícitamente, estos trabajos para la realización de una tesis doctoral***. Además el doctorando durante su doctorado ha colaborado con la realización de 5 artículos más (1-5), uno de los cuales es una revisión, dos se han utilizado en otras tesis doctorales, pero 2 son originales y no se han utilizado en ninguna tesis. Todo ello demuestra la implicación del doctorando con el grupo y su trabajo de tesis durante este periodo.

Barcelona, 20 de Diciembre de 2018

Firma del co-director de Tesis	Firma del co-director de Tesis	Firma del doctorando
Dra. Maria Rosa Faner Canet	Dr. Àlvar Agustí	Guillaume Noell

Otros artículos que no conforman el trabajo de la tesis, pero en los que ha participado el doctorando:

1: Faner R*, Noell G*, Badia JR, López-Giraldo A, Bakke P, Silverman EK, Tal-Singer R, Agustí A. Distribution, temporal stability and association with all-cause mortality of the 2017 GOLD groups in the ECLIPSE cohort. Respir Med. 2018 Aug;141:14-19. PubMed PMID:30053959. *** co-primary authors. No usado en ninguna tesis.**

2: Toledo-Pons N, Noell G, Jahn A, Iglesias A, Duran MA, Iglesias J, Rios A, Scrimini S, Faner R, Gigirey O, Agustí A, Cosío BG. Bone marrow characterization in COPD: a multi-level network analysis. Respir Res. 2018 Jun 15;19(1):118. PubMed PMID: 29903047. **Segundo autor, articulo que forma parte de otra tesis doctoral.**

3: Noell G, Faner R, Agustí A. From systems biology to P4 medicine: applications in respiratory medicine. Eur Respir Rev. 2018 Feb 7;27(147). pii: 170110. doi: 10.1183/16000617.0110-2017. Print 2018 Mar 31. Review. PubMed PMID: 29436404. **Primer autor, artículo de revision que no forma parte de otra tesis doctoral.**

4: Agustí A, Compte A, Faner R, Garcia-Aymerich J, Noell G, Cosio BG, Rodriguez-Roisin R, Celli B, Anto JM. The EASI model: A first integrative computational approximation to the natural history of COPD. PLoS One. 2017 Oct 10;12(10):e0185502. doi: 10.1371/journal.pone.0185502. eCollection 2017. PubMed PMID: 29016620; PubMed Central PMCID: PMC5634586. **Quinto autor, artículo original que no forma parte de otra tesis doctoral.**

5: Faner R, Cruz T, Casserras T, López-Giraldo A, Noell G, Coca I, Tal-Singer R, Miller B, Rodriguez-Roisin R, Spira A, Kalko SG, Agustí A. Network Analysis of Lung Transcriptomics Reveals a Distinct B-Cell Signature in Emphysema. Am J Respir Crit Care Med. 2016 Jun 1;193(11):1242-53. PMID: 26735770. **Quinto autor, articulo que forma parte de otra tesis doctoral.**

Bibliography

1. Andersen, K. and V. Gudnason, *[Chronic non-communicable diseases: a global epidemic of the 21st century]*. Laeknabladid, 2012. **98**(11): p. 591-5.
2. *WHO, The global burden of disease: 2004 update*. Available from: http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/.
3. Unwin, N. and K.G. Alberti, *Chronic non-communicable diseases.* Ann Trop Med Parasitol, 2006. **100**(5-6): p. 455-64.
4. Ajay, V.S., D.A. Watkins, and D. Prabhakaran, *Relationships among Major Risk Factors and the Burden of Cardiovascular Diseases, Diabetes, and Chronic Lung Disease*, in *Cardiovascular, Respiratory, and Related Disorders*, rd, et al., Editors. 2017: Washington (DC).
5. Xu, X., G.D. Mishra, and M. Jones, *Mapping the global research landscape and knowledge gaps on multimorbidity: a bibliometric study.* J Glob Health, 2017. **7**(1): p. 010414.
6. Goh, K.-I., et al., *The human disease network.* 2007. **104**(21): p. 8685-8690.
7. ; Available from: https://ourworldindata.org/life-expectancy.
8. Mathers, C.D., et al., *Causes of international increases in older age life expectancy.* Lancet, 2015. **385**(9967): p. 540-8.
9. Bynum, W.F., et al., *The Western Medical Tradition: 1800-2000*. 2006: Cambridge University Press.
10. Hunter, L. and K.B. Cohen, *Biomedical language processing: what's beyond PubMed?* Mol Cell, 2006. **21**(5): p. 589-94.
11. Chakma, J., et al., *Asia's ascent--global trends in biomedical R&D expenditures.* N Engl J Med, 2014. **370**(1): p. 3-6.
12. Noell, G., R. Faner, and A.J.E.R.R. Agustí, *From systems biology to P4 medicine: applications in respiratory medicine.* 2018. **27**(147): p. 170110.
13. Mardis, E.R., *The $1,000 genome, the $100,000 analysis?* Genome Med, 2010. **2**(11): p. 84.
14. Phillips, K.A., M.J. Pletcher, and U. Ladabaum, *Is the ``$1000 Genome'' really $1000? Understanding the full benefits and costs of genomic sequencing.* Technol Health Care, 2015. **23**(3): p. 373-9.
15. Noble, W.S., *How does multiple testing correction work?* Nat Biotechnol, 2009. **27**(12): p. 1135-7.
16. Macklem, P.T.J.J.o.A.P., *Emergent phenomena and the secrets of life.* 2008. **104**(6): p. 1844-1846.
17. Agustí, A., J.J.A.j.o.r. Vestbo, and c.c. medicine, *Current controversies and future perspectives in chronic obstructive pulmonary disease.* 2011. **184**(5): p. 507-513.
18. Russell, C.D. and J.K.J.C.O.i.S.B. Baillie, *Treatable traits and therapeutic targets: goals for systems biology in infectious disease.* 2017. **2**: p. 140-146.
19. *Global Initiative for Obstructive Lung Disease (GOLD); global strategy for the diagnosis, management and prevention of COPD; 2019 report*. https://goldcopd.org/wp-content/uploads/2018/11/GOLD-2019-v1.7-FINAL-14Nov2018-WMS.pdf.
20. Ntritsos, G., et al., *Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis.* Int J Chron Obstruct Pulmon Dis, 2018. **13**: p. 1507-1514.
21. Sørheim, I.-C., et al., *Gender differences in COPD: are women more susceptible to smoking effects than men?* 2010. **65**(6): p. 480-485.

22. Mannino, D.M., et al., *Chronic obstructive pulmonary disease surveillance-United States, 1971-2000.* 2002. **47**(10): p. 1184-1199.
23. Ford, E.S., et al., *Total and state-specific medical and absenteeism costs of COPD among adults aged >/= 18 years in the United States for 2010 and projections through 2020.* Chest, 2015. **147**(1): p. 31-45.
24. Faner, R., et al., *Distribution, temporal stability and association with all-cause mortality of the 2017 GOLD groups in the ECLIPSE cohort.* 2018.
25. Hangaard, S., et al., *Causes of misdiagnosis of chronic obstructive pulmonary disease: A systematic scoping review.* 2017. **129**: p. 63-84.
26. Çolak, Y., et al., *Prognosis of asymptomatic and symptomatic, undiagnosed COPD in the general population in Denmark: a prospective cohort study.* 2017. **5**(5): p. 426-434.
27. Fletcher, C. and R. Peto, *The natural history of chronic airflow obstruction.* Br Med J, 1977. **1**(6077): p. 1645-8.
28. Lange, P., B. Celli, and A. Agusti, *Lung-Function Trajectories and Chronic Obstructive Pulmonary Disease.* N Engl J Med, 2015. **373**(16): p. 1575.
29. Salvi, S.S. and P.J. Barnes, *Chronic obstructive pulmonary disease in non-smokers.* Lancet, 2009. **374**(9691): p. 733-43.
30. Lange, P., et al., *Lung-function trajectories leading to chronic obstructive pulmonary disease.* 2015. **373**(2): p. 111-122.
31. Hernández, M., et al., *Impact of using the new GOLD classification on the distribution of COPD severity in clinical practice.* 2018. **13**: p. 351.
32. Agusti, A., et al., *Characterisation of COPD heterogeneity in the ECLIPSE cohort.* 2010. **11**(1): p. 122.
33. Lundbäck, B., et al., *Not 15 but 50% of smokers develop COPD?—report from the obstructive lung disease in Northern Sweden studies.* 2003. **97**(2): p. 115-122.
34. Bridevaux, P.-O., et al., *Prevalence of airflow obstruction in smokers and never smokers in Switzerland.* 2010.
35. Marsh, S., et al., *Smoking and COPD: what really are the risks?* 2006. **28**(4): p. 883-884.
36. Afonso, A.S., et al., *COPD in the general population: prevalence, incidence and survival.* 2011. **105**(12): p. 1872-1884.
37. Trupin, L., et al., *The occupational burden of chronic obstructive pulmonary disease.* 2003. **22**(3): p. 462-469.
38. Lopez, A.D., et al., *Global burden of disease and risk factors*. 2006: The World Bank.
39. Li, J., et al., *Major air pollutants and risk of COPD exacerbations: a systematic review and meta-analysis.* 2016. **11**: p. 3079.
40. Sunyer, J.J.E.R.J., *Urban air pollution and chronic obstructive pulmonary disease: a review.* 2001. **17**(5): p. 1024-1033.
41. Allwood, B.W., L. Myer, and E.D.J.R. Bateman, *A systematic review of the association between pulmonary tuberculosis and the development of chronic airflow obstruction in adults.* 2013. **86**(1): p. 76-85.
42. Raynaud, C., N. Roche, and C.J.R.r. Chouaid, *Interactions between HIV infection and chronic obstructive pulmonary disease: Clinical and epidemiological aspects.* 2011. **12**(1): p. 117.
43. Andréjak, C., et al., *Chronic respiratory disease, inhaled corticosteroids and risk of non-tuberculous mycobacteriosis.* 2013. **68**(3): p. 256-262.
44. Kim, J.-H., et al., *Inhaled corticosteroid is associated with an increased risk of TB in patients with COPD.* 2013. **143**(4): p. 1018-1024.

45. van Zyl Smit, R., et al., *Global lung health: the colliding epidemics of tuberculosis, tobacco smoking, HIV and COPD.* 2010. **35**(1): p. 27-33.
46. Terzikhan, N., et al., *Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam Study.* 2016. **31**(8): p. 785-792.
47. Li, L.S.K., et al., *"what are my chances of developing COPD if one of my parents has the disease?" A systematic review and meta-analysis of prevalence of co-occurrence of COPD diagnosis in parents and offspring.* 2017. **12**: p. 403.
48. Rahaghi, F.F., et al., *The prevalence of alpha-1 antitrypsin deficiency among patients found to have airflow obstruction.* 2012. **9**(4): p. 352-358.
49. Pillai, S.G., et al., *Loci identified by genome-wide association studies influence different disease-related phenotypes in chronic obstructive pulmonary disease.* 2010. **182**(12): p. 1498-1505.
50. Wain, L.V., et al., *Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank.* 2015. **3**(10): p. 769-781.
51. Li, Y., M.H. Cho, and X.J.C.S. Zhou, *What do polymorphisms tell us about the mechanisms of COPD?* 2017. **131**(24): p. 2847-2863.
52. Qiao, D., et al., *Whole exome sequencing analysis in severe chronic obstructive pulmonary disease.* 2018. **27**(21): p. 3801-3812.
53. Kong, X., et al., *Genome-wide association study identifies BICD1 as a susceptibility gene for emphysema.* Am J Respir Crit Care Med, 2011. **183**(1): p. 43-9.
54. Alder, J.K., et al., *Telomere length is a determinant of emphysema susceptibility.* 2011. **184**(8): p. 904-912.
55. Rizvi, S., S.T. Raza, and F. Mahdi, *Telomere length variations in aging and age-related diseases.* Curr Aging Sci, 2014. **7**(3): p. 161-7.
56. Stanley, S.E., et al., *Telomerase mutations in smokers with severe emphysema.* 2015. **125**(2): p. 563-570.
57. Sethi, S.J.A.o.t.A.T.S., *Chronic obstructive pulmonary disease and infection. Disruption of the microbiome?* 2014. **11**(Supplement 1): p. S43-S47.
58. Shaheen, S.O., et al., *The relationship of dietary patterns with adult lung function and COPD.* 2010. **36**(2): p. 277-284.
59. Sorli-Aguilar, M., et al., *Dietary patterns are associated with lung function among Spanish smokers without respiratory disease.* 2016. **16**(1): p. 162.
60. Kaluza, J., et al., *Fruit and vegetable consumption and risk of COPD: a prospective cohort study of men.* 2017: p. thoraxjnl-2015-207851.
61. Faner, R. and Á.J.A.o.t.A.T.S. Agustí, *Multilevel, dynamic chronic obstructive pulmonary disease heterogeneity. A challenge for personalized medicine.* 2016. **13**(Supplement 5): p. S466-S470.
62. Agusti, A., et al., *Characterisation of COPD heterogeneity in the ECLIPSE cohort.* Respiratory research, 2010. **11**: p. 122.
63. Agusti, A. and J. Vestbo, *Current controversies and future perspectives in chronic obstructive pulmonary disease.* American journal of respiratory and critical care medicine, 2011. **184**(5): p. 507-13.
64. Agusti, A. and W. Macnee, *The COPD control panel: towards personalised medicine in COPD.* Thorax, 2013. **68**(7): p. 687-90.
65. Galban, C.J., et al., *Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression.* Nat. Med, 2012. **18**: p. 1711-1715.
66. Bafadhel, M., et al., *The Role of CT Scanning in Multidimensional Phenotyping of COPD.* Chest, 2011. **140**(3): p. 634-642.

67. MacNee, W., *Pathogenesis of chronic obstructive pulmonary disease.* Proc Am Thorac Soc, 2005. **2**(4): p. 258-66; discussion 290-1.
68. Hoogendoorn, M., et al., *Association between lung function and exacerbation frequency in patients with COPD.* 2010. **5**: p. 435.
69. Laratta, C.R. and S.J.B.r.i. Van Eeden, *Acute exacerbation of chronic obstructive pulmonary disease: cardiovascular links.* 2014. **2014**.
70. Hallin, R., et al., *Nutritional status, dietary energy intake and the risk of exacerbations in patients with chronic obstructive pulmonary disease (COPD).* 2006. **100**(3): p. 561-567.
71. Spruit, M., et al., *Muscle force during an acute exacerbation in hospitalised patients with COPD and its relationship with CXCL8 and IGF-I.* 2003. **58**(9): p. 752-756.
72. Chen, Y.-W.R., J.M. Leung, and D.D.J.P.O. Sin, *A systematic review of diagnostic biomarkers of COPD exacerbation.* 2016. **11**(7): p. e0158843.
73. Freeman, C.M., et al., *Acute exacerbations of chronic obstructive pulmonary disease are associated with decreased CD4+ & CD8+ T cells and increased growth & differentiation factor-15 (GDF-15) in peripheral blood.* 2015. **16**(1): p. 94.
74. Qiu, Y., et al., *Biopsy neutrophilia, neutrophil chemokine and receptor gene expression in severe exacerbations of chronic obstructive pulmonary disease.* 2003. **168**(8): p. 968-975.
75. Bathoorn, E., et al., *Airways inflammation and treatment during acute exacerbations of COPD.* 2008. **3**(2): p. 217.
76. Saha, S. and C.E.J.I.j.o.c.o.p.d. Brightling, *Eosinophilic airway inflammation in COPD.* 2006. **1**(1): p. 39.
77. Kerkhof, M., et al., *Blood eosinophil count and exacerbation risk in patients with COPD.* 2017. **50**(1): p. 1700761.
78. Bafadhel, M., et al., *Blood eosinophil guided prednisolone therapy for exacerbations of COPD: a further analysis.* 2014. **44**(3): p. 789-791.
79. Wedzicha, J.A. and T.A.J.T.L. Seemungal, *COPD exacerbations: defining their cause and prevention.* 2007. **370**(9589): p. 786-796.
80. White, A., S. Gompertz, and R.J.T. Stockley, *Chronic obstructive pulmonary disease• 6: The aetiology of exacerbations of chronic obstructive pulmonary disease.* 2003. **58**(1): p. 73-80.
81. Wedzicha, J.A.J.P.o.t.A.T.S., *Role of viruses in exacerbations of chronic obstructive pulmonary disease.* 2004. **1**(2): p. 115-120.
82. Mallia, P., et al., *Rhinovirus infection induces degradation of antimicrobial peptides and secondary bacterial infection in chronic obstructive pulmonary disease.* 2012. **186**(11): p. 1117-1124.
83. Wedzicha, J.A., et al., *Mechanisms and impact of the frequent exacerbator phenotype in chronic obstructive pulmonary disease.* 2013. **11**(1): p. 181.
84. Arostegui, I., et al., *Subtypes of patients experiencing exacerbations of COPD and associations with outcomes.* 2014. **9**(6): p. e98580.
85. Vermeeren, M., et al., *Prevalence of nutritional depletion in a large out-patient population of patients with COPD.* 2006. **100**(8): p. 1349-1355.
86. Rawal, G. and S.J.J.o.t.i.m. Yadav, *Nutrition in chronic obstructive pulmonary disease: A review.* 2015. **3**(4): p. 151-154.
87. Lakhdar, R. and R.A.J.J.o.t.d. Rabinovich, *Can muscle protein metabolism be specifically targeted by nutritional support and exercise training in chronic obstructive pulmonary disease?* 2018. **10**(Suppl 12): p. S1377.
88. Spruit, M.A., et al., *COPD and exercise: does it make a difference?* 2016. **12**(2): p. e38.

89. Putcha, N., et al. *Comorbidities and chronic obstructive pulmonary disease: prevalence, influence on outcomes, and management*. in *Seminars in respiratory and critical care medicine*. 2015. NIH Public Access.
90. Control, C.f.D. and Prevention, *How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the surgeon general.* 2010.
91. Luu, T.M., et al., *Preterm birth: risk factor for early-onset chronic diseases.* 2016. **188**(10): p. 736-746.
92. Kotecha, S.J., et al., *Effect of preterm birth on later FEV1: a systematic review and meta-analysis.* 2013. **68**(8): p. 760-766.
93. Van Remoortel, H., et al., *Risk factors and comorbidities in the preclinical stages of chronic obstructive pulmonary disease.* 2014. **189**(1): p. 30-38.
94. Wauters, E., et al., *The TERT-CLPTM1L locus for lung cancer predisposes to bronchial obstruction and emphysema.* 2011: p. erj01871-2010.
95. Young, R., et al., *Chromosome 4q31 locus in COPD is also associated with lung cancer.* 2010. **36**(6): p. 1375-1382.
96. Smolonska, J., et al., *Common genes underlying asthma and COPD? Genome-wide analysis on the Dutch hypothesis.* 2014: p. erj00019-2014.
97. Rubio-Perez, C., et al., *Genetic and functional characterization of disease associations explains comorbidity.* Sci Rep, 2017. **7**(1): p. 6207.
98. Faner, R., et al., *Molecular and clinical diseasome of comorbidities in exacerbated COPD patients.* Eur Respir J, 2015. **46**(4): p. 1001-10.
99. Benjamini, Y. and Y.J.J.o.t.r.s.s.S.B. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* 1995: p. 289-300.
100. Wedzicha, J.A., R. Singh, and A.J. Mackay, *Acute COPD exacerbations.* Clin Chest Med, 2014. **35**(1): p. 157-63.
101. Davies, L., R. Angus, and P.J.T.L. Calverley, *Oral corticosteroids in patients admitted to hospital with exacerbations of chronic obstructive pulmonary disease: a prospective randomised controlled trial.* 1999. **354**(9177): p. 456-460.
102. Bafadhel, M., et al., *Acute exacerbations of chronic obstructive pulmonary disease: identification of biologic clusters and their biomarkers.* 2011. **184**(6): p. 662-671.
103. Bafadhel, M., et al., *Blood eosinophils to direct corticosteroid treatment of exacerbations of chronic obstructive pulmonary disease: a randomized placebo-controlled trial.* 2012. **186**(1): p. 48-55.
104. Perera, W.R., et al., *Inflammatory changes, recovery and recurrence at COPD exacerbation.* 2007. **29**(3): p. 527-534.
105. Miravitlles, M., et al., *Is it possible to identify exacerbations of mild to moderate COPD that do not require antibiotic treatment?* 2013. **144**(5): p. 1571-1577.
106. Hurst, J.R., et al., *Use of plasma biomarkers at exacerbation of chronic obstructive pulmonary disease.* 2006. **174**(8): p. 867-874.
107. Dev, D., et al., *Value of C-reactive protein measurements in exacerbations of chronic obstructive pulmonary disease.* 1998. **92**(4): p. 664-667.
108. Alotaibi, N.M., et al., *Phenotyping COPD exacerbations using imaging and blood-based biomarkers.* 2018. **13**: p. 217.
109. Karadeniz, G., et al., *C-reactive protein measurements as a marker of the severity of chronic obstructive pulmonary disease exacerbations.* 2013. **36**(4): p. 948-953.
110. Montes de Oca, M. and M.J.M.S. Laucho-Contreras, *Is It Time to Change the Definition of Acute Exacerbation of Chronic Obstructive Pulmonary Disease? What Do We Need to Add?* 2018. **6**(2): p. 50.

111. Macklem, P.T., A.J.P.i.B. Seely, and Medicine, *Towards a definition of life.* 2010. **53**(3): p. 330-340.
112. Hahn, M.W., A.D.J.M.b. Kern, and evolution, *Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks.* 2004. **22**(4): p. 803-806.
113. Yu, H., et al., *The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.* 2007. **3**(4): p. e59.
114. Faner, R., et al., *Network analysis of lung transcriptomics reveals a distinct B-cell signature in emphysema.* 2016. **193**(11): p. 1242-1253.
115. Divo, M.J., et al., *Chronic obstructive pulmonary disease comorbidities network.* 2015: p. ERJ-01716-2014.
116. Bush, A.J.A.o.t.A.T.S., *Lung development and aging.* 2016. **13**(Supplement 5): p. S438-S446.
117. Barker, D.J., et al., *Relation of birth weight and childhood respiratory infection to adult lung function and death from chronic obstructive airways disease.* 1991. **303**(6804): p. 671-675.
118. Stern, D.A., et al., *Poor airway function in early infancy and lung function by age 22 years: a non-selective longitudinal cohort study.* 2007. **370**(9589): p. 758-764.
119. Berry, C.E., et al., *A distinct low lung function trajectory from childhood to the fourth decade of life.* 2016. **194**(5): p. 607-612.
120. Owens, L., et al., *Airway function in infancy is linked to airflow measurements and respiratory symptoms from childhood into adulthood.* 2018.
121. Svanes, C., et al., *Early life origins of chronic obstructive pulmonary disease.* 2010. **65**(1): p. 14-20.
122. McGeachie, M.J., et al., *Patterns of growth and decline in lung function in persistent childhood asthma.* 2016. **374**(19): p. 1842-1852.
123. Bui, D.S., et al., *Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life.* 2018.
124. Bui, D.S., et al., *Childhood lung function predicts adult chronic obstructive pulmonary disease and asthma–chronic obstructive pulmonary disease overlap syndrome.* 2017. **196**(1): p. 39-46.
125. Henderson, A.J., *The child is father of the man: the importance of early life influences on lung development*. 2014, BMJ Publishing Group Ltd.
126. Bolton, C.E., et al., *Lung consequences in adults born prematurely.* Thorax, 2015. **70**(6): p. 574-80.
127. Barker, D.J.J.C.s., *In utero programming of chronic disease.* 1998. **95**(2): p. 115-128.
128. Guo, Y.I., et al., *A predictive model for the development of chronic obstructive pulmonary disease.* Biomed Rep, 2015. **3**(6): p. 853-863.
129. Balte, P., et al., *Relationship between birth weight, maternal smoking during pregnancy and childhood and adolescent lung function: A path analysis.* Respir Med, 2016. **121**: p. 13-20.
130. Kellesarian, S.V., et al., *Association between prenatal maternal cigarette smoking and early childhood caries. A systematic review.* J Clin Exp Dent, 2017. **9**(9): p. e1141-e1146.
131. Hanrahan, J.P., et al., *The effect of maternal smoking during pregnancy on early infant lung function.* Am Rev Respir Dis, 1992. **145**(5): p. 1129-35.
132. Martinez, F.D.J.N.E.J.o.M., *Early-life origins of chronic obstructive pulmonary disease.* 2016. **375**(9): p. 871-878.
133. Thacher, J.D., et al., *Tobacco smoke exposure in early life and adolescence in relation to lung function.* Eur Respir J, 2018. **51**(6).

134. Allinson, J.P., et al., *Combined impact of smoking and early-life exposures on adult lung function trajectories.* 2017. **196**(8): p. 1021-1030.
135. Mathew, A.R., et al., *Life-course Smoking Trajectories and Risk of Emphysema in Middle Age: The CARDIA Lung Study.* 2018(ja).
136. Vasquez, M.M., et al., *Low Lung Function in Young Adult Life Is Associated with Early Mortality.* Am J Respir Crit Care Med, 2017. **195**(10): p. 1399-1401.
137. Humphreys, J., et al., *Early-life predictors of future multi-morbidity: results from the Hertfordshire Cohort.* Age Ageing, 2018. **47**(3): p. 474-478.
138. Palatianou, M.E., et al., *Long-term metabolic effects of high birth weight: a critical review of the literature.* Horm Metab Res, 2014. **46**(13): p. 911-20.
139. Yu, Z.B., et al., *Birth weight and subsequent risk of obesity: a systematic review and meta-analysis.* Obes Rev, 2011. **12**(7): p. 525-42.
140. Newton, K.P., et al., *Low and High Birth Weights Are Risk Factors for Nonalcoholic Fatty Liver Disease in Children.* J Pediatr, 2017. **187**: p. 141-146 e1.
141. Boeri, L., et al., *Low Birth Weight Is Associated with a Decreased Overall Adult Health Status and Reproductive Capability - Results of a Cross-Sectional Study in Primary Infertile Patients.* PLoS One, 2016. **11**(11): p. e0166728.
142. de Wijs-Meijler, D.P., et al., *Oxidative injury of the pulmonary circulation in the perinatal period: Short- and long-term consequences for the human cardiopulmonary system.* Pulm Circ, 2017. **7**(1): p. 55-66.
143. Padmanabhan, V., R.C. Cardoso, and M. Puttabyatappa, *Developmental Programming, a Pathway to Disease.* Endocrinology, 2016. **157**(4): p. 1328-40.
144. Hardin, M. and E.K. Silverman, *Chronic Obstructive Pulmonary Disease Genetics: A Review of the Past and a Look Into the Future.* Chronic Obstr Pulm Dis, 2014. **1**(1): p. 33-46.
145. Yuan, C., et al., *Genetic polymorphism and chronic obstructive pulmonary disease.* Int J Chron Obstruct Pulmon Dis, 2017. **12**: p. 1385-1393.
146. Artigas, M.S., et al., *Targeted sequencing of lung function loci in chronic obstructive pulmonary disease cases and controls.* 2017. **12**(1): p. e0170222.
147. Wain, L.V., et al., *Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets.* 2017. **49**(3): p. 416.
148. McGeachie, M.J., et al., *Genetics and genomics of longitudinal lung function patterns in individuals with asthma.* 2016. **194**(12): p. 1465-1474.
149. Miller, S., et al., *The Ser82 RAGE Variant Affects Lung Function and Serum RAGE in Smokers and sRAGE Production In Vitro.* PLoS One, 2016. **11**(10): p. e0164041.
150. Miller, S., et al., *Genes associated with polymorphic variants predicting lung function are differentially expressed during human lung development.* Respir Res, 2016. **17**(1): p. 95.
151. Peng, C., et al., *Epigenome-wide association study of total serum immunoglobulin E in children: a life course approach.* Clin Epigenetics, 2018. **10**: p. 55.
152. Liang, L., et al., *An epigenome-wide association study of total serum immunoglobulin E concentration.* 2015. **520**(7549): p. 670.
153. Ek, W.E., et al., *Epigenome-wide DNA methylation study of IgE concentration in relation to self-reported allergies.* Epigenomics, 2017. **9**(4): p. 407-418.
154. Chen, W., et al., *An epigenome-wide association study of total serum IgE in Hispanic children.* J Allergy Clin Immunol, 2017. **140**(2): p. 571-577.
155. Harju, M., et al., *The burden of childhood asthma and late preterm and early term births.* J Pediatr, 2014. **164**(2): p. 295-9 e1.

156. Cui, Y., et al., *The Exposome: Embracing the Complexity for Discovery in Environmental Health.* Environ Health Perspect, 2016. **124**(8): p. A137-40.
157. Malin, B.A., K.E. Emam, and C.M. O'Keefe, *Biomedical data privacy: problems, perspectives, and recent advances*, in *J Am Med Inform Assoc*. 2013: England. p. 2-6.
158. Obermeyer, Z. and E.J. Emanuel, *Predicting the Future - Big Data, Machine Learning, and Clinical Medicine.* N Engl J Med, 2016. **375**(13): p. 1216-9.
159. Koren, G., et al., *Machine learning of big data in gaining insight into successful treatment of hypertension.* Pharmacol Res Perspect, 2018. **6**(3): p. e00396.