Utilizing Genetic Techniques to Understand Relationships Between Lung Disease and Comorbid Conditions

By

Victoria L. Martucci

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Human Genetics
August 31, 2021
Nashville, Tennessee

Approved:

Melinda C. Aldrich, PhD, MPH

Nancy J. Cox, PhD

Bradley Richmond, MD, PhD

Digna Velez Edwards, PhD

Lea K. Davis, PhD

David C. Samuels, PhD

Timothy S. Blackwell, MD

To my parents and my partner for their constant love and support

Acknowledgements

Pursuing an MD/PhD degree is a large undertaking, and it is not a solitary journey by any means. I could not have completed this dissertation without the constant support of mentors, family, and friends. First and foremost, I would like to thank my PhD advisor, Melinda Aldrich, who has supported me both personally and professionally throughout my years at Vanderbilt. She was always there to celebrate my successes and to encourage me to keep going after the failures. She challenged me to think carefully about my projects, to consider the broader implications of my work, and to identify areas for improvement. Melinda also ensured that my training was well-rounded, encouraging me to pursue professional development opportunities and explore projects outside the scope of my dissertation. I have grown so much as a scientist during my time in the lab, and I could not have done that without Melinda's mentorship.

I have been fortunate to have a wide network of mentors and support throughout my training. My dissertation committee members, Nancy Cox, Lea Davis, Digna Velez Edwards, David Samuels, Bradley Richmond, and Timothy Blackwell, have provided valuable perspectives and helpful advice to strengthen my dissertation. In a similar vein, the members of the Aldrich lab and the TREAT team provided me countless opportunities to present research and receive constructive criticism. I am also grateful for the support of the leadership teams of both the MSTP and the Human Genetics Training Program. I would especially like to thank Roz Johnson for her patience and endless help navigating the administrative aspects of PhD training and Melissa Krasnove for being a truly wonderful listener, supporter, and friend.

The journey to an MD/PhD degree is a long one, full of ups and downs, and I have been fortunate to have wonderful colleagues on the ride with me. I'm grateful for my MSTP classmates and "family", who understand the unique challenges of a dual degree. My fellow

Human Genetics graduate students have also been invaluable, helping me troubleshoot methods and commiserating with me over the frustrations of graduate school life. I would especially like to thank Abin Abraham, Ayesha Muhammad, and Jessica Brown for always being there when I needed friends, and for giving me an outlet to talk endlessly about my cats. I'm also grateful for my friends outside of Vanderbilt, who provided much-needed opportunities to interact with the world outside of school.

Finally, none of this would have been possible without the love and support of my family and my partner. I am who I am today because of my parents, Pat and Mick, and my siblings, Emily, Anna, and Evan. They may not understand the science, but they have always made sure I know how proud they are of me. Last but certainly not least, I want to thank my partner, Alex. A therapist once encouraged me to think of my MD/PhD as a long car trip, and I could not have asked for a better, more supportive co-pilot. From listening to me vent when I needed a sympathetic ear, to providing input as I talked through scientific problems, and to making sure I was taking care of myself, Alex has done it all without complaint. I cannot possibly find the words to express how much I love and appreciate them for all their support.

Of course, I cannot forget to acknowledge my funding support: the MSTP and Human Genetics Training Grants (T32GM007347 and T32GM080178, respectively) and my F30 from the National Heart, Lung, and Blood Institute (F30HL140756). I also received financial support from the Vanderbilt Institute for Clinical and Translational Research (CTSA grant UL1TR002243). This work would not have been possible without the resources of the Advanced Computing Center for Research and Education at Vanderbilt University.

Table of Contents

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Chronic Obstructive Pulmonary Disease

1.1.1 Epidemiology and Risk Factors

Chronic obstructive lung disease (COPD) is a major cause of morbidity and mortality worldwide. The global prevalence of COPD is approximately 11.4%, affecting an estimated 328 million people.[1,2] In the United States, approximately 15 million adults report being diagnosed with COPD, for an overall prevalence of 6.3%.[3] Based on the most recently available estimates from the Centers for Disease Control (CDC), COPD is the fourth leading cause of death in the United States, accounting for 5.6% of all deaths in 2016.[4,5] It is also responsible for approximately 2.9 million deaths worldwide per year, making it the fourth leading cause of death worldwide.[1,6] The mortality rate from COPD is expected to rise worldwide, and it is estimated that COPD will be the third leading cause of death by 2030.[6–10] True estimates of COPD prevalence are likely higher, as many individuals with COPD are unaware of their disease. In studies comparing self-reported diagnoses and clinical data, patient self-report has consistently low sensitivity rates of 26-32%, indicating that the majority of individuals are unaware of their disease status.[11–13] Underdiagnosis of COPD, particularly in asymptomatic patients, is also a major challenge. Pulmonary function tests performed in individuals with other chronic medical conditions have found that 60-90% of patients who meet diagnostic criteria for COPD were never diagnosed.[14–23] The death rates from COPD are also likely underestimated.[24–27] Analyses of death certificates of individuals with confirmed COPD found that the diagnosis is often not included on the certificate, even when COPD is the primary cause of death. Among deaths directly caused by COPD exacerbations, 34% did not list COPD as the primary cause of death, and 21% had no mention of COPD on the death certificate.[25] Given its prevalence, mortality, and underdiagnosis, COPD represents a major public health challenge, and

additional research on COPD is needed to address one of the leading causes of mortality both in the United States and worldwide.

COPD prevalence varies by age, sex, and race/ethnicity. Prevalence of COPD increases with age. Based on self-reported data from the 2019 Behavioral Risk Factor Surveillance System survey, only 2.3% of adults 18-44 have been diagnosed with COPD, compared to 11.9-12.1% of individuals over age 65.[28] This trend also holds when using pulmonary function to diagnose COPD, with between 9.2-15.6% of adults age 40-59 years meeting criteria for COPD versus 13.3-31.2% of adults aged 60-79.[29] Based on self-reported data, women have a slightly higher prevalence than men, with 7.2% of women in the United States reporting a COPD diagnosis compared to 5.6% of men.[28] However, men have a higher prevalence when using lung function measures, with 11.4-24.8% of men and 5.4-17.3% of women in the United States meeting diagnostic criteria for COPD.[29,30] A comprehensive meta-analysis examined gender-specific prevalence and found a slightly higher prevalence in men, with an estimated prevalence of 8.07% in men and 7.30% in women in North America.[31] In terms of race, non-Hispanic White and Black individuals in the United States both report a prevalence of 7.0%. Hispanic individuals have a lower prevalence of 3.5% in self-reported data.[28] Similar trends are found using pulmonary function measures to estimate COPD prevalence. In the United States, 9.5-22.9% of non-Hispanic Whites, 6.9-18.0% of non-Hispanic Blacks, and 2.7-10.4% of Mexican Americans meet pulmonary function definitions for COPD.[29,30]

The major risk factor for developing COPD is cigarette smoking. Approximately 80% of individuals with COPD in the developed world are current or former smokers.[32] Ever-smokers have a 2.89 increased odds of developing COPD compared to non-smokers.[33] These odds increase for individuals with longer smoking durations.[34] Current smokers are at increased risk of developing disease compared to former smokers, especially compared to former smokers who have abstained for ten or more years.[44,45] Exposure to second-hand smoke has also been associated with increased odds of developing COPD.[32,35] However, the extent to which COPD is attributable to cigarette smoking varies greatly based on geographic location. In the developed world, 77-84% of COPD mortality in men and 61-62% in women

can be attributed to smoking, whereas in developing countries, only 45-49% of COPD mortality in men and 12-20% in women is attributed to smoking.[32] While smoking is the leading risk factor for COPD, it is clearly not the only risk factor.

Other environmental exposures are also associated with COPD, including air pollution and airborne particulates.[32] Occupational exposure to vapor dust, gas, or fumes has been associated with airflow limitation, with longer and more severe exposure conferring higher odds of developing the disease.[32,36,37] Worldwide, exposure to smoke from indoor cooking fires is a common risk factor for COPD.[32,38] Burning of biomass fuel as a risk factor for COPD is more prevalent in women, as they typically spend more time indoors.[32] Respiratory infections such as tuberculosis and viral infections in childhood have also been implicated in COPD risk.[35,39,40] In a study of incident cases of COPD in young adults, 8% were estimated to be due to a history of childhood respiratory infections.[40] Asthma has also been identified as a risk factor for developing COPD, and there is overlap between the two disease phenotypes, with some patients having dual diagnoses.[32,40,41] As smoking rates decrease, these additional factors are likely to play an increasingly large role in the risk of developing COPD.

1.1.2 Definition Using Pulmonary Function Tests

COPD is defined by reduced pulmonary function, which is measured using spirometry. To perform a pulmonary function test, an individual is connected to a machine that measures airflow in the lungs. The individual is then asked to take a deep breath and blow out as hard as they can until they can no longer exhale into the connected tube. The spirometer then calculates a number of lung function measurements. This process is typically repeated three times, and the "best" performing measure is recorded. If there is concern about airway obstruction, the test may be repeated after administration of an inhaled bronchodilator to determine if the obstruction is reversible.[42,43] Spirometry is only ordered when there is clinical suspicion of a respiratory issue. Routine screening with spirometry is not currently recommended.[42]

A number of measures are calculated during pulmonary function testing to capture information about an individual's lung capacity and respiratory performance. During the expiration phase of the spirometry test, both the rate of flow over time and the amount of air being exhaled are recorded in a flow-volume loop (Figure 1-1). In healthy lungs, the expiration phase of the flow-volume loop shows an initial rapid increase in flow followed by a linear decrease until no further air can be expired. In individuals with COPD, the initial increase in flow is followed by a non-linear decrease due to reduced elasticity of lung tissue and increased airflow resistance. This decreased elasticity and increased airflow resistance also reduce the amount of volume that can be exhaled, leading to a smaller loop overall (Figure 1-1).[42,44] The forced expiratory volume in one second (FEV_1) captures the amount of air that an individual is able to exhale in the first second after beginning to exhale. The peak expiratory flow (PEF) records the maximum rate of air movement during the expiratory phase. The overall amount of air expired is called the forced vital capacity (FVC). Individuals can never fully empty their lungs, so calculations are performed using the measured variables to estimate an individual's total lung capacity (TLC). The ratio of FEV_1/FVC is calculated as a way to determine the degree of airflow obstruction.[42,43] The clinical definition for COPD is set by the Global Obstructive Lung Disease (GOLD) consortium. Per GOLD guidelines, COPD is defined as a post-bronchodilator FEV_1/FVC ratio < 0.7. The requirement for a post-bronchodilator measurement is to ensure that the airflow obstruction is irreversible, which is a key biologic difference between COPD and asthma.[45]

Historically, COPD was thought to result from rapid decline in lung function in individuals susceptible to smoking-related damage.[46] However, a seminal study by Lange et al. in 2015 found that 52% of individuals with COPD did not experience a rapid decline in FEV_1.[47] Instead, these individuals started from a lower baseline FEV_1 and then experienced a normal rate of decline with aging.[47] This suggested that lung development and early childhood factors play a role in the development of COPD later in life. Childhood conditions such as asthma, lower respiratory infections, and exposure to air pollution have all been associated with decreased maximum lung function in adulthood.[48–50] Early smoking behavior in adolescence may also reduce peak lung function in adulthood.[51] Longitudinal studies

of lung function have found that 4-13% of young adults have an FEV_1 less than 80% of predicted, and that the increased risk of comorbidities and early mortality is further compounded by smoking.[52]

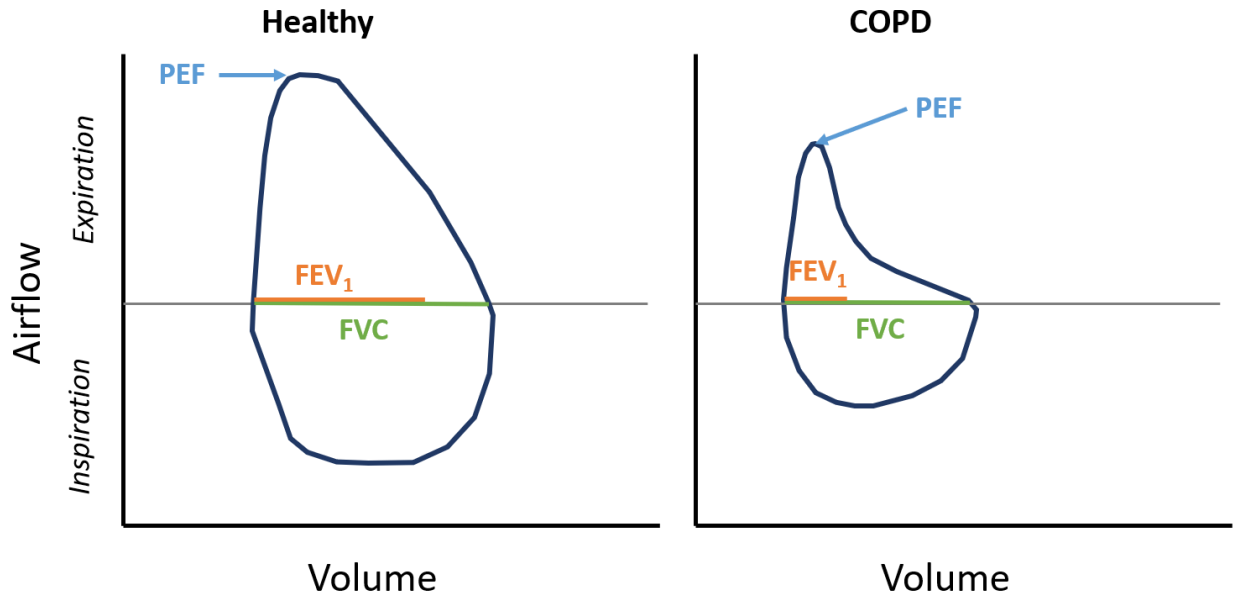

Figure 1-1. Flow-volume loops for healthy lungs (left) and COPD (right). Corresponding pulmonary function measurements are annotated on the flow-volume loops. FEV_1: forced expiratory volume in one second; FVC: forced vital capacity; PEF: peak expiratory flow. Adapted from Niewoehner.[44]

COPD severity is defined by post-bronchodilator FEV_1 percent predicted. Severity is ranked from GOLD stage I to GOLD stage IV, with stage I representing mild disease and stage IV indicating very severe disease. However, FEV_1 is only weakly correlated with levels of lung impairment, so the GOLD guidelines recommend these severity categories only be used for overall patient prognosis rather than to guide clinical treatment. Treatment recommendations are based on symptoms rather than pulmonary function test measurements.[45] Since COPD is defined using pulmonary function tests, lung function measurements are highly related to COPD, and research on the biology and genetics of lung function can provide insight into COPD biology as well.

1.1.3 Genetics of COPD

The first gene to be associated with COPD was *SERPINA1*, which encodes alpha-1-antitrypsin (AAT).[53,54] The AAT protein has an essential role in maintaining the balance between protease and anti-

protease activity in the lung, particularly by inhibiting the activity of neutrophil elastase.[53–55] While many mutations have been identified in the *SERPINA1* gene, only a few have been associated with reduced AAT function. To develop AAT deficiency, individuals must be homozygous for deleterious variants, the most common of which are the Z and S alleles. Both the Z and S alleles are associated with decreased levels of AAT. The Z allele leads to the most severe decline in AAT levels, with a reduction in AAT levels of approximately 85% compared to the most common AAT allele (the M allele). The S allele is associated with a milder 40-50% reduction in AAT levels.[56] This leads to increased protease activity, causing destruction of alveolar walls and development of emphysema. The prevalence of these deleterious variants varies across the world, with the highest frequencies found in individuals of European descent.[53–57] The estimated percentage of individuals in the United States homozygous for the Z allele, which is associated with a more severe clinical phenotype, is 0.036% in Whites but only 0.002% in African Americans. A similar trend is seen with individuals homozygous for the S allele or individuals heterozygous for the Z and S alleles, with SZ frequencies of 0.022% in Whites and 0.006% in African Americans and SS frequencies of 0.052% in whites and 0.022% in African Americans.[54] Individuals with COPD due to AAT deficiency, estimated to be approximately 1% of all individuals with COPD, often present before 45 years of age and with more severe emphysema than individuals without AAT deficiency.[53–55,58]

Even in the absence of AAT deficiency, individuals with a family history of COPD are at increased risk of developing the disease themselves. The odds of developing COPD are approximately 1.7-5.4 times higher in individuals with a first-degree relative with COPD, even when accounting for smoking history.[59–62] The SNP-based genetic heritability of COPD is estimated to be between 30-40%, though twin studies have calculated heritability as high as 60%.[53,63–65] To identify additional genetic risk factors for COPD, several observational cohorts have been developed. The largest, COPDGene, recruited over 10,000 smokers, with approximately one-third of those enrolled identifying as African American.[66] The Subpopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS) and the Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE) studies both

enrolled close to 3,000 individuals, largely of European descent.[67,68] To increase the diversity of research participants, researchers have also leveraged pulmonary function data from other large epidemiologic cohorts developed to study atherosclerosis and heart disease such as the Atherosclerosis Risk in Communities Study (ARIC),[69] the Multi-Ethnic Study of Atherosclerosis (MESA),[70] and Coronary Artery Risk Development in Young Adults (CARDIA).[71]

Multiple genome-wide association studies (GWAS) have been performed to identify potential genetic risk factors for COPD.[53] In the largest GWAS published to date in a largely European population, Sakornsakolpat et al. identified 82 loci associated with COPD, 35 of which were novel.[72] Several loci and genes have been consistently implicated across GWAS, including the *CHRNA3/CHRNA5/IREB2* locus on chromosome 15q25, *HHIP* on chromosome 4q31, and *FAM13A* on chromosome 4q22.[53,72–81] Unsurprisingly, many of the loci associated with COPD have also been identified as significant associations in GWAS of pulmonary function, including the *CHRNA3/CHRNA5/IREB*, *HHIP*, and *FAM13A* loci.[53,82–87] GWAS have also been conducted to look at response to COPD treatment, though these have had small sample sizes and have detected few genome-wide significant signals.[88–92] GWAS have also examined specific phenotypes in COPD, such as airway responsiveness,[93] pulmonary artery enlargement,[94] and percent emphysema on imaging.[95,96] These genetic signals can provide insight into the biology of COPD.

The majority of the GWAS conducted to date have been in European populations, providing limited information on genetic risk factors in individuals of African descent.[97] GWAS conducted in individuals with African ancestry have primarily focused on pulmonary function measures.[86,98] Signals have been detected in the region of *CHRNA3/CHRNA5/IREB* in African ancestry GWAS of lung function, but they did not reach genome-wide significance, likely due to lack of statistical power resulting from smaller sample sizes.[98] Novel signals identified in African ancestry GWAS have also failed to replicate in other cohorts, though this may also be a result of small sample sizes.[86] Cross-ancestry meta-analyses including African ancestry GWAS data identified 47 novel variants, some of which replicated in European populations.[86] Polygenic risk scores (PRS), developed from European descent GWAS, applied

to African ancestry populations have shown weaker associations than observed in European ancestry populations, though the effect sizes have been in the same direction in both populations.[87,99] This suggests that there may be some overlap in the genetic risk factors for COPD development across diverse ancestry populations, but population-specific SNPs or differences in effect sizes may remain. Research studies including larger numbers of African American participants are needed to address these questions.

1.1.4 COPD Comorbidities

Comorbidities are common in COPD, with 86-98% of individuals with COPD reporting at least one comorbid condition.[100–102] These comorbidities include cardiovascular, cerebrovascular, neurological, psychiatric, gastrointestinal, renal, musculoskeletal, or respiratory diseases and disorders.[102–104] Individuals with COPD develop more comorbidities than other adults, often at a younger age.[30,100,105–108] Comorbidity profiles for individuals aged 56-65 with COPD have been found to be similar to those of individuals without COPD who are 15 to 20 years older.[105] Individuals with comorbid conditions report decreased quality of life and increased symptom severity and exacerbation frequency,[109–114] and the presence of multiple comorbidities can increase mortality rates by as much as 400%.[115] COPD has been associated with increased mortality from cardiovascular disease, pneumonia, obstructive sleep apnea, chronic kidney disease, and lung cancer.[115–123] COPD is also an independent risk factor for lung cancer, and patients having concomitant COPD and lung cancer have worse outcomes than those with lung cancer alone.[124–133]

The mechanisms underlying the increased risk of comorbidities for COPD patients are not well understood. Systemic inflammation has been hypothesized as a potential link, though whether the systemic inflammation is the cause of COPD or is caused by COPD remains controversial.[103,104,134–136] In support of this explanation, individuals with COPD and heart disease had significantly increased levels of several inflammatory markers compared to individuals with COPD without heart disease, even after adjusting for age, gender, and smoking history.[115] Furthermore, individuals with COPD and elevated inflammatory markers were at increased risk of heart disease, diabetes, lung cancer, and pneumonia

compared to individuals with COPD whose inflammatory markers were within normal limits.[137] The increased presence of comorbidities in COPD could also be explained by shared environmental risk factors, such as smoking and age.[103,104,134] However, studies of COPD and common comorbid conditions have found that the increased risk of comorbidities with COPD persists even when smoking history is accounted for, suggesting that shared environmental factors alone may not fully explain the relationship.[138] On the other hand, compared to individuals who develop COPD due to biomass smoke exposure, individuals with smoking-related COPD have increased risk of developing ischemic heart disease, suggesting that smoking tobacco may play a role in the development of certain comorbidities.[139] Another explanation for the relationship between COPD and its comorbidities is shared underlying genetic risk factors.[103,140] Recent studies of the relationship between COPD and several cardiovascular comorbidities have identified genetic correlations between these traits.[141] The finding that individuals with poor lung development also have higher rates of other cardiovascular and metabolic disorders suggests that genetic risk factors and gene-environment interactions even in early life may contribute to the development of both COPD and comorbid conditions.[140] Lung cancer and COPD are also known to share genetic risk factors. Loci such as *CHRNA3/CHRNA5, HHIP,* and *FAM13A* have been identified in GWAS for both disorders.[142] Estimates of genetic correlation between COPD and lung cancer are high, even when excluding genomic regions associated with smoking behaviors, suggesting a genetic link beyond smoking.[143] These studies have provided valuable insight into the relationship between COPD, cardiovascular disease, and lung cancer. However, the relationship between COPD and other common comorbidities is still unknown.

1.2 Lung Cancer

1.2.1 Epidemiology

Lung cancer is the leading cause of cancer death worldwide, accounting for approximately 18% of all cancer deaths in 2020. It is also the second most commonly diagnosed cancer, with an estimated 2.2

million new cases of lung cancer diagnosed worldwide in 2020.[144] In the United States, lung cancer represents 12.7% of all new cancer diagnoses and accounted for an estimated 135,720 deaths in 2020.[145] The highest lung cancer incidence rates are in the Southeastern United States, with incidence rates as high as 87.0 per 100,000 individuals in Kentucky compared to the overall incidence rate of 55.2 per 100,000 in the United States.[146] Since smoking is a major risk factor for lung cancer, it is unsurprising that the highest incidence rates are located in states with high smoking rates.[147]

Lung cancer occurs in both men and women, though the incidence rate in men is higher at 62.8 per 100,000 compared to 49.4 per 100,000 for women.[146,148,149] While lung cancer incidence rates continue to decline, the rate of decrease is twice as fast in men than in women.[148,149] This is likely due to differences in smoking uptake over time, as smoking rates in men peaked approximately two decades earlier than in women.[150] Death rates have also declined at different rates in men and women, with men experiencing a 40% reduction in death rate from lung cancer between 1990 and 2016 compared to a 23% reduction in women.[148,149] However, the overall death rate per 100,000 people remains higher in men (44.5) than in women (30.6).[146,149]

While overall lung cancer incidence rates are similar in non-Hispanic White and Black populations, disparities can be observed when stratifying by sex. Black men have the highest incidence rates of lung cancer at 71.7 per 100,000, and mortality rates for Black men are higher than for any other group.[146,148,149] Part of this race-sex disparity may be attributable to differences in socioeconomic status, as lung cancer mortality rates are much higher in poor counties, especially for men.[148] Several studies examining lung cancer survival rates in populations with similar socioeconomic status and access to health care have found no difference in lung cancer survival rates among Blacks and Whites.[151–154]

1.2.2 Histologic Subtypes of Lung Cancer

Lung cancer is broadly classified into either small cell cancer or non-small cell lung cancer (NSCLC) due to differences in pathology, prognosis, and treatment. NSCLC accounts for 80-85% of all lung cancers and can be further classified based on tumor histology.[155–158] The most common histologic

subtype is adenocarcinoma, accounting for approximately 60% of NSCLC cases, followed by squamous cell carcinoma.[149,157] Large cell carcinoma is a relatively rare subtype of NSCLC, accounting for less than 3% of all lung cancer diagnoses.[156,157,159] Distinguishing between NSCLC subtypes can help guide molecular testing and treatment options, as the frequency of genetic mutations with targeted therapeutic options varies between histologic subtypes.[158]

1.2.3 Immunotherapy in Lung Cancer

While the 5-year survival rate for NSCLC remains low, the development of novel targeted therapy options has prolonged survival, particularly for advanced stage disease.[149,160] Immunotherapy has become a promising treatment strategy for metastatic NSCLC. Typically, the immune system detects and eliminates cancerous cells through the detection of cancer antigens and activation of T cells to destroy the abnormal cells. However, some cancer cells develop mutations that allow them to evade detection or promote immunosuppression. The immune system has built-in checkpoint mechanisms, which can be upregulated by tumor cells to prevent an appropriate immune response. The programmed cell-death protein 1 (PD-1) receptor, which is expressed on activated T cells, serves as a modulator of immune response. In response to inflammatory signals induced by immune system activation, tissue cells express programmed cell-death protein ligand 1 (PD-L1), which binds to PD-1 and downregulates T cell activity, preventing excessive tissue damage. Mutations allowing overexpression of PD-L1 in tumor cells can therefore limit the immune response to the tumor. Cytotoxic T-lymphocyte antigen-4 (CTLA-4) is another receptor expressed on T cells that plays an important role in immune response regulation. In addition to recognizing a foreign antigen, T cells also require a costimulatory signal from a second protein, B7, expressed on the antigen-presenting cell in order to become active. CTLA-4 binds to B7, blocking its interaction with the stimulatory CD28 receptor and preventing T cell activation (Figure 1-2).[161–163] Both PD-1/PD-L1 and CTLA-4 play important roles in modulating the immune response to cancer, which has made them attractive candidates for targeted therapy.

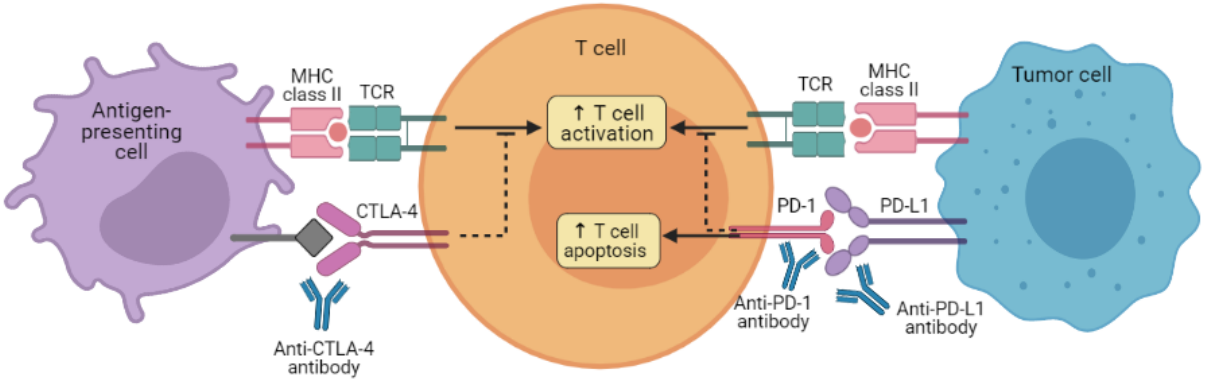

Figure 1-2. Mechanisms of anti-PD-1/PD-L1 and anti-CTLA-4 immunotherapy. Adapted from Ramos-Casals et al.[174] Created with BioRender.com.

Antibodies targeting PD-1/PD-L1 and CTLA-4 have demonstrated efficacy in treating several cancer types, including NSCLC. Multiple clinical trials have shown that these immune checkpoint inhibitors (CPI) can increase progression-free survival and overall survival in patients with metastatic NSCLC compared to traditional cytotoxic chemotherapy.[158,161,162] In studies comparing nivolumab, a PD-1 CPI, to docetaxel, a cytotoxic chemotherapy, for the treatment of non-squamous and squamous NSCLC, the overall survival in the nivolumab group was 12.2 months compared to 9.4 months in the docetaxel group for non-squamous NSCLC and 9.2 months compared to 6.0 months for squamous NSCLC.[164,165] Nivolumab has also been tested in combination with a CTLA-4 antibody, ipilimumab. Individuals treated with combination nivolumab and ipilimumab had a median overall survival of 17.1 months compared to individuals receiving chemotherapy (median overall survival 13.9 months).[166] In comparison to nivolumab monotherapy, individuals treated with nivolumab and ipilimumab had higher rates of objective response, particularly if their tumor had high levels of PD-L1 expression. In individuals whose PD-L1 expression was at least 50%, 92% of individuals treated with combination nivolumab and ipilimumab experienced a partial or complete response compared to 50% of individuals treated with nivolumab alone.[167] However, overall survival was fairly similar between nivolumab monotherapy and nivolumab and ipilimumab combination therapy (median 5.7 months vs. 4.7 months).[168] Another PD-1 CPI,

pembrolizumab, demonstrated improved progression-free survival, with a median progression-free survival of 10.3 months in the pembrolizumab group compared to 6.0 months in individuals treated with platinum-based chemotherapy.[169] Studies comparing combination pembrolizumab and chemotherapy to chemotherapy alone have demonstrated similar survival benefits with the addition of pembrolizumab.[170,171] The PD-L1 CPI atezolizumab also led to improved overall survival compared to docetaxel, with median overall survival of 13.8 months and 9.6 months, respectively.[172] When combined with chemotherapy, atezolizumab improved progression-free survival to a median 8.3 months compared to 6.8 months for chemotherapy alone.[173] Based on the results of these clinical trials, several CPI have been approved for use as first- or second-line therapy in the treatment of NSCLC. Additional clinical trials are ongoing to determine whether combining CPI with cytotoxic chemotherapy, targeted therapy, or other CPI confers additional survival benefits compared to single agent CPI therapy.[155,161]

While CPI represent a promising treatment option, they can also trigger immune-related adverse events (irAEs). The severity of irAEs can vary from mild, transient conditions to severe, permanent disorders.[174] They can affect any organ system and occur at any point in treatment, though the median onset is typically 2-16 weeks after the initial CPI dose.[174] The frequency of particular irAEs varies based on CPI mechanism, with colitis, hypophysitis, and rash more frequently seen with CTLA-4 inhibitors while pneumonitis, hypothyroidism, arthralgia, and vitiligo are more common with PD-1/PD-L1 inhibitors. The mechanisms underlying these differences are not well understood.[174,175] The frequency of severe irAEs is higher in individuals treated with CTLA-4 antibodies than in those treated with PD-1/PD-L1 antibodies, and the overall rate of irAEs is higher in individuals who receive combination CTLA-4 and PD-1/PD-L1 therapy.[174,175] Risk factors for the development of irAEs are still being elucidated, but a personal or family history of autoimmune disease has been associated with increased irAE rates, suggesting a potential genetic mechanism.[174]

Although irAEs can lead to lifelong conditions, they may be a positive prognostic marker. The development of irAEs has been associated with increased progression-free survival, overall survival, and overall response rate.[174,176–178] This finding has been more consistently demonstrated in individuals treated

with PD-1/PD-L1 antibodies.[176–178] Research is ongoing to determine whether the timing, severity, or type of irAE are associated with differences in survival.[174,176] Recent meta-analyses have found that skin, endocrine, and gastrointestinal irAEs are associated with increased OS compared to liver and lung irAEs, though gastrointestinal irAEs did not appear to increase PFS.[177,178] Given both the prognostic and long-term health consequences of irAEs, there is a great deal of interest in identifying predictors of irAEs. Small biomarker studies have identified laboratory values and pre-existing antibodies that may be predictors of increased irAEs, but these findings have not been replicated in larger populations and their predictive utility is unclear.[179,180]

1.3 Statistical Techniques to Assess Shared Genetic Architecture

1.3.1 Polygenic Risk Scores

Polygenic risk scores (PRS) are a valuable tool in genetics, with applications in both clinical and research settings. PRS aggregate multiple genetic variants from genome-wide association studies (GWAS) to summarize the genetic risk for a particular trait. Multiple techniques have been developed to build PRS, which rely on clumping and thresholding or beta shrinkage approaches (Figure 1-3). Clumping and thresholding approaches first use a clumping algorithm to identify a relatively independent set of single nucleotide polymorphisms (SNPs) by removing SNPs in linkage disequilibrium (LD). The thresholds used for clumping vary, but one of the most popular methods, PRSice, uses a default clumping r^2 of 0.1.[181,182] Various p-value thresholds are used to select SNPs from the base GWAS for inclusion in the PRS, and the association between the PRS and the trait of interest can then be ascertained in an independent target dataset. Shrinkage approaches use LD reference panels and statistical methods to estimate the true effect sizes of SNPs in the original GWAS, and the estimated effect sizes are then used as weights for the PRS.[182–185]

While PRS are being developed to predict disease risk or prognosis in clinical settings, PRS are also a useful research tool for identifying shared genetic architecture between traits. PRS have been successfully employed to understand relationships between related and comorbid conditions.[185,186] PRS

for a number of psychiatric conditions have been used to identify genetic relationships between conditions such as schizophrenia, bipolar disorder, attention deficit-hyperactivity disorder, and major depressive disorder.[187–193] Similar analyses have identified shared genetic risk factors across cancer types and between a variety of cardiovascular, neurological, psychiatric, and inflammatory conditions.[194–199] Broad scale analyses of PRS using electronic health record (EHR) data have also been used to identify associations between conditions across a variety of phenotypes through the use of phenome-wide association studies (see 1.3.2).

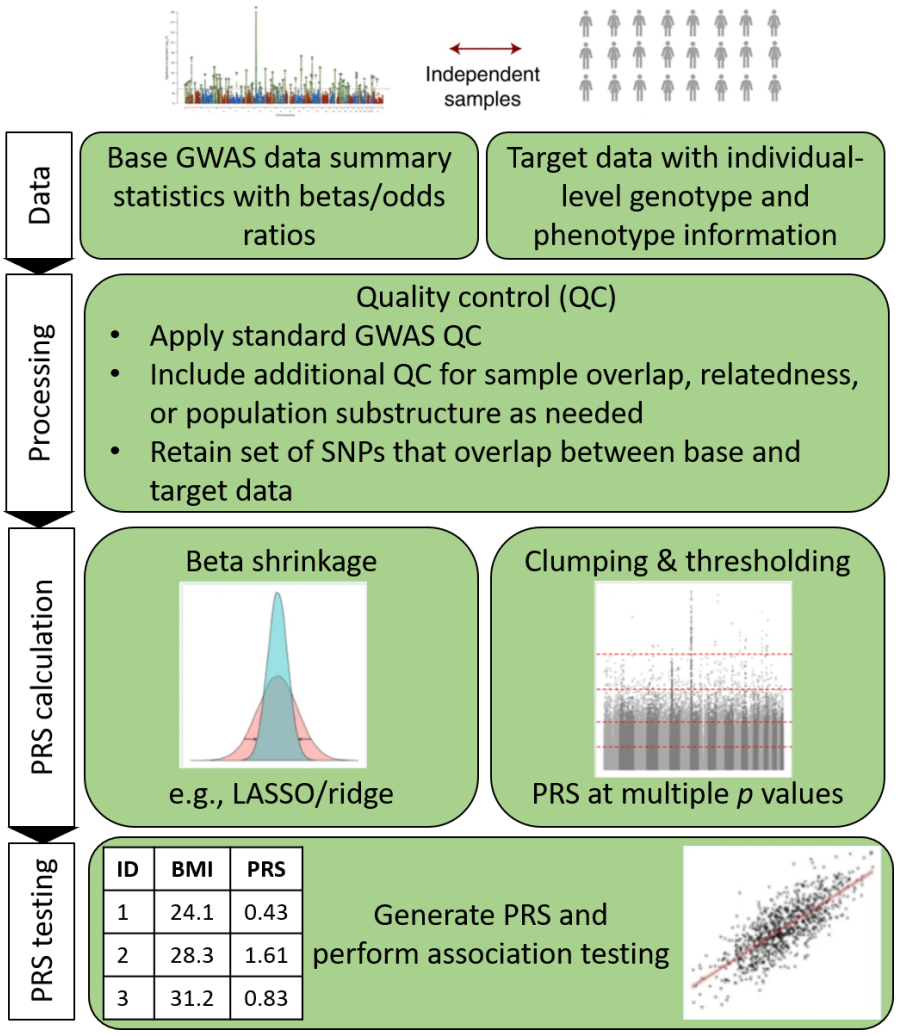

Figure 1-3. Overview of polygenic risk score development. Adapted from Choi et al.[182]

1.3.2. Phenome-wide Association Studies

Phenome-wide association studies (PheWAS) are tools that leverage the density of information available in EHRs. PheWAS test for associations with multiple clinical phenotypes documented within the EHR, defined by International Classification Disease (ICD) codes. However, ICD codes are designed to be granular for billing purposes, so a particular phenotype may be captured through a variety of related codes. To account for this relatedness and decrease multiple testing burden, a classification scheme that collapses similar ICD codes into a single phecode has been developed. Phecodes have the added benefit of allowing for a specification of a minimum number of ICD codes before an individual can be classified as a case (by default, two or more codes) and incorporate exclusion codes to ensure the control population for each phecode is not contaminated by individuals with similar phenotypes. Phecodes have been shown to be an accurate tool for classifying disease phenotypes, performing better than ICD codes alone.[200,201]

PheWAS are particularly valuable tools in settings with EHR data linked to a DNA biobank, as they allow the investigation of genetic relationships across the medical phenome.[202–205] Initial PheWAS used single SNPs as predictors, proving the utility of the technique and confirming previously identified associations found in prior GWAS studies.[206,207] These studies were also able to highlight novel associations between genetic markers for one trait and other phenotypes in the EHR, suggesting shared biology.[202–207] More recently, PheWAS using PRS have expanded upon the early studies and broadened our understanding of the genetic relationship between a variety of medical conditions.[208–217] PheWAS are a discovery tool able to systematically investigate relationships between PRS for one disorder and other traits that may otherwise have been overlooked in studies using only traits of interest selected *a priori*.

1.4 Motivation for the Research

1.4.1 Lack of COPD Research in EHR Biobanks

COPD is a major public health concern due to the high prevalence, mortality, and economic burden of the disease.[218–220] EHRs represent a valuable research resource for COPD, as highlighted by

national strategic plans released by the Centers for Disease Control and Prevention and the National Institutes of Health.[219,220] Research studies using EHR data can evaluate diagnoses, treatments, and clinical outcomes of COPD, as well as the relationship between COPD and common comorbidities. Furthermore, the depth of information available in EHRs can be leveraged to identify sub-phenotypes of COPD, a disease that is otherwise quite heterogeneous and therefore difficult to treat or predict prognoses and outcomes.[221,222]

Another major issue in research is the lack of racial diversity in most biomedical research studies.[223,224] This is true for COPD research as well, where Black and African American populations have been poorly represented in COPD studies.[97] Black and African American individuals have been shown to develop COPD with lower smoking amounts than other racial/ethnic populations, which may result from differences in smoking behaviors or nicotine metabolism.[97,225–227] Furthermore, Black individuals have been found to have lower baseline lung function levels than non-Hispanic Whites, which may also contribute to increased COPD susceptibility at lower smoking levels.[97,228–231] Black individuals are more likely to be undiagnosed and less likely to have access to appropriate COPD care than non-Hispanic Whites, meaning that the true burden of disease in Black populations may not be fully understood.[18,97,232,233] Studies also suggest that Black individuals with COPD experience more severe exacerbations, decreased quality of life, and poorer outcomes from comorbid conditions compared to their non-Hispanic White counterparts.[97,233–235] Research studies inclusive of underrepresented populations would help address these important research gaps and disparities.

1.4.2 Unknown Mechanism for Relationship Between COPD and Major Depressive Disorder

Psychiatric comorbidities are a common occurrence in individuals with COPD.[100,101,236–241] In particular, individuals with COPD have an increased prevalence of major depressive disorder (MDD), with estimated prevalence rates ranging from 8-80%.[236–241] This wide range of prevalence estimates is largely due to differences in how depression was defined and assessed between studies and the population being studied.[239] A large multi-national meta-analysis restricting to studies using a clinically validated

depression screening instrument demonstrated an overall depression prevalence of 27.1% in individuals with COPD compared to 10.0% in individuals without COPD.[239] Individuals with COPD and MDD have higher rates of exacerbations and hospital re-admissions, decreased medication adherence, poorer quality of life, and increased mortality.[236–238,242–247] While the frequent co-occurrence of COPD and MDD is well-established, the mechanisms underlying this relationship are unknown. Several mechanisms have been proposed, including systemic inflammation, hypoxemia and oxidative stress, or shared risk factors.[238,248–250] Smoking plays an important role in both COPD and MDD, though the direction of the relationship between smoking and MDD remains controversial.[250–252] COPD and MDD may also be linked through shared genetic risk factors.[87,253–255] Previous genetic studies of COPD and lung function have identified SNPs that are also associated with MDD,[87,253] but a GWAS of depressive symptoms in smokers with COPD failed to identify any significant loci.[254] Further studies are therefore needed to investigate the relationship between COPD and MDD.

1.4.3 Limited Study of the Role of Genetics in irAEs

While immune-related adverse events (irAEs) have been largely associated with improved outcomes in lung cancer patients treated with immunotherapy,[174,176–178] the biological mechanisms of irAEs are still under investigation. Previous research on irAE mechanisms have suggested a potential role for increased proliferation and activation of immune cells, leading to the production of auto-reactive cells and antibodies.[256,257] However, it is currently unknown whether irAEs result from a pre-existing underlying susceptibility to autoimmune disease. In small studies of individuals with pre-existing autoimmune conditions who received immunotherapy, up to 75% experienced a disease flare and/or developed an irAE.[258,259] These findings suggest that irAEs are more common in individuals predisposed to autoimmune conditions, but it is unknown whether genetic factors associated with autoimmunity are involved in the development of irAEs.[260] Furthermore, prior studies have largely focused on all irAEs collectively rather than looking at specific irAEs. Investigation of genetic risk factors for specific irAEs

could elucidate not only biologic mechanisms but also predict individuals more likely to respond to immunotherapy.

1.5 Research Aims to Be Addressed

This dissertation will leverage genetic techniques to understand the relationship between lung cancer, COPD, and comorbid conditions via the following aims:

1. Develop a phenotyping algorithm to identify COPD cases in EHRs (Chapter 2)
2. Examine the genetic relationship between COPD and MDD within an EHR (Chapter 3)
3. Investigate genetic predictors of thyroid irAEs in lung cancer patients treated with immunotherapy in an EHR (Chapter 4)

Through these research aims, we will be able to improve our understanding of the biology underlying COPD and lung cancer and related conditions. By building a phenotyping algorithm to identify individuals with COPD in electronic health records, we will open numerous opportunities for biomedical research using large-scale digitized health records. Our investigations of the genetic relationship between COPD and MDD will elucidate the potential shared biology of these disorders which may allow for better identification and future treatment of individuals suffering from both conditions. Finally, our study of thyroid irAEs in lung cancer patients will provide needed information on the potential mechanisms of irAEs among patients receiving immunotherapy. Overall, this research will address important gaps in our current understanding of COPD and lung cancer, two leading causes of morbidity and mortality worldwide.

CHAPTER 2

Clinical Features of COPD Patients in Electronic Health Records

2.1 Introduction

Chronic obstructive pulmonary disease (COPD) is a leading cause of death globally.[261] The clinical gold standard for COPD diagnosis is demonstration of irreversible airflow limitation assessed via pulmonary function testing (PFT).[45] However, routine screening for COPD is not recommended, particularly among asymptomatic patients,[262] and PFTs are underutilized in clinical settings.[263] Discovering a reliable means to identify COPD cases in electronic health records (EHRs) in the absence of PFTs would enable a wide variety of research applications.

EHRs provide a valuable tool to expand and accelerate COPD research.[220] To date, most research studies have relied on well-curated cohorts to further our understanding of COPD risks and outcomes.[66,68,264–266] These observational cohorts are costly and time-consuming to develop, hampering rapid advances in research. Furthermore, these cohorts often lack representation from diverse populations, recruiting primarily individuals of European descent. Limited research has been conducted on COPD in Black individuals, despite evidence suggesting differences in disease presentation, progression, and outcomes compared to Whites.[97] EHRs also allow opportunities for pragmatic clinical trials, with an unprecedented depth of digitized information that can be used to rapidly and cost-effectively identify study participants from diverse populations, investigate medical interventions and outcomes in real-world settings, and conduct personalized medicine research.[267–271] EHRs have been identified as a valuable research tool to address research gaps in patient-centered outcomes, COPD sub-phenotypes, and comorbidities.[220,271,272] Linking EHR data with genetic information provides additional opportunities to investigate genetic factors involved in COPD development, sub-phenotypes, responses to treatment, and clinical outcomes. EHRs contain the necessary information to evaluate potential laboratory and imaging biomarkers for COPD diagnosis and prognosis.[271] EHR data can also be used to identify system-wide

gaps in COPD management and initiate targeted quality improvement projects to increase compliance with recommended guidelines.[273–275] While several algorithms have been developed for COPD,[276–290] to our knowledge only one was developed in a large-scale EHR system, and its performance metrics were not reported.[85]

Vanderbilt University Medical Center (VUMC) has a well-characterized EHR of clinical data captured over several decades through routine care. We used de-identified data from the VUMC clinical population to develop and evaluate EHR-based COPD phenotyping algorithms.

2.2 Methods

2.2.1 Vanderbilt University Medical Center Synthetic Derivative

We used clinical data from the Synthetic Derivative, a de-identified version of the VUMC EHR data warehouse, containing data on over 2.1 million adult patients and over 1 billion unique observations dating back to the 1980s.[291] Details regarding Synthetic Derivative development have been previously published.[291] Extractable PFT data have been available in the EHR since 2011. The Vanderbilt University Institutional Review Board approved this study.

The study population consisted of adult patients over 45 years of age at last clinic visit who visited VUMC prior to March 8, 2019. We then filtered to identify a medical home population, defined as patients having a minimum record length of 180 days (6 months). Demographic data, International Classification of Disease (ICD)-9 and ICD-10 codes, laboratory data, and PFTs were obtained from structured fields in the Synthetic Derivative. Quality control was implemented to remove individuals having ages inconsistent with their record length (defined as the number of days between the first clinical encounter and last clinical encounter, n = 109). Smoking information (ever/never) was collected from unstructured clinical notes.

2.2.2 Algorithm Development

PFTs were used as the clinical gold standard, with COPD defined as the ratio of forced expiratory volume in one second (FEV_1) to forced vital capacity (FVC) < 0.7 after bronchodilator administration. We chose to focus on patients with PFTs during the algorithm development since PFTs are an objective measurement required for definitive COPD diagnosis in clinical settings.[292]

We developed computable phenotyping algorithms for COPD case status combining ICD codes and additional clinical information extracted from "problem lists", "radiology reports", and "medication lists". To evaluate the performance of the phenotyping algorithms, comparisons were made with PFT defined COPD case status (post-bronchodilator $FEV_1/FVC < 0.7$). Cases were required to have at least one ICD-9 (i.e. 491.x, 492.x, 496.x) or ICD-10 code (i.e. J41.x, J42.x, J43.x, or J44.x) alone or in combination with "oxygen" or "O2" on the "problem list". This resulted in two algorithms: a code algorithm and a code+keyword algorithm. Controls were required to have no ICD codes for COPD, asthma, sarcoidosis, and idiopathic pulmonary fibrosis (IPF). We calculated performance metrics, including sensitivity, specificity, PPV, negative predictive value (NPV), and F measure for each algorithm.

2.2.3 Algorithm Validation

To internally validate our phenotyping algorithms, we tested them in an independent random sample of 200 clinical charts. Stratified random sampling was used to select 100 individuals with two or more COPD ICD-9 or ICD-10 codes and 100 individuals with fewer than two COPD ICD codes to ensure potential COPD cases were well-represented in the chart review set. Gold standard chart review was performed by two independent reviewers with clinical training. Discrepancies were adjudicated by a pulmonary physician. We calculated sensitivity, specificity, PPV, NPV, and F measure, as above. Kappa statistics between reviewers and percent agreement were calculated using the R package *irr*.[293]

2.2.4 Genetic Data and Polygenic Risk Score Development

Our study population included European and African ancestry participants with available Illumina MEGA-Ex array genotyping data.[291] Data from the Illumina MEGA-Ex array were subjected to quality control to remove individuals and single nucleotide polymorphisms (SNPs) with <98% call rates, SNPs with minor allele frequency <1%, and SNPs not in Hardy-Weinberg equilibrium (p-value < $1x10^{-6}$). Imputation was performed using the Michigan Imputation Server with the Haplotype Reference Consortium reference panel.[294] Principal component analysis was performed using EIGENSTRAT and a set of SNPs pruned for linkage disequilibrium using a window size of 50 kb, a step size of 5 kb, and a r^2 threshold of 0.2.[295] Ancestry was defined by proximity to European and African ancestry reference populations on principal component plots. After quality control, genotyping data were available on 48,147 European ancestry and 5,852 African ancestry adults over 45 years of age.

We tested a polygenic risk score (PRS) for COPD previously developed in a genome-wide association study of lung function traits in 400,102 individuals of European ancestry.[87] Risk scores were calculated for each individual in our dataset using reference alleles and weights from the original study and the score function in Plink v1.9.[296] Of the 279 SNPs included in the original PRS, 188 were directly genotyped and passed quality control in the European ancestry dataset and 172 were directly genotyped and passed quality control in the African ancestry dataset. We used LDproxy[297] to identify proxy variants for the missing SNPs, using 1000 Genomes reference populations[298] CEU for the European ancestry and YRI and ASW for the African ancestry study populations. Using an r^2 threshold of 0.8 to identify proxy SNPs resulted in a lung function PRS comprised of 247 SNPs in our European ancestry dataset and 225 SNPs in our African ancestry dataset. We tested for associations between this lung function PRS and COPD case status in both the African ancestry and European ancestry study populations using a logistic regression model, adjusted for age at last clinic visit, sex, ever/never smoking status, and the first 3 principal components.

2.2.5 Alpha-1-antitrypsin Laboratory Data

Since alpha-1-antitrypsin testing is recommended for all individuals undergoing diagnostic evaluation for COPD,[299,300] we explored how frequently this lab test was ordered for individuals in our COPD case and control sets. Laboratory data was extracted from structured fields in the Synthetic Derivative. We identified all individuals with a reported laboratory value alpha-1-antitrypsin, regardless of the actual value.

2.3 Results

2.3.1 Study Population

We identified 1,008,661 individuals age 45 or older at last clinic visit. Quality control removed 109 individuals, leaving 1,008,552 eligible individuals. The median age at last clinic visit was 61 years (interquartile range 53-71), with a median record length of 1.9 years (interquartile range 0.1-8.2 years). The study population included a greater percentage of females than males (54.1% vs 45.9%) and was primarily observer-reported European descent (69.7%). Prevalence of ever smoking in the population was 18.9%.

2.3.2 Algorithm Development and Validation

We selected a computable phenotyping algorithm that met at least an 80% positive predictive value based on the clinical gold standard of PFT-defined COPD. The algorithm required cases to be at least 45 years of age. The code+keyword algorithm required COPD cases to have ten or more COPD ICD codes (491.x, 492.x, 496.x, J41.x, J42.x, J43.x, or J44.x) OR three to nine COPD ICD codes AND a mention of oxygen on the problem list (Figure 2-1). The optimal numbers of ICD codes for the code-only and code+keyword algorithms was determined based on the performance of code-only algorithms consisting of one to fifteen ICD codes for COPD (Figure 2-2). Controls were required to have no ICD codes for COPD, asthma, sarcoidosis, or IPF (Figure 2-1).

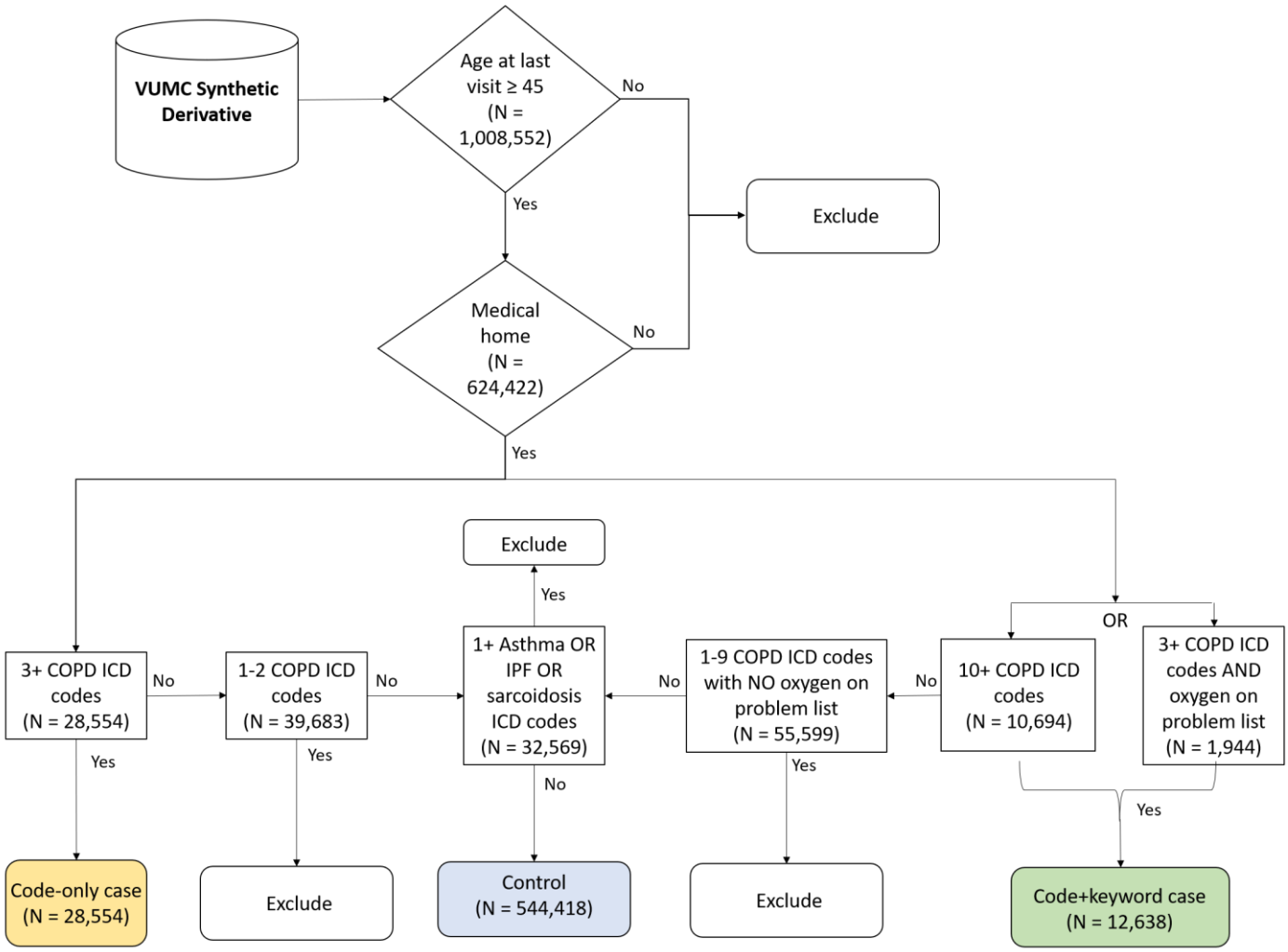

Figure 2-1. Phenotyping algorithm schematic for chronic obstructive pulmonary disease (COPD).

We evaluated the internal validity of our electronic phenotyping algorithms in a comprehensive chart review set of 200 charts. The kappa statistic between the two clinical reviewers was 0.75 and the percent agreement was 91%. Eighteen discrepancies were adjudicated by the third reviewer. Both phenotyping algorithms had comparable sensitivity and NPV in the development and validation sets, but the specificity and PPV were higher for the code+keyword algorithm (

Table 2-1).

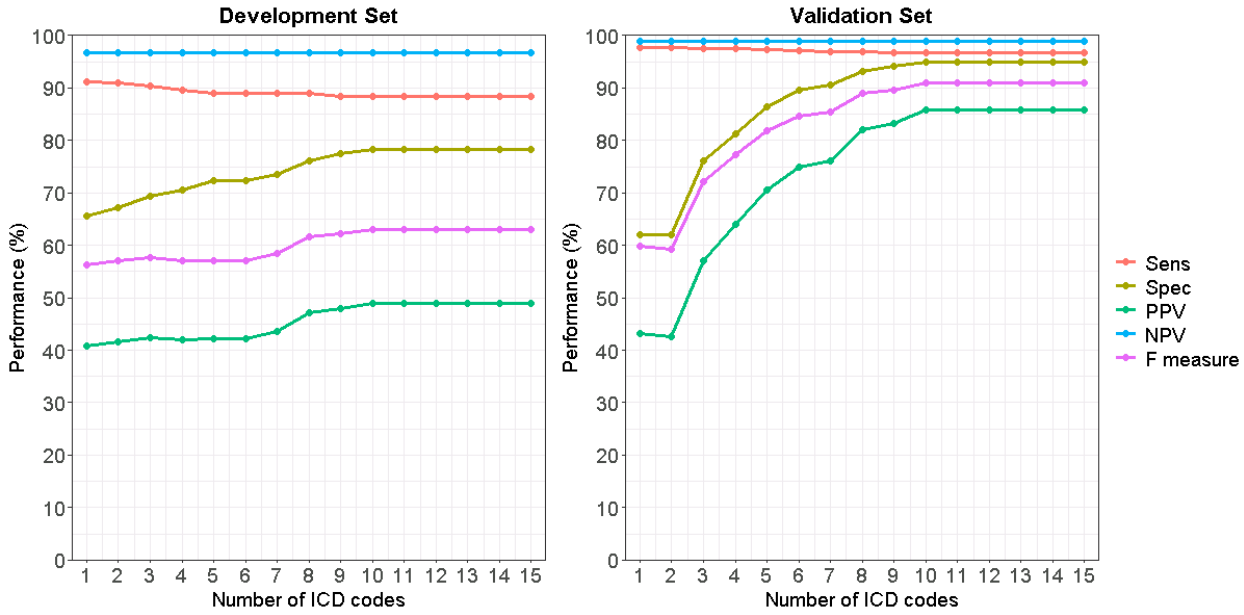

Figure 2-2. Performance of code-only algorithms in the development (left) and validation (right) sets. Sens = sensitivity, spec = specificity, PPV = positive predictive value, NPV = negative predictive value.

Table 2-1. Clinical validity test properties for each phenotyping algorithm and prevalence within the electronic health record among participants in the Vanderbilt University Medical Center Synthetic Derivative, November 2020.

Metric	Code-only		Code+Keyword	
	Development	Validation	Development	Validation
Sensitivity (%)	90.3	100	88.9	100
Specificity (%)	68.7	71.9	76.7	94.1
PPV (%)	43.8	60.9	50.0	88.6
NPV (%)	96.3	100	96.3	100
F measure	0.59	0.76	0.64	0.94
COPD prevalence based on algorithm (%)	32.0	32.0	24.0	17.5

PFT = pulmonary function test, PPV = positive predictive value, NPV = negative predictive value

2.3.3 Application of Phenotyping Algorithms to Electronic Health Records

To demonstrate the utility of our algorithms to identify COPD cases within a large-scale EHR database, we applied them to an independent adult population over 45 years of age at last visit with a record length floor of 6 months (≥180 days, N = 623,986). The code-only algorithm identified 28,544 COPD cases (4.6% COPD prevalence), while the code+keyword algorithm identified 12,638 COPD cases (2.0% COPD prevalence). Both COPD case sets had similar median ages at last clinic visit and distributions of sex and race. The prevalence of COPD by race was similar in both algorithm definitions, with a 3.5% prevalence in Whites and 3.6% prevalence in Blacks for the code-only algorithm and a 1.6% prevalence in Whites and a 1.7% prevalence in Blacks for the code+keyword algorithm. Code+keyword COPD cases had a longer median record length (8.3 years) than code-only cases (7.3 years), and the percentage of ever smokers was higher among code+keyword cases (60.8%) than code-only cases (55.8%). Controls were younger than COPD cases, less likely to be smokers (23.0%), and more likely to be female (55.2%) and non-White than cases (21.3%) (Table 2-2). Among individuals with available PFT data (N=14,281), code+keyword cases had lower median FEV_1 and FVC percent predicted values in both pre- and post-bronchodilator measures than code-only cases, while controls had higher PFT values versus both case groups (Figure 2-3). Using the GOLD definition of COPD, i.e. fixed ratio of post-bronchodilator $FEV_1/FVC < 0.7$, we identified 2,226 PFT-defined COPD cases (44.8% COPD prevalence among individuals with available PFT data).[292]

Table 2-2. Demographic characteristics of adults over age 45 years, Vanderbilt University Medical Center Synthetic Derivative, November 2020.

Characteristic	**Code-only cases N = 28,554**	**Code+keyword cases N = 12,638**	**Controls N = 544,418**	**PFT cases N = 2,226**	**PFT controls N = 11,553**
Median age at last clinic visit, years (IQR)	69 (60-76)	70 (62-77)	62 (53-71)	69 (61-75)	64 (55-71)
Sex, N (%)					
Female	13,369 (46.8)	6,167 (48.8)	300,575 (55.2)	1,068 (48.0)	6,347 (54.9)
Male	15,185 (53.2)	6,471 (51.2)	243,775 (44.8)	1,158 (52.0)	5,206 (45.1)
Unknown	0	0	68	0	0
Race, N (%)					
White	24,821 (86.9)	11,025 (87.2)	428,615 (78.7)	1,991 (91.9)	9,861 (85.4)
Black	2,689 (9.4)	1,251 (9.9)	46,021 (8.5)	158 (7.3)	1,240 (10.7)
Other	120 (0.4)	47 (0.4)	6,350 (1.2)	18 (0.8)	169 (1.5)
Unknown	924 (3.2)	315 (2.5)	63,432 (11.7)	59 (2.7)	283 (2.4)
Median record length, years (IQR)	7.3 (3.2-13.0)	8.3 (4.0-13.9)	6.2 (2.4-11.5)	8.8 (3.1-15.1)	7.5 (2.7-14.3)
Smoking status, N (%)					
Ever smoker	15,929 (55.8)	7,683 (60.8)	125,325 (23.0)	1,747 (78.5)	5,642 (48.8)
Never smoker	2,846 (10.0)	1,000 (7.9)	214,314 (39.4)	462 (20.8)	5,819 (50.4)
Unknown	9,779 (34.2)	3,955 (31.3)	204,779 (37.6)	17 (0.8)	92 (0.8)

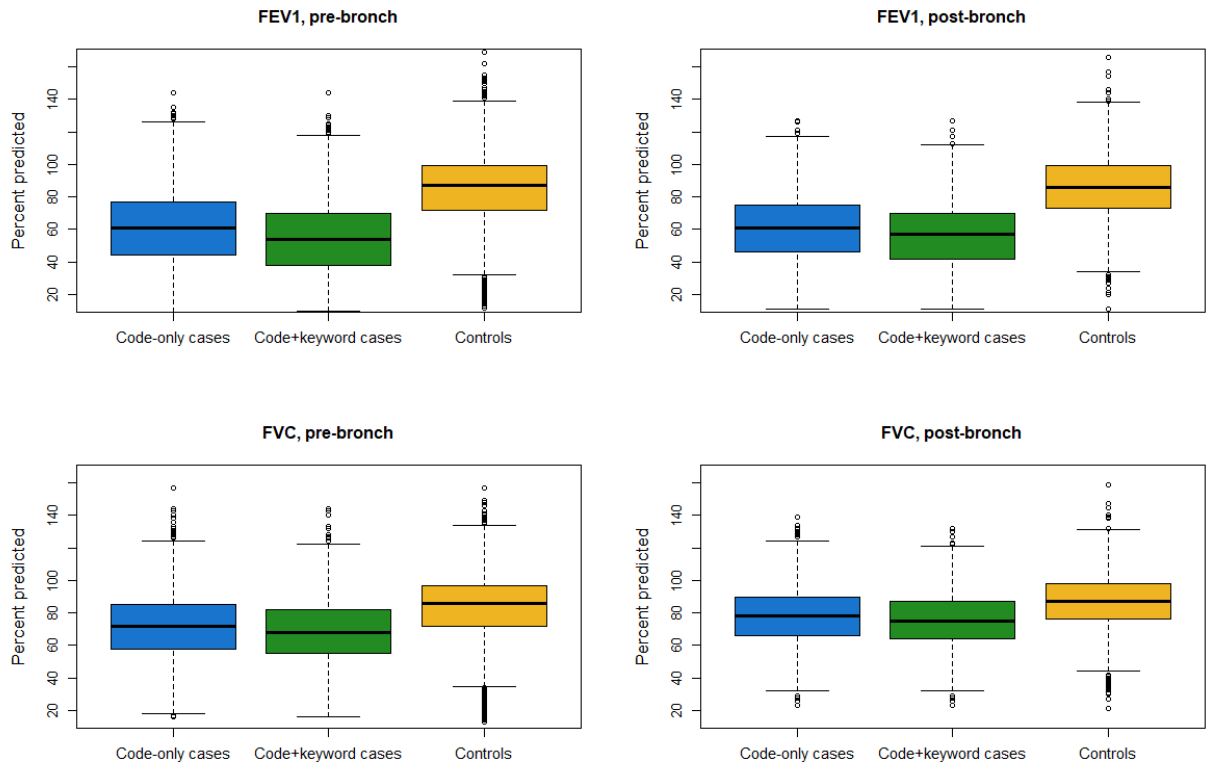

Figure 2-3. Pulmonary function test values by algorithm set among Vanderbilt University Medical Center Synthetic Derivative participants with pulmonary function test data (2011-2020) (N = 14,281). Abbreviations: FEV1: forced expiratory volume in one second; FVC: forced vital capacity; ICD: International Classification of Disease, pre-bronch: pre-bronchodilator, post-bronch: post-bronchodilator.

2.3.4 Alpha-1-Antitrypsin Tests

We characterized the frequency of alpha-1-antitrypsin tests among our COPD race-sex groups. Only 3.6% of COPD cases identified by the code-only algorithm and 5.1% of cases identified by the code+keyword algorithm had an available alpha-1-antitrypsin measurement (Table 2-3). Among PFT-defined cases, the percentage of individuals with alpha-1-antitrypsin tests was higher at 7.6%. The percentage of individuals with alpha-1-antitrypsin labs varied by sex and race (Table 2-3).

Table 2-3. Percentage of individuals with alpha-1-antitrypsin labs by case status and demographic characteristics, Vanderbilt University Medical Center Synthetic Derivative, November 2020.

Group	Code-only Cases (N, %)	Code+keyword Cases (N, %)	Algorithm Controls (N, %)	PFT Cases (N, %)	PFT Controls (N, %)
Overall	1,027 (3.6)	640 (5.1)	6,595 (1.2)	169 (7.6)	756 (6.5)
White women	466 (4.1)	292 (5.5)	2,851 (1.2)	91 (9.7)	341 (6.5)
White men	466 (3.5)	289 (5.0)	2,861 (1.5)	62 (5.9)	354 (7.7)
Black women	36 (2.6)	29 (4.2)	138 (0.5)	9 (9.6)	27 (3.3)
Black men	26 (2.1)	20 (3.7)	141 (0.7)	2 (3.1)	13 (3.1)

PFT: pulmonary function test; PFT case/control definition: post-bronchodilator forced expiratory volume/forced vital capacity < 0.7

2.3.5 Polygenic Risk Score Performance Across Ancestry Groups

We replicated a previously developed lung function PRS[87] and tested for associations with algorithm-defined COPD status in European and African ancestry adults. The lung function PRS was significantly associated with code-only case status (odds ratio (OR) = 1.11, 95% confidence interval (CI): 1.07-1.15), code+keyword case status (OR = 1.15, 95% CI: 1.10-1.20), and PFT case status (OR = 1.32, 95% CI: 1.19-1.46) (Figure 2-4). None of the three COPD case definitions were significantly associated with the lung function PRS in individuals of African descent (Figure 2-4).

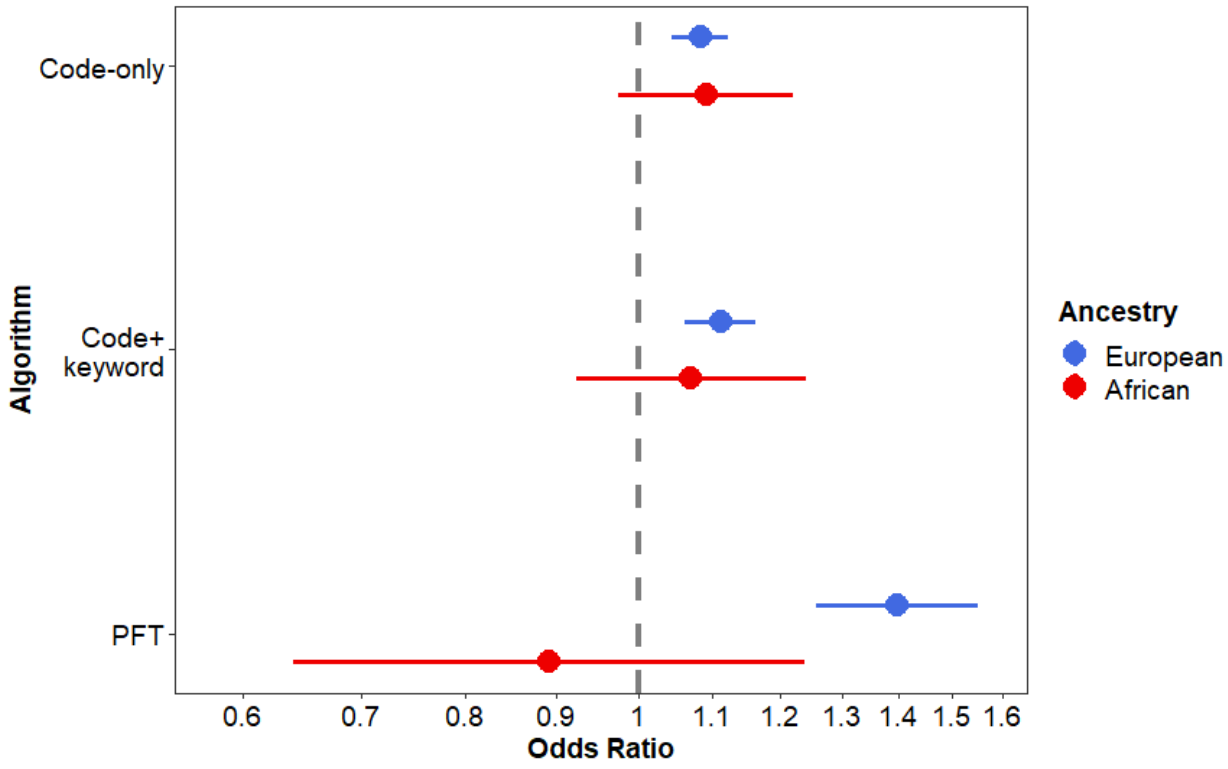

Figure 2-4. Validation of COPD algorithm demonstrated by logistic regression of the association between FEV_1/FVC polygenic risk score and COPD case-control status among Vanderbilt University Medical Center Synthetic Derivative participants (2007-2020).

FEV_1/FVC: forced expiratory volume in one second/forced vital capacity; PFT: pulmonary function test

2.4 Discussion

We sought to develop automated COPD phenotyping algorithms to facilitate biomedical research and quality improvement within EHR systems. We developed a code-only ICD code-based algorithm and a code+keyword algorithm combining ICD code information with details from the clinical record. Consistent with known COPD epidemiology,[261] the prevalence of COPD in both the code-only and code+keyword algorithms was the same across racial groups. However, the overall prevalence of COPD in both case sets (4.6% in the code-only group and 2.0% in the code+keyword group) was lower than the national average of 6.3%, suggesting additional unrecognized COPD cases may remain in our dataset. While we identified more cases with the code-only definition than the code+keyword definition, the PPV (89%) was high for the code+keyword algorithm, suggesting that the code+keyword algorithm has high accuracy for identifying individuals with COPD in EHR. Only two previous studies of COPD algorithms

achieved similar or higher PPV (89.4%[282] and 99.5%[283]), and both were developed in primary care systems and required access to medication information, which may not be reliably available in US-based EHR. Thus, adding keywords improved identification of COPD cases.

For clinical EHR systems without available clinical notes, COPD cases can still be identified, although with less accuracy. Despite this reduced accuracy, the lung function PRS, which summarizes genetic information in a potentially clinically meaningful way, provided a unique opportunity to validate our phenotyping algorithms. The lung function PRS was associated with our algorithm case definitions and PFT-defined COPD. Our genetic analyses were limited to a European ancestry-derived PRS due to the lack of an available lung function PRS developed in an African ancestry population. The Eurocentric bias in available PRS can result in poor transferability of PRS across ancestrally diverse populations.[301]

We also used our algorithm to identify compliance with recommended alpha-1-antitrypsin testing guidelines[299,300] and found that the lab was performed in less than 10% of COPD cases regardless of case definition. Black men had particularly low rates of testing. Racial disparities have been previously described in lab ordering,[302–304] but further confirmation of our findings is needed. Overall, the low rates of alpha-1-antitrypsin testing represent an opportunity for quality improvement. The World Health Organization, American Thoracic Society, and European Respiratory Society have all recommended that individuals with COPD undergo this test to screen for alpha-1-antitrypsin deficiency.[299,300] As a tertiary referral center, it is possible that some individuals with COPD at VUMC received the test at outside institutions, but our findings raise concern that this lab is being overlooked by most providers. A similar approach could be used to identify other potential gaps in COPD clinical evaluation and treatment that should be addressed at a system level.

This study has some limitations. The use of EHR in biomedical research has inherent challenges due to inconsistent documentation, missing data, and inaccuracies.[305–313] One important factor to consider when evaluating our algorithm performance is the impact of disease prevalence. In particular, PPV and NPV are dependent on prevalence of the disease of interest.[314,315] Thus, these measures of clinical validity are likely influenced by the prevalence of COPD in our population. In the validation chart review set, the

prevalence of COPD was 22.5%, yet the U.S. prevalence of COPD among adults is 6.3%, much lower than our development or validation sets ascertained from our tertiary care hospital. In populations with COPD prevalence closer to the national average, we expect that the PPV of the algorithm will decrease, though we cannot predict the magnitude of that decrease. Another consideration is that tertiary care centers often have sicker patients with denser clinical documentation and more complete data than other hospital settings, potentially impacting generalizability to non-tertiary settings.[313] COPD cases identified by our code+keyword algorithm have more severe disease compared to code-only cases based on FEV_1 and FEV_1/FVC measurements. Our algorithm used ICD code information, which can be inaccurate, particularly for secondary research use; however, this is balanced by efficiency and reduced costs within an established EHR.[316–318] Information on whether ICD codes were assigned in outpatient or inpatient settings could also increase algorithm performance. This strategy has been used to identify disease status for several conditions, including COPD, though performance metrics were not reported.[319] Future studies that incorporate this information could improve algorithm performance. However, our analyses of known COPD risk factors and previous genetic associations suggest that the population identified by our algorithm was similar to previously studied COPD populations.[33,87]

The primary motivation for this study was to develop a phenotyping algorithm for COPD that could be implemented to conduct biomedical research and pragmatic trials. It was therefore of utmost importance to identify algorithms that were easily implemented in EHR across multiple settings. ICD codes are widely available as a structured data field, which makes them easy to obtain even in institutions without advanced informatics infrastructure. One concern is the possibility for COPD phenotype misclassification. However, since COPD is both over- and under-diagnosed, we anticipate any misclassification to be non-differential and effect estimates to be biased towards the null.[320] In pragmatic clinical trials, it is important to identify a broadly representative patient population to determine real-world generalizability. Sensitivity is therefore more important than PPV in this setting.[321,322] Our algorithm has a high sensitivity, which makes it advantageous over using PFTs alone to identify individuals with COPD. Underutilization of PFTs in clinical settings is well-documented,[263,323–332] and as

few as one third of individuals with a clinical definition of COPD actually receive PFTs.[263,325–327] Furthermore, due to patient transfer between health care systems, individuals may have received PFTs at an outside institution. By using widely available ICD codes, we allow individuals without PFTs yet having COPD to be included in potential pragmatic trials within EHR.

In conclusion, we identified and characterized COPD cases using digitized EHR data and computable phenotyping algorithms. As methods evolve for computational phenotyping, our algorithm is a step towards efficient identification of large-scale populations for clinical and genetic research studies within EHR that may facilitate accelerated scientific discoveries and personalized medicine opportunities for this devastating disease.

CHAPTER 3

Fate or Coincidence: Do COPD and Major Depression Share Genetic Risk Factors?[1]

3.1 Introduction

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality globally, affecting 328 million people and causing 3 million deaths per year.[333] Comorbidities are common among COPD patients.[100,101] Individuals with comorbid conditions report decreased quality of life,[109–112] and the presence of multiple comorbidities can increase mortality rates by as much as 400%.[115] Therefore, understanding the relationship between COPD and its comorbidities is a research priority.[220]

Psychiatric comorbidities are commonly reported in COPD patients. Individuals with COPD have an increased prevalence of major depression, with estimates ranging from 8-80%.[236–241] The prevalence of depression is higher in individuals with more severe disease.[236,237,240] Among individuals with COPD, depression is associated with greater exacerbation, higher rates of hospital re-admission, decreased medication adherence, poorer quality of life, and increased mortality.[236–238,242–247]

The biologic mechanism underlying the relationship between COPD and depression is unknown. Both disorders are highly heritable, with an estimated genetic heritability of 25-37% for COPD[65] and 28-51% for major depressive disorder (MDD).[334–336] Heritability of lung function traits such as forced expiratory volume in one second (FEV_1) and forced vital capacity (FVC), which are the basis for COPD diagnosis, are also high, with estimated heritability ranging from 18-50%.[337–339] Systemic inflammation, hypoxemia and oxidative stress, and shared environmental risk factors, such as smoking, have been proposed as possible mechanisms linking these two conditions.[238,248–250] Smoking is a major risk factor for

[1] This chapter is adapted from "Fate or Coincidence: Do COPD and Major Depression Share Genetic Risk Factors?" published in *Human Mol. Gen.* and has been reproduced with the permission of the publisher and my co-authors Bradley Richmond, Lea K. Davis, Timothy S. Blackwell, Nancy J. Cox, David Samuels, Digna Velez Edwards, and Melinda C. Aldrich.

COPD, and it may also be an independent risk factor for depression, though the direction of this relationship is still debated.[250–252] Shared genetic risk factors have been investigated in a small number of studies.[87,253–255] A candidate gene study for depression identified a small number of single nucleotide polymorphisms (SNPs) associated with increased COPD risk.[253] A large-scale phenome-wide association study (PheWAS) conducted in the UK Biobank detected associations between lung function genomic loci and depressive symptoms.[87] These studies suggest that the relationship between COPD and MDD may be due to pleiotropy, where a single SNP affects two or more distinct traits.[186] However, a genome-wide association study (GWAS) of depressive symptoms in smokers with COPD did not identify any significant loci.[254] A polygenic score (PRS) built from a genome-wide gene-by-environment interaction study of depressive symptoms identified a significant association with COPD, but the underlying model assumed an interaction between SNPs and stressful life events and therefore did not examine purely genetic effects.[255] Further complicating the relationship between COPD and MDD is the presence of sex differences in both disorders. MDD is more prevalent in women, and women typically experience more severe depressive symptoms than men.[340] Genetic studies of MDD have identified evidence of sex-specific risk variants and transcriptional signatures.[341,342] Women develop COPD at lower smoke exposure than men and may experience more severe disease and rapid respiratory decline compared to men with similar smoking exposure.[343–345] We investigated the genetic relationship between COPD and MDD, using existing GWAS summary statistics to test for genetic correlation and pleiotropy between the traits. We leveraged electronic health records (EHR) linked to genotyping data to explore shared genetic associations between COPD and MDD using a PheWAS, an approach often used to examine relationships between comorbid conditions.[346–348] We also performed sex-stratified analyses to investigate possible sex differences in the relationship between MDD and COPD. An overall schematic of our study design and methods is provided in Appendix 1, Figure 6-1.

3.2 Methods

3.2.1 Study Population

Our study population included participants in the Vanderbilt University Medical Center BioVU clinical repository (2007-2019). BioVU is a DNA biobank linked to de-identified EHR clinical data, dating back to the 1980s.[291] We limited our study population to BioVU individuals of European ancestry defined by principal component analysis previously genotyped on the Illumina Infinium Multi-Ethnic Genotyping Array. Demographic data (sex, age at last record), smoking, International Classification of Disease (ICD)-9 and ICD-10 codes, and pulmonary function data (2011-2019) were extracted from structured fields in the EHR using natural language processing.

We selected individuals of European ancestry using principal component analysis implemented in EIGENSTRAT.[295,349] We performed standard quality control and imputed genotypes to the Haplotype Reference Consortium with the Michigan Imputation Server[294]. Genotypes were hard-called using default settings (probability greater than 0.1) in PLINK 1.9.[296,350]

3.2.2 GWAS Summary Statistics

To investigate potential pleiotropy between lung function and MDD, we used publicly available summary statistics from previously performed GWAS in individuals of European ancestry. Summary statistics were obtained from a large-scale GWAS of lung function (forced expiratory volume in one second [FEV_1], forced vital capacity [FVC], FEV_1/FVC, and peak expiratory flow [PEF])[87] and from a meta-analysis of two genome-wide studies of MDD.[351,352]

3.2.3 Genetic Correlation

We calculated the overall genetic correlations (Rg) between traits using Linkage Disequilibrium Score Regression (LDSC) software and a reference linkage disequilibrium (LD) score panel derived from European 1000 Genomes populations.[353,354] To calculate local genetic correlation, we used ρ-HESS with a European LD reference panel provided by the software authors.[355]

3.2.4 Polygenic Risk Scores

To build PRS, we used PRS-CS (polygenic risk score-continuous shrinkage) to estimate posterior effect sizes of SNPs with continuous shrinkage priors in each GWAS.[356] We then applied the score function in PLINK 1.9[296,350] to calculate a PRS for each individual in BioVU. PRS were normalized by subtracting the mean and dividing by the standard deviation.

3.2.5 PheWAS

We explored the relationship between each PRS and EHR phenotypes in a PheWAS.[205] We performed logistic regression analysis to examine associations between PRS and 1,857 phecodes. Phecodes are defined by aggregating similar ICD-9 and ICD-10 billing codes[200,201] and have been used extensively in prior studies.[202,203,207,211,213–217,357–361] We mapped extracted ICD-9 and ICD-10 billing codes from BioVU to phecodes using the PheWAS R package.[362] Phecodes with fewer than 20 cases were excluded from analyses. Models were adjusted for age at last visit, sex, smoking (ever/never), and 3 PCs estimated using EIGENSTRAT[295,349] to adjust for potential confounding by genetic ancestry. We also performed sex-stratified PheWAS using the same parameters and covariates as in the main analysis, with the exception of sex as a covariate. A type 1 error rate of alpha = 0.05/1,857 phecodes = 2.69×10^{-5} was set for inference of statistical significance.

3.2.6 Multi-trait Conditional Analysis

We performed multi-trait-based conditional and joint analysis (mtCOJO) to investigate cross-phenotype effects.[363] We evaluated the change in effect size for SNPs in the FEV_1/FVC GWAS before and after conditioning on MDD. We also implemented heterogeneity in dependent instrument outlier approach (HEIDI-outlier), incorporated into mtCOJO methods, to detect potentially pleiotropic SNPs.[363] We used the NHGRI-EBI GWAS Catalog[364] and the NHLBI Genome-Wide Repository of Associations Between SNPs and Phenotypes[365] to look up prior associations for identified SNPs.

3.3 Results

3.3.1 Study Population

Our BioVU study population consisted of 72,447 European ancestry individuals with 9,386,383 SNPs. Approximately 5% of the BioVU population had a COPD phecode. COPD individuals were older (median age 68 years) and more male (53.5%) than the overall study population (median age 56 years and 44.0% male). COPD patients had a higher prevalence of ever smoking (87.6%) than the overall BioVU population (49.8%). The prevalence of major depression (one or more depression phecodes) was higher in COPD patients (8.8%) than among patients without a diagnosis of COPD (3.5%) (Table 3-1).

Table 3-1. Demographics of European ancestry BioVU population (2007 – 2019).

Characteristic	**COPD Phecode (N = 3,466)**	**No COPD Phecode (N = 68,981)**	**Total (N = 72,447)**
Median age (IQR)	68 (60-76)	55 (35-68)	56 (36-68)
Gender (N, %)			
Female	1,615 (46.6)	38,969 (56.5)	40,584 (56.0)
Male	1,851 (53.4)	30,010 (43.5)	31,861 (44.0)
Missing	0	2	2
Smoking status (N, %)			
Ever	2,435 (83.9)	20,861 (41.2)	23,296 (43.5)
Never	455 (16.1)	29,741 (58.8)	30,207 (56.5)
Missing	565	18,379	18,944
Major depressive disorder (N, %)	305 (8.8)	2,385 (3.5)	2,690 (3.7)

COPD: chronic obstructive pulmonary disease

3.3.2 Genetic Correlation Between MDD and Lung Function

We found low genetic correlations between MDD and lung function traits using LDSC. None of the genetic correlations between MDD and lung function were statistically significant. The strongest correlation between MDD and lung function was with PEF (Rg = -0.035, p=0.07). In contrast, we observed strong and statistically significant correlation between lung function traits (Table 3-2). Local genetic correlation showed statistically significant peaks in Rg on chromosome 6 for both FEV_1/FVC (Bonferroni-corrected p-value = 8.62×10^{-3}) and FEV_1 and MDD (Bonferroni-corrected p-value = $4.38 \times 10^{-}$

[6]). However, the maximum correlation values were still small (maximum Rg for FEV_1/FVC and MDD: 3.86×10^{-4}, maximum Rg for FEV_1 and MDD: 3.36×10^{-4}).

Table 3-2. Genetic correlation between major depressive disorder and lung function traits.

Phenotype 1	**Phenotype 2**	**Rg**	***P* value**
MDD	FEV_1/FVC	-0.0011	0.95
MDD	FEV_1	-0.0325	0.07
MDD	FVC	-0.0307	0.10
MDD	PEF	-0.0351	0.07
FEV_1/FVC	FEV_1	0.4046	2.66×10^{-89}
FEV_1/FVC	FVC	-0.0841	3.20×10^{-5}
FEV_1/FVC	PEF	0.6273	0
FEV_1	FVC	0.877	0
FEV_1	PEF	0.7058	0
FVC	PEF	0.4351	1.28×10^{-136}

FEV_1: forced expiratory volume in one second, FVC: forced vital capacity, MDD: major depressive disorder, PEF: peak expiratory flow

3.3.3 PheWAS Analyses with Lung Function and MDD PRS

We built PRS for lung function (818,738 SNPs) and MDD (803,205 SNPs) from publicly available GWAS summary statistics for lung function measures and MDD. To confirm expected associations with lung function, we used linear regression to test for the association between the lung function PRS and their corresponding pre-bronchodilator lung function traits in a subset of BioVU patients with available lung function data. The PRS were robustly associated with the corresponding lung function traits (Appendix 1, Table 6-1). We performed a PheWAS using logistic regression models to examine associations between PRSs and 1,857 phecodes in the entire study population. Cases and controls were defined independently for each phecode, and phecodes with less than 20 cases were excluded (N = 438 phecodes). The lung function PRS were consistently associated with decreased COPD in the PheWAS (Table 3-3). Similar associations were observed in sex-stratified analyses, though the significance of the association varied between lung function phenotypes (Appendix 1, Figure 6-2 to Figure 6-5 and Table 6-11 to Table 6-18). The MDD-PRS was significantly associated with increased risk

of mood disorders (odds ratio [OR]=1.28, 95% confidence interval [CI]: 1.25-1.32; $p=6.42x10^{-76}$) and MDD (OR=1.27, 95% CI: 1.22-1.32; $p=1.41x10^{-31}$) when adjusting for age, sex, and the first 3 PCs (Table 3-3). In sex-stratified analyses, the MDD-PRS was also significantly associated with mood disorders and MDD (Appendix 1, Figure 6-6 and Table 6-9, Table 6-10).

Table 3-3. Association of lung function and MDD PRS with COPD and MDD in European BioVU participants (2007 – 2019).

	COPD				**MDD**			
PRS	**OR[1]**	**95% CI[1]**	**OR[2]**	**95% CI[2]**	**OR[1]**	**95% CI[1]**	**OR[2]**	**95% CI[2]**
FEV_1	0.87	0.84-0.90	0.87	0.84-0.90	1.00	0.96-1.04	0.99	0.95-1.03
FVC	0.94	0.91-0.98	0.95	0.91-0.99	1.00	0.96-1.04	0.99	0.95-1.03
FEV_1/FVC	0.83	0.81-0.86	0.83	0.80-0.87	1.00	0.96-1.04	1.01	0.97-1.05
PEF	0.89	0.86-0.92	0.88	0.85-0.92	1.03	0.99-1.07	1.03	0.99-1.07
MDD	1.13	1.09-1.17	1.07	1.03-1.12	1.27	1.22-1.32	1.24	1.19-1.30

[1]Model adjusted for age, sex, first 3 principal components (N=72,445)
[2]Model adjusted for age, sex, first 3 principal components, and ever smoking (N=53,503)
COPD: chronic obstructive pulmonary disease, FEV_1: forced expiratory volume in one second, FVC: forced vital capacity, MDD: major depressive disorder, PEF: peak expiratory flow; OR = odds ratio; CI = confidence interval

In addition to the expected phenotype associations, we observed a significant association between the MDD-PRS and COPD when adjusting for age, sex, and the first 3 PCs (OR=1.13; 95% CI: 1.09-1.17; *P* value $=3.72 \times 10^{-12}$) (Table 3-3, **Error! Reference source not found.**A, Appendix 1, Table 6-3). Adjusting for smoking attenuated the association and was no longer statistically significant (OR=1.09; 95% CI: 1.04-1.13; $p=8.07x10^{-5}$) (Table 3-3, **Error! Reference source not found.**B, Appendix 1, Table 6-4). Similar patterns were observed for both men and women in the sex-stratified analyses of PRS-MDD (Appendix 1, Figure 6-6 and Table 6-2, Table 6-9, Table 6-10). None of the lung function PRS were associated with MDD in the smoking adjusted or smoking unadjusted analyses (Table 3-3, Figure 3-1, Appendix 1, Table 6-5 to Table 6-8). Similarly, no significant associations between any of the lung

function PRS and MDD were observed in the sex-stratified analyses (Appendix 1, Figure 6-2 to Figure 6-5 and Table 6-2, Table 6-11 to Table 6-18).

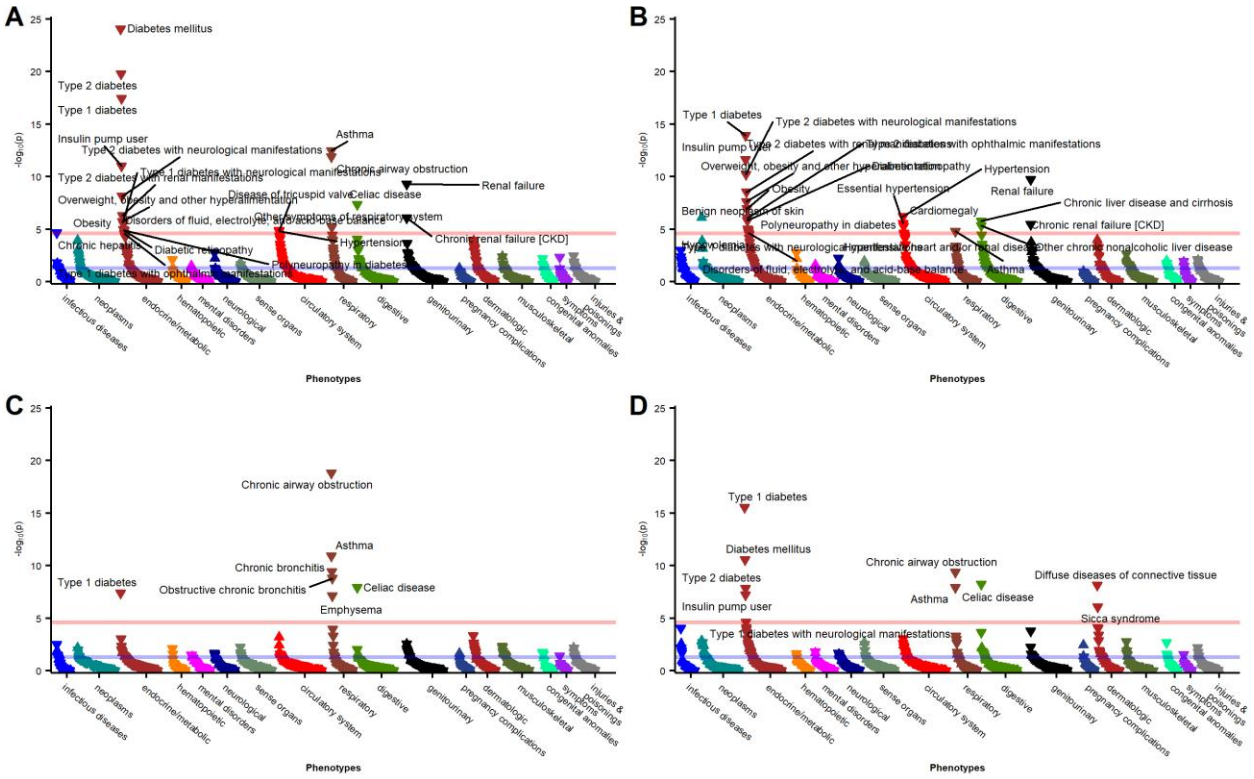

Figure 3-1. Phenome-wide association study among BioVU participants (2007-2019) of (a) FEV_1, (b) FVC, (c) FEV_1/FVC, and (d) PEF polygenic scores, adjusted for age, sex, first three principal components, and ever smoking.

3.3.4 Multi-trait Conditional Analysis to Detect Potential Pleiotropy

We used mtCOJO to adjust MDD for the genetic effects of FEV_1/FVC. The majority of SNPs showed little to no change in the effect estimate. The median percent change in the beta before and after conditioning was 0%, with an inter-quartile range of -6% to 5%. However, HEIDI-outlier identified three SNPs (rs12040241, rs7617480, rs12967855) with evidence of pleiotropy between MDD and FEV_1/FVC (Appendix 1, Table 6-19).

3.4 Discussion

We evaluated the potential for shared genetic architecture between lung function and MDD. We did not observe a significant global genetic correlation between lung function traits and MDD, consistent

with prior work.[72] In contrast, genetic correlations between lung function traits ranged from -0.08 to 0.87, similar to previous studies.[87] Local genetic correlation did identify a small but statistically significant increase in genetic correlation on chromosome 6 in the human leukocyte antigen (HLA) region. This finding is consistent with the known role of inflammation and the immune system in both COPD[366,367] and MDD[368,369]. We found that the PRS-MDD was significantly associated with COPD in our PheWAS, but this association was no longer statistically significant when controlling for smoking. Conversely, none of the lung function PRS showed a significant association with MDD in PheWAS analyses, suggesting little shared genetic architecture between lung function and MDD. However, using multi-trait conditional analysis, we identified three potentially pleiotropic SNPs. Interestingly, two of these SNPs were associated with both mood and smoking traits in a prior GWAS.[370–375] An intronic variant in *KLHDC8B,* rs7617480, was previously identified as genome-wide significant in GWAS of smoking cessation[370] and subjective well-being.[371] The second SNP, rs12967855, an intronic variant in *CELF4*, was previously found to have genome-wide significant associations with lifetime smoking index[372] and unipolar depression.[373–375]

While we identified three potentially pleiotropic variants, our findings do not provide strong evidence for a shared genetic architecture between MDD and COPD. Smoking behaviors may contribute to the relationship between MDD and COPD.[251,252,376] Cigarette smoking and nicotine dependence have been identified as potential confounding factors of the relationship between COPD and mood disorders,[376] and smoking may modify associations between COPD and depression.[250] Among individuals with COPD, current smokers report higher rates of depression symptoms and have increased mortality risks compared to former smokers and individuals without depression.[377,378] Previous studies have also shown that smokers with mental illness have higher mortality rates, particularly from respiratory conditions.[377,379–381] Further study is needed to understand the underlying mechanisms linking smoking, COPD, and MDD.(59–61)

This study has several strengths and considerations. We used available summary statistics from large, well-powered GWAS to conduct our analyses.[87,351,352] We also used the rich BioVU resource with

extensive clinical data allowing us to examine multiple phenotypes. Our study is limited by the inclusion of only European ancestry participants. PRS performance decreases in cross-ancestry analysis,[382,383] and the limited number of lung function GWAS that have been conducted in African Americans have had small sample sizes with few genome-wide significant findings.[86,98] Further research is needed to understand the genetic relationship between COPD and MDD in non-European descent populations. Another limitation of our study is the lack of a replication population to validate our findings. However, our findings are consistent with prior research.[72,87] Finally, our study relied on EHR data, which can present challenges due to data missingness and misclassification.[308,309,311–313] We chose to use phecodes to define phenotypes in our study, as previous research has demonstrated that phecodes better capture clinical disease than ICD codes alone.[200] For the majority of phenotypes, we expect the effects of misclassification to be minimal or biased toward the null.[384,385] We also encountered challenges due to missingness, particularly for smoking data (Table 3-1), which is prone to high rates of missingness and inaccuracies in EHRs.[386–389] Individuals who were missing smoking data were younger and had a lower prevalence of COPD than those with available smoking information (Table 6-20), thus relying on complete case analysis may limit the generalizability of our findings.

In conclusion, we found that the elevated prevalence of MDD in COPD cannot be solely explained by shared genetic risk factors. Our findings suggest a role for shared environmental or behavioral risk factors, such as smoking. We identified three potentially pleiotropic SNPs that can be prioritized in future studies of MDD and COPD. These findings require further investigation into the biological underpinnings between MDD and COPD to elucidate the causal mechanism underlying their relationship.

CHAPTER 4

Immunotherapy-Mediated Thyroid Dysfunction: Genetic Risk and Impact on Outcomes with PD-1 Blockade in Non-Small Cell Lung Cancer

4.1 Introduction

Immune checkpoint inhibitor (CPI) based therapy, which targets the adaptive immune system, is associated with remarkable long-term responses in a subset of cancers.[390–392] Benefit from CPI therapy is affected by tumor-specific features, such as PD-L1 expression[393,394] or tumor mutational burden[395] and host-specific features associated with underlying immunity such as the microbiome[396–398] and, possibly, the germline genetics of the host.[399–403]

The dual role of CPIs in promoting T cell activation but also autoimmunity leads to a variety of clinically significant systemic autoinflammatory responses (immune related adverse events, irAEs) in a subset of individuals.[404] The presentations of irAEs often mimic autoimmune conditions that occur spontaneously, but it is unclear if the underlying mechanism is shared or distinct despite the phenotypic similarity. Furthermore, it is uncertain whether and which irAEs are associated with CPI benefit. While some studies show an association between an irAE and improved outcomes,[405–408] others do not,[409,410] and some show worse outcomes.[411,412] These conflicting findings may be due to factors such as (a) survivor bias (patients who respond to therapy and are on therapy longer are more likely to develop irAEs) and (b) heterogeneous cohorts (combining cancer types, irAEs with different pathophysiology, ranges of presentation/ severity and different types of treatment of irAEs). We hypothesized that examining a specific irAE in a single cancer type that is routinely observed early in the CPI treatment course would yield clarity on the relationship between irAE development, immunity, and CPI benefit.

Thyroid irAEs occur early in the course of CPI exposure[413,414] and are among the most common irAE, with a cumulative incidence of approximately 10% of patients on PD-1 blockade therapy.[415,416] Spontaneous autoimmune hypothyroidism is common and appears to be at least partially heritable.[417,418] Genome-wide association studies (GWAS) have identified many common variants contributing to risk.[419]

It is unclear whether hypothyroidism that occurs following PD-1 blockade therapy is genetically similar to hypothyroidism that occurs in the general population. Additionally, whether this genetic risk impacts PD-1 blockade benefit is unknown.

To further understand the relationships between underlying autoimmunity, thyroid irAEs, and immunotherapy outcomes, we examined patients with non-small cell lung cancer (NSCLC) receiving PD-1 blockade-based therapy in cohorts from 3 academic medical centers. To investigate the genetics of thyroid irAEs, we first developed a genetic predictor of spontaneous hypothyroidism in non-cancer patients. We used the UK Biobank, an open access resource containing genetic and health related traits from ~500,000 volunteers in the UK and built a polygenic risk score[420] which integrates the information from many polymorphisms in the genome. We then validated that score in non-cancer patients in the Vanderbilt University biobank (BioVU). Finally, we evaluated whether genetic risk of sporadic hypothyroidism by GWAS is associated with incident hypothyroidism following PD-1 blockade and also evaluated the association between genetic risk and survival.

4.2 Methods

4.2.1 Patients

This retrospective study was approved by the institutional review board at Memorial Sloan Kettering (MSK), Vanderbilt University Medical Center (VUMC), and Dana Farber Cancer Institute (DFCI). All patients at MSK and DFCI provided written informed consent for blood banking. Patients in this study had advanced NSCLC and received PD-1 blockade-based therapy (either PD-1 blockade monotherapy or in combination with CTLA-4 blockade). In the MSK cohort, 551 patients who received CPIs between 2011-2018 and had an available baseline blood sample were included (Table 4-1). In the VUMC population, 195 individuals who had received CPIs between 2009-2019 were identified from BioVU, VUMC's DNA biobank linked to de-identified electronic health records.[291] At DFCI, 561 patients with NSCLC who received CPIs between 2013-2020 were examined.

For the germline genetic analysis, patients from MSK and VUMC (MSK and VUMC cohort) were analyzed together as they were both genotyped using the same GWAS array. Patients from DFCI were analyzed separately as an independent cohort.

Table 4-1. Baseline patient characteristics and treatment details.

Characteristic	**MSK (n = 551)**	**VUMC (n = 193)**	**DFCI (n = 561)**
Median age (IQR) – yr	67 (59-73)	63 (57-69)	67 (60-74)
Biological sex – no. (%)			
Female	291 (53)	74 (38)	313 (56)
Male	260 (47)	119 (62)	248 (44)
Race, self-reported – no./total no. (%)			
White	458/533 (86)	176/191 (92)	506/551 (92)
Black	39/533 (7)	11/191 (6)	22/551 (4)
Asian	34/533 (6)	2/191 (1)	17/551 (3)
Other	2/553 (<1)*	2/191 (1)	6/551 (1)
Unknown	18	2	10
Ethnicity, self-reported – no./total no. (%)			
Hispanic or Latino	18/545 (3)	2/186 (1)	10/561 (2)
Non-Hispanic or Latino	527/545 (96)*	184/186 (99)	551/561 (98)
Unknown	6	7	0
Cigarette smoking status – no./total no. (%)			
Former or current	472/551 (86)	166/188 (88)	363/421 (86)
Never	79/551 (14)	22/188 (12)	58/421 (14)
Unknown	0	5	140
Histology – no. (%)			
Adenocarcinoma	427 (78)	120 (62)	416 (74)
Non-adenocarcinoma	124 (22)	73 (38)	145 (26)
Treatment – no. (%)			
Anti-PD-(L)1 monotherapy	480 (87)	179 (93)	527 (94)
Anti-PD-(L)1 + CTLA-4 combination	71 (13)	14 (7)	34 (6)

*Percentages may not add up to 100 due to rounding. IQR = interquartile range, PD-(L)1 = programmed cell death protein (ligand) 1, CTLA-4 = cytotoxic T-lymphocyte-associated protein 4

4.2.2 Clinical Variables

MSK medical records and pharmacy records were reviewed for age, self-reported demographics, smoking status, lung cancer histology, past medical history, and CPI treatment history. These elements were abstracted and entered into a clinical data sheet. VUMC data were extracted using MedEx or from

structured tables within existing de-identified medical records.[421] All treatment dates underwent manual chart review by a trained thoracic oncology nurse, and data were entered into a REDCap database.

4.2.3 Thyroid irAE Event

A thyroid event after the start of CPI therapy was defined as either (1) incident hypothyroidism or (2) transient incident hyperthyroidism followed by incident hypothyroidism. Incident hypothyroidism was defined as (a) a TSH of ≥ 10 mU/L or (b) TSH of ≥ 5 mU/L with a new prescription of levothyroxine ≥ 50 mcg. Incident hyperthyroidism was defined as TSH < 0.05 mU/L. Individuals with incident hyperthyroidism without subsequent hypothyroidism (N = 2) was excluded from the analysis due to testing shortly before patients transitioned to hospice with no additional follow up thereafter. A baseline history of hypothyroidism or hyperthyroidism was defined as documentation of a thyroid condition prior to the start of CPI therapy. We excluded these patients from the analysis due to the challenge in defining whether a thyroid irAE occurred.

In the MSK cohort, medical records were manually reviewed to identify cases of thyroid events. Extracted laboratory and medication data were used to identify thyroid events in the VUMC population, with manual chart review confirmation. In the DFCI cohort, extracted laboratory and medication data were similarly used to identify thyroid events.

4.2.4 Response Assessment

In the MSK cohort, best overall response (BOR) was assessed by investigator-assessed Response Evaluation Criteria in Solid Tumors (RECIST) v1.1 from the start of CPI therapy to the date of tumor progression or death due to any cause. Progression-free survival (PFS) and overall survival (OS) was assessed from the date of start of CPI therapy.

Due to the nature of the VUMC cohort data, direct PFS data were not available. Time on treatment, defined as the date of the last dose of CPI minus the date of the first dose of CPI, was used as a

proxy for PFS. OS data was not available due to de-identification of data precluding linkage to the National Death Index (NDI).

In the DFCI cohort, overall survival (OS) was assessed from the date of start of CPI therapy, with linkage to the NDI up to 12/2019. Patients treated after the NDI check were censored at last contact at DFCI. The VUMC and DFCI cohorts did not have response assessment available.

4.2.5 Genotyping of MSK and VUMC Samples

Blood DNA from MSK and VUMC were genotyped on the Affymetrix Axiom Precision Medicine Diversity Array. Imputation was performed on the Michigan Imputation Server using the 1000 Genomes Phase 3 v5 reference panel. Standard quality control measures were implemented to remove poorly genotyped samples (call rate < 95%), SNPs (genotyping rate < 95%), and rare variants (minor allele frequency < 0.005) prior to imputation. After dropping samples with missing genotypes or low-quality genotypes, we had a total of N = 729 (N = 536 from MSKCC and N = 193 from VUMC) in the genetics dataset.

4.2.6 Genotype Imputation of DFCI Samples from Tumor Sequencing

Samples from DFCI had tumor panel sequencing as part of routine clinical care from which germline variants were imputed using off-target reads. OncoPanel, a custom targeted hybrid capture sequencing platform, was used to assay genomic variation from tumor biopsies. Germline variant imputation was performed across all samples from raw sequence reads using the STITCH imputation software, which leverages ultra-low coverage read data together with the 1000 Genomes Phase 3 v5 reference panel to infer probabilistic germline calls for the autosomal chromosomes. Quality control was performed to remove poorly imputed variants (INFO < 0.4) and rare variants (minor allele frequency < 0.01).

4.2.7 Polygenic Risk Assessment

Summary statistics were obtained from previous genome-wide association studies of 2 thyroid-related phenotypic outcomes in the UK Biobank - self-reported hypothyroidism (22,500 individuals, 436,818 controls)[422] and thyroid medication use (H03A medication category; 24,835 individuals, 280,750 controls).[423] Polygenic risk scores (PRS) were constructed using LDpred, which estimates posterior mean effect sizes using Bayesian priors and linkage disequilibrium information.[424] PRS weights will be made available at time of manuscript publication.

4.2.8 Quality control of PRS in DFCI Samples

A partially overlapping set of 833 patients with a variety of solid tumors seen at the DFCI had both OncoPanel next generation sequencing (as part of the same clinical cohort) and direct germline genotyping (as part of an orthogonal biobank). This set was used to validate the tumor imputed PRS. DNA samples were processed from whole blood and genotyped on the Illumina Multi-Ethnic Genotyping Array (MEGA), the Expanded Multi-Ethnic Genotyping Array (MEGA-Ex) array, and the Multi-Ethnic Global (MEG) BeadChip; imputed to the Haplotype Reference Consortium reference panel; and then restricted to ~1.1 million HapMap3 variants that typically exhibit high imputation accuracy across genotyping platforms. PRSs were inferred using the imputed variants that passed quality control in each respective study (i.e. no harmonization across the two platforms was imposed).

4.2.9 External Validation of Polygenic Risk Score

External validation of the PRS was performed in a population of 51,070 individuals of European descent with no cancer diagnosis in BioVU. European ancestry was determined using principal components analysis (PCA). Individuals in this cohort were genotyped on the Illumina MEGA-Ex array and subjected to standard quality control. Imputation was performed with the Haplotype Reference Consortium reference panel on the Michigan Imputation Server.[294] PRS were calculated using the previously derived weights from LDpred and the score function in PLINK 1.9.[296,350] Spontaneous hypothyroidism cases and controls were defined using phecodes, which aggregate similar ICD-9-CM and

ICD-10-CM. Individuals must have at least 2 ICD codes for hypothyroidism to be assigned a phecode, and individuals with other thyroid diseases were excluded from the control set.

4.2.10 Ancestry Analysis

We determined genetic ancestry using PCA in PLINK 1.9.[296,350] PCA was conducted in each genotype data separately: the BioVU dataset without cancer and CPI (MEGA-Ex array), the MSK and VUMC cohorts (Affymetrix PMDA array), and the DFCI dataset which included imputed SNPs from low coverage sequencing.

4.2.11 Statistical Analyses

De-identified data on duration of treatment and vital status at the time of the database lock on October 1, 2020 for the MSK cohort was used for the survival analysis. Patients who did not experience the event of interest at the database lock were censored at the time of the last follow up date. Survival analyses for PFS and OS were performed using Kaplan-Meier time-to-event estimates. We also determined the association between incident thyroid events and PFS and OS using multivariate Cox regression. We ran the Cox regression encoding thyroid events as a dichotomous variable counting any thyroid event during therapy. In addition, to account for potential survival bias, we also conducted analyses in which thyroid events were tested as time-dependent covariates. As a sensitivity analysis, landmark analysis in patients with OS of at least 90 days was performed in the MSK cohort. Ninety days was chosen as it includes more than half of patients with a thyroid irAE event (n=37/65, 57%). All models included age, sex, and concurrent use of anti-CTLA-4 therapy as covariates.

To test the association between prevalent hypothyroidism and PRS in BioVU (non-cancer patients), we used logistic regression models and adjusted for age at last visit, sex, and the first 10 principal components to account for genetic ancestry.

The association of PRS with incident hypothyroidism in the context of CPI therapy was tested using multivariate Cox regression with covariates for age, sex, and the first 10 principal components, and

hazard ratios (HRs) were computed per standard deviation of the PRS. In the DFCI cohort, additional technical covariates were included for sequencing panel version, whether the patient was sequenced after the start of therapy, and number of prior therapy lines.

Receiver operating characteristic (ROC) curves were used to visualize how well the PRS model discriminated hypothyroidism events and area under the ROC curve (AUROC) was estimated to quantify the overall prediction accuracy of the PRS. All p-values were two-sided. Statistical analysis and data visualization was performed using R v4.0.3 (R foundation for Statistical Computing, Vienna, Austria) with RStudio v1.3.1093.

A summary of the methods is illustrated in Appendix 2, Figure 6-7.

4.2.12 Individual Variant Testing

To better understand the shared biology between thyroid irAEs and spontaneous hypothyroidism, we analyzed the association with individual SNPs and thyroid irAEs. We focused on 16,751 SNPs that were noted to be genome-wide significant ($p<5x10^{-8}$) from the UK Biobank dataset in the GWAS for self-reported hypothyroidism since the UK Biobank is so well-powered. Of these, we filtered on minor allele frequency >0.01 and imputation quality $R^2>0.5$ in the MSKCC and VUMC GWAS dataset, leaving 16,132 SNPs. We tested the association of these SNPs with thyroid irAEs using Cox models, adjusted for age, sex, and ancestry as estimated by principal components (Appendix 2, Table 6-21). To determine the appropriate level for multiple hypothesis testing correction, we used the LD clump feature in PLINK and used a threshold of $R^2>0.5$ to identify the number of effectively independent regions tested, with a result of N = 1,057. Therefore, we used a multiple hypothesis testing correction of 0.05/1057 or 4.73×10^{-5} to account for multiple testing. All p-values were two-sided. To characterize potential function of a variant we used an online resource which includes distance from transcriptional start site, expression quantitative trait locus (eQTL) data and promoter Hi-C results (https://genetics.opentargets.org/)[425] and the GTEx consortium data (https://www.gtexportal.org/home/).[426]

4.3 Results

4.3.1 Study Population for Analysis

We identified 3 cohorts of patients with NSCLC who received CPI for our analysis: 551 individuals from MSK (MSK cohort), 193 individuals from VUMC (VUMC cohort), as well as 561 individuals from DFCI (DFCI cohort). The median follow-up time was 18.7 months (IQR: 7.5-44.4 months) for the MSK cohort, 11.1 months (IQR: 4.1-31.8 months) for the VUMC cohort, and 12.0 months (IQR: 4.3-26.2 months) for the DFCI cohort. Thyroid dysfunction, as defined by either hypothyroidism or hyperthyroidism with progression to hypothyroidism, occurred in 12% (n=65/551) of patients in the MSK cohort and was found to be an early event after CPI start, occurring within a median of 2.4 months (IQR: 1.4-4.2 months). Thyroid dysfunction occurred in 16% (n= 31/195, median of 4.1 months, IQR: 1.6-7.9 months) and 7% (n=42/561, median of 5.8 months, IQR: 4.8-13.1 months) of the patients in the VUMC and DFCI cohorts, respectively.

Age (median, interquartile range [IQR]: 67, 59-73 MSK; 63, 57-69 VUMC; 67, 60-74 DFCI), race and ethnicity, and smoking status (former or current: 86% MSK, 88% VUMC, 86% DFCI of patients where information is available) was similar across sites (Table 4-1). There were slight differences in gender (Female: 53% MSK, 38% VUMC, 56% DFCI). Adenocarcinoma was the most common histology (78% MSK, 62% VUMC, 74% DFCI) at the three sites, and most patients received anti-PD-(L)1 monotherapy (87% MSK, 93% VUMC, 94% DFCI). A small fraction of patients (13% at MSK, 7% at VUMC, 6% at DFCI) received anti-PD-(L)1 and anti-CTLA-4 combination therapy.

4.3.2 Thyroid irAEs Are an Early Event and Associated with Longer Survival

We first examined the association between thyroid irAEs and response in the MSK (BOR, PFS, and OS) and VUMC (PFS) cohorts, as these cohorts had manually abstracted outcome data. We did this by comparing patients with thyroid irAEs and those who neither had a history of hypothyroidism nor thyroid irAEs.

Table 4-2. Hazard ratios of the effect of thyroid irAEs on progression-free survival in the combined MSK+VUMC cohort. Models were adjusted for age, sex, and combined anti-PD-(L)1 + anti-CTLA-4 therapy.

	Progression-Free Survival		
Analysis	**aHR**	**95% CI**	**p-value**
Standard Cox regression analysis	0.42	0.33-0.54	$6x10^{-12}$
Time-dependent analysis	0.68	0.52-0.88	$4x10^{-3}$

aHR = adjusted hazard ratio

Compared to those without a thyroid event, individuals with thyroid irAEs had a higher objective response rate (MSK cohort: 38/65, 58% vs 110/486, 23%, p<0.0001 Fisher's exact) and longer PFS (combined MSK+VUMC cohort: 54% vs 19% progression-free at 1 year, 13% vs 3% progression-free at 3 years) and OS (MSK cohort: 85% vs 56% alive at 1 year, 35% vs 14% alive at 3 years) (Figure 4-1). The adjusted HR for PFS for thyroid irAEs in the combined cohort were 0.42 (95% CI: 0.33-0.54) and 0.68 (95% CI: 0.52-0.88) in our standard Cox and time-dependent analysis, respectively (Table 4-2). We found that 57% (n=37/65) of thyroid dysfunction events occurred within 90 days of CPI start in the MSK cohort. As a sensitivity analysis, we examined whether the same relationship held by only examining the patients who had OS longer than 90 days in the MSK cohort (86%, 475/551). Thyroid irAEs remained significantly associated with PFS (49% vs 20% at 1 year, 14% vs 4% at 3 years, HR: 0.50, 95% CI: 0.33-0.74) and OS (73% vs 62% alive at 1 year, 30% vs 15% alive at 3 years, HR: 0.59, 95% CI: 0.38-0.94) (Appendix 2, Figure 6-8). In individuals with available PD-L1 expression data (n = 320), thyroid irAEs remained significantly associated with PFS (HR: 0.39, 95% CI: 0.26-0.58) and OS (HR: 0.36, 95% CI: 0.22-0.58) when adjusting for PD-L1 expression.

Given this highly significant relationship between development of thyroid irAEs and multiple objective measures of response demonstrating clinical benefit of CPIs, we evaluated whether a hereditary predisposition to development of hypothyroidism could predict thyroid irAEs as well as CPI benefit.

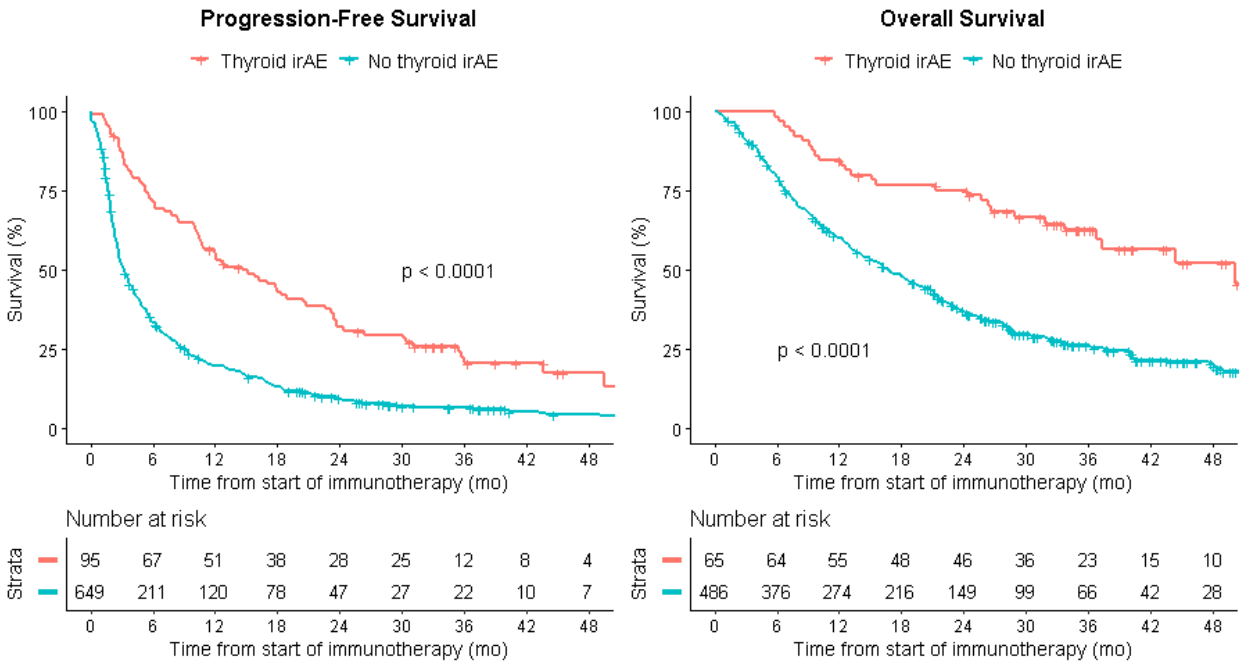

Figure 4-1. Thyroid irAEs as a predictor of PFS in the combined MSK+VUMC cohort and OS in the MSK cohort. Kaplan-Meier survival curves are unadjusted and compare those who had a thyroid irAE to those who did not have a thyroid irAE. The x-axis reflects time from start of CPI therapy.

4.3.3 Polygenic Risk Score for Thyroid Disorders Is Associated with Developing Thyroid irAEs

We next asked whether predisposition to a baseline thyroid condition is associated with developing thyroid irAEs. We developed a PRS for thyroid disorders using the UK Biobank. We used two different phenotypic proxies for thyroid disease: self-reported hypothyroidism and thyroid medication use and developed PRS for each of these using LDpred. The SNP weights for these PRSs were highly correlated (Pearson correlation = 0.9). We validated these PRSs using participants who were not included in our VUMC lung cancer and immunotherapy cohort but had genotypes available in the VUMC BioVU population (non-cancer patients). Since the UK Biobank is predominantly of European ancestry, we restricted the VUMC BioVU to those who were of European ancestry. Overall, there was a strong association of each individual PRS with hypothyroidism (AUROC = 0.6 for both) (Appendix 2, Figure 6-9A) with an odds ratio/standard deviation of 1.33 (95% CI 1.29-1.37) for the PRS for self-reported hypothyroidism and an increased odds ratio by decile (Appendix 2, Figure 6-9B). We then applied these

PRSs to our discovery cohort. Since the effect sizes for these 2 PRSs were highly correlated, we focused primarily on the PRS for self-reported hypothyroidism (hypothyroidism PRS) from the UK Biobank.

The hypothyroidism PRS was significantly associated with development of thyroid irAEs in the MSK+VUMC cohort (Figure 4-2A), and it had similar performance for predicting thyroid irAEs (AUROC = 0.6) in comparison with its performance in the VUMC BioVU cohort (spontaneous hypothyroidism). The thyroid medication PRS from UK Biobank had similar performance (Appendix 2, Figure 6-10). In a Cox regression model adjusted for age, sex, and the first ten principal components, the HR per standard deviation for the hypothyroidism PRS was 1.34 (95% CI: 1.08-1.66; $p = 8.73 \times 10^{-3}$), with similar effect sizes seen for the other PRS (Table 4-3). The hypothyroidism PRS results remained significant after removing individuals who received combination CPI therapy (HR: 1.34, 95% CI: 1.07-1.69, p = 0.01) and only including individuals of European ancestry (HR: 1.27, 95% CI: 1.02-1.59, p = 0.03). Additionally, when examining the PRS by tertile, individuals in the highest tertile PRS scores had higher rates of hypothyroidism events than the first and second tertile (Figure 4-2A; Appendix 2, Figure 6-10).

We sought to replicate these PRS associations in the DFCI cohort. First, we confirmed that the PRS was imputed accurately from tumors using a separate set of 833 benchmark individuals with both tumor sequencing and germline SNP array data available (see Methods): the correlation between the tumor imputed PRS and the germline ground truth PRS was >0.87 for all PRSs evaluated, with no visible outliers (Appendix 2, Figure 6-11). Next, we tested each PRS for association with the time to thyroid irAE: both PRSs were significantly associated, with the PRS for self-reported hypothyroidism yielding an HR of 1.39 (95% CI 1.07-1.82; p=0.01 by Cox regression) (AUROC=0.64) with similar effect sizes seen for the other PRS (Figure 4-2B; Appendix 2, Figure 6-12).

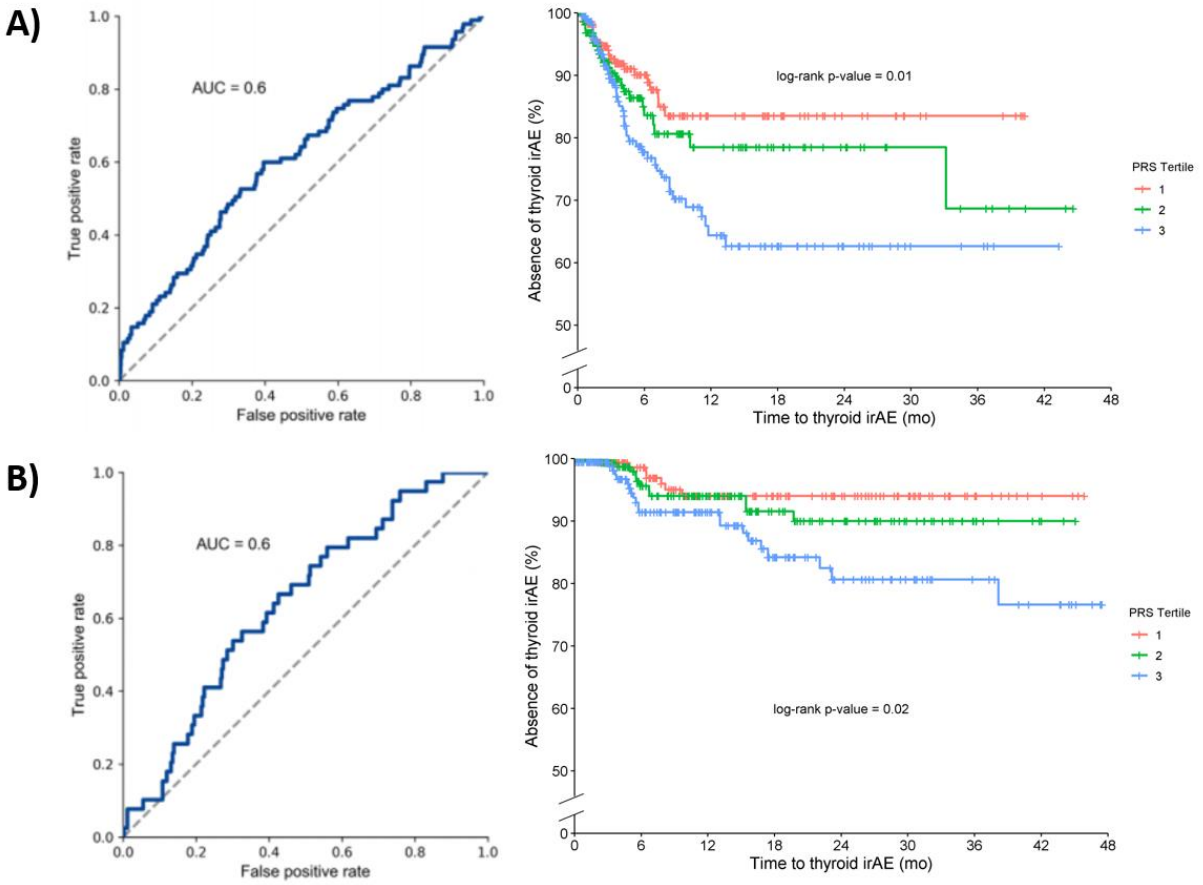

Figure 4-2. Hypothyroidism PRS (using self-reported hypothyroidism) as a predictor of CPI-induced thyroid irAEs in the (A) MSK+VUMC cohort and (B) the DFCI cohort. The left panel shows the ROC curve and right panel shows time to event by PRS tertile. P-values for the three curves are calculated using a log-rank test.

Table 4-3. PRS as a predictor of CPI-induced thyroid irAEs in the MSK+VUMC cohort. HRs of the effect of PRS as a predictor of thyroid irAEs in the combined MSK+VUMC cohort. Cox regression model was adjusted for age, sex, and the first ten principal components.

PRS Phenotype	aHR	95% CI	p-value
Hypothyroidism	1.34	1.08-1.66	8.73×10^{-3}
Thyroid medications	1.32	1.07-1.63	9.98×10^{-3}

aHR = adjusted hazard ratio

4.3.4 Analysis of individual loci associated with thyroid irAEs

As an exploratory analysis, we evaluated whether any of the genome-wide significant associations identified in the UK Biobank for self-reported hypothyroidism are also individually associated with thyroid irAEs. After filtering on minor allele frequency and imputation quality, we tested each of 16,132 significant UK Biobank SNPs for association with hypothyroidism irAEs in the MSK+VUMC cohort (Appendix 2, Table 6-21). Of these, 1,502 of were nominally associated ($p<0.05$) with thyroid irAEs and exhibited significant sign consistency (1,143/1,502 were associated in the same direction) underscoring the shared genetic effect observed through the PRS. One SNP, rs9268543 ($p=7.5x10^{-7}$), surpassed stringent Bonferroni correction, even though the MSK+VUMC cohort was orders of magnitude smaller than the UK Biobank and thus underpowered for individual variant discovery. This SNP falls within the HLA locus and has been associated with numerous other autoimmune traits.[427,428] This variant is also an expression quantitative trait locus (eQTL) for several genes at this locus, most strongly for HLA-DQA2.[426]

4.3.5 PRS for Hypothyroidism Is Not Associated with PFS or OS

Lastly, we assessed whether PFS for hypothyroidism was associated with CPI response. Despite the association between both PRSs and thyroid irAEs, neither PRS was significantly associated with PFS or OS in the MSK cohort (Appendix 2, Table 6-22; Appendix 2, Figure 6-13A). Likewise, in the DFCI cohort no significant association between PRS and OS was observed (Appendix 2, Figure 6-13B).

4.4 Discussion

We examined the associations between hereditary predisposition for spontaneous hypothyroidism, thyroid irAEs, and benefit from CPI treatment with the goal of inferring whether there may be shared biology. Thyroid irAEs occurred early during treatment, and those who developed this irAE were more likely to have a beneficial initial response and to achieve a longer duration of response to

CPI therapy compared to those who did not. We developed a PRS for hypothyroidism that performed similarly in predicting development of spontaneous hypothyroidism in the general population and thyroid irAEs in multiple patient cohorts. However, the PRS did not predict benefit from immunotherapy.

We saw a rapid onset of thyroid irAEs in the MSK and VUMC datasets. In the DFCI dataset, thyroid events occurred on average later, but this could have been due to the later and less frequent TSH measurements in this cohort and despite this the PRS performed similarly in the DFCI cohort. The rapid onset of thyroid irAEs in the MSK dataset and VUMC datasets in a genetically predisposed population and our previous work on finding shared auto-antibodies[413] suggests that the checkpoint inhibitors unmasked a pre-existing subclinical autoimmune condition suppressed by immune checkpoints.

A potential limitation of our study is survivor bias, or time from CPI start to development of thyroid irAEs. However, this irAE consistently occurred early after CPI start and survivor bias was adjusted for in time-dependent and landmark analyses. We found thyroid irAEs significantly associated with multiple markers of response: BOR, PFS, and OS. Additionally there is plausibility that tipping immunologic homeostasis of immune tolerance with CPI therapy not only leads to indirect cancer cell cytotoxicity but also potentiates (or drives) development of anti-thyroid antibodies leading to thyroid irAEs. Therefore, we infer there exists a shared immunologic mechanism that drives both thyroid irAEs and CPI benefit leading to improved survival. Furthermore, there are other cancer therapy settings in which development of thyroid dysfunction is associated with response to treatment (e.g. IL-2).[429]

Another limitation of our analyses is that our PRS was developed in UK Biobank which includes participants who are predominantly of European ancestry. The cohorts of CPI-treated patients we tested were also predominantly of European ancestry. Since PRS may not generalize well across different ancestry populations[301] the PRS we tested may not work to predict irAEs in non-European ancestry populations. Larger studies of patients on CPI in non-European ancestry populations are needed to understand the genetics of treatment benefit and irAEs in these populations.

The underlying factor leading to thyroid irAEs and CPI benefit may have a hereditary origin, an environmental origin, or may be a combination of multiple factors. Our PRS developed for spontaneous

hypothyroidism, a diagnosis that generally has an autoimmune etiology, was able to perform similarly well in predicting thyroid irAEs. This supports a shared biological mechanism encapsulated by the PRS driving autoimmune hypothyroidism and thyroid irAEs. In short, this suggests there exists genetic polymorphism-based hereditary factors that contribute to developing thyroid irAEs. Our results suggest that other irAEs may be predictable using analogous PRS for similar clinical phenotypes.

Our analyses may also be useful to help understand the mechanisms that underlie thyroid irAEs. To examine shared loci, we focused on the large, well-powered UK Biobank as the discovery GWAS, and investigated which SNPs replicated in our MSK+VUMC cohort. We found that one loci was associated with thyroid irAEs after stringent multiple test correction. The HLA locus with the top variant (rs9268543) was previously associated with many autoimmune diseases, including rheumatoid arthritis,[427] inflammatory bowel disease[428] and with hypothyroidism and type 1 diabetes in the UK Biobank. The overlap between HLA genetic variants for spontaneous hypothyroidism and thyroid irAEs suggests that at least some of the antigens underlying the disorder overlap. This variant is an attractive target for downstream experimental analysis to understand the mechanisms of irAEs. Future studies with larger sample sizes may help to identify additional shared loci between spontaneous hypothyroidism and thyroid irAEs and possibly to identify new loci at which the genetics do not overlap.

Our hypothyroidism PRS did not distinguish between those benefiting and not benefiting from CPIs. This suggests that in contrast to inherited factors, metabolic, epigenetic or environmental factors (e.g. shared exogenous exposures among those who benefit/do not benefit from CPI that shape the adaptome over time) play a larger role in driving the mechanism behind thyroid irAEs and CPI benefit. Additionally, the present analysis may be underpowered to detect the difference. Since the PRS for hypothyroidism had an AUROC of ~0.6, larger cohort sizes may be needed to overcome the heterogeneity of environmental factors. The association between thyroid irAEs and treatment response could also be due to inherited genetic factors that are not captured by the PRS. The PRS developed from UK Biobank was developed exclusively from common genetic variants and does not capture the effect of rare genetic variants. Since rare variants may contribute a large fraction of heritability for some complex

traits,[430,431] future studies using whole genome sequencing data may reveal a relationship between heritable genetic variation for thyroid irAEs and CPI benefit. An alternative explanation may be that the genetic contribution captured by the PRS may solely reflect a "branch" off target effect of CPI therapy (e.g. cross presentation of shared antigens that are not associated with CPI benefit) decoupled from the adaptive immune mechanisms of CPI therapy that lead to improved outcomes. If the latter explanation is correct, it would suggest that the genetic component of an individual's cancer-immune set point may not be the same factors that confer autoimmune thyroid disease risk.

Individual irAEs have distinct disease kinetics that differentially affect survival bias and varied severity ranging from subclinical to life-threatening[404] - unsurprisingly, not all are consistently associated with CPI benefit.[409,432,433] Our finding that development of thyroid irAEs secondary to PD-1 blockade-based therapy is associated with CPI benefit is similar to what has been seen in the literature.[434] Thyroid irAEs occur early, are common, and can be treated with thyroid hormone supplementation, making it an early clinical signal suggestive of long-term benefit from CPI therapy.

Genetic predisposition to irAEs has previously been examined in studies associating irAE risk with human leukocyte antigen (HLA) genes[435–438] and/or individual SNPs.[439] Only one other study has examined these associations using PRSs, which is more reflective of underlying susceptibility to autoimmune disease.[440,441] Our findings differ from this study of genetics of autoimmune skin conditions and survival.[442] Khan et al. developed a PRS of autoimmune dermatologic conditions and applied the PRSs to a clinical trial of PD-1 blockade in bladder cancer. Similar to our results, they found an association between the autoimmune PRSs and irAE. In contrast to our findings where hypothyroid PRSs were not associated with CPI benefit, dermatologic PRSs were associated with overall survival in patients receiving CPI therapy.[442] This suggests predisposition for specific autoimmune diseases differentially impacts immunotherapy benefit.

In summary, in this study of patients from three academic centers, we found that developing thyroid irAEs was robustly associated with benefit from CPI therapy. Genetic predisposition measured by PRSs for spontaneous - immune-mediated - hypothyroidism was associated with the irAE but did not

predict CPI benefit, suggesting distinct immune pathways driving underlying genetic risk for hypothyroidism and benefit from CPI. Large scale and mechanistic studies are needed to elucidate underlying pathways linking genetic risk, specific irAEs, and therapeutic response. Understanding the relationships and biological underpinnings of these processes are critical for both advancement of precision immunotherapy and development of new therapies for our patients.

CHAPTER 5

Conclusion and Future Directions

5.1 Summary of Findings

The goal of this dissertation was to utilize genetic techniques to understand the relationships between lung cancer, COPD, and comorbid conditions using DNA biobanks linked to electronic health records. The three aims were as follows:

1. Develop a phenotyping algorithm to identify COPD cases in EHR (Chapter 2)
2. Examine the genetic relationship between COPD and MDD within an EHR (Chapter 3)
3. Investigate genetic predictors of thyroid irAEs in lung cancer patients treated with immunotherapy in an EHR (Chapter 4)

In the first aim, we used the Synthetic Derivative, a de-identified version of Vanderbilt University Medical Center's EHR, to develop two phenotyping algorithms for COPD. The code-only algorithm used ICD codes to define cases and controls, while the code+keyword algorithm included text from the problem list in addition to ICD codes to define cases. Both algorithms have unique advantages and disadvantages. The code-only algorithm relied solely on structured data, making it easy to implement. In addition, ICD codes are widely used across the world, making the algorithm easily portable between EHR systems.[443] However, the code-only algorithm introduced the potential for more misclassification, as the specificity (71.9%) and PPV (60.9%) in the validation set were lower than the code+keyword algorithm which displayed improved specificity (94.1%) and PPV (88.6%). The more stringent criteria reduced the number of COPD cases identified by the algorithm in the Synthetic Derivative from 28,520 with the code-only algorithm to 12,622 cases. Furthermore, the use of "oxygen" as a keyword inherently biased our cases to those with more severe disease (Figure 2-3). The choice of which algorithm to use will ultimately depend on the setting, as the balance between sample size, ease of implementation, and potential misclassification will vary across research applications.

The second aim focused on the relationship between COPD and MDD. We used summary statistics from previous GWAS to evaluate the genetic correlation between these often comorbid conditions and found the genetic correlation to be low. We leveraged BioVU, a DNA biobank linked to the Vanderbilt Synthetic Derivative, to develop PRSs for both lung function and MDD and to perform PheWAS. Although the MDD-PRS was significantly associated with COPD, the strength of the association was attenuated when controlling for smoking. Furthermore, the lung function PRS were not associated with MDD. Overall, our findings suggest that the relationship between COPD and MDD is not due to shared genetic risk factors. The influence of smoking on our findings supports further research into the relationship between smoking, COPD, and depression. In addition to being a major risk factor for COPD, smoking has been found to be independently associated with MDD.[250–252] By demonstrating a weak genetic relationship between COPD and MDD, our findings encourage research into other explanations for the comorbidity between the two traits, which may ultimately help determine better diagnostic or treatment methods for individuals with both COPD and MDD.

In the final aim, we explored the role of genetics in the development of irAEs in individuals with lung cancer treated with PD-1/PD-L1 immunotherapy with or without combined CTLA-4 immunotherapy. We focused on hypothyroid events, as these are one of the most common irAEs,[415,444] typically occur early after treatment,[413–415] and can be easily diagnosed with routine lab tests.[175,445] In our cohort, hypothyroid irAEs were associated with improved progression-free and overall survival, consistent with previous studies on the positive prognostic impact of irAEs.[174,176–178] We built a PRS using summary statistics from a GWAS of hypothyroidism in the general population, and we found that the PRS was significantly associated with the development of a thyroid irAE. This suggests that the development of thyroid irAEs is at least in part due to an underlying increased genetic risk for hypothyroidism. However, our PRS was not associated with progression-free or overall survival, indicating that the positive prognostic impact of thyroid irAEs is not solely driven by genetic risk factors. Overall, these findings provide valuable insight into the biology of thyroid irAEs in NSCLC individuals treated with immunotherapy.

5.2 Limitations

One major limitation of our study is the lack of population diversity. Our analyses were primarily limited to individuals of European descent. Poor representation of non-European populations is a widespread problem in biomedical and genetics research.[223,224,446] Due to the demographics of the Vanderbilt University Medical Center patient population, our non-European sample sizes are small, limiting our ability to perform robust statistical analyses in these populations. Research in diverse populations is a priority in biomedical research. Racial and ethnic disparities have been identified across a range of clinical phenotypes and health outcomes.[447,448] Differences in allele frequencies between ancestry groups can have important ramifications for clinical practice, as variants associated with severe drug reactions have been identified at high frequencies in certain genetic ancestries.[223,224] Linkage disequilibrium patterns also vary between populations. In GWAS, variants identified as significant are not necessarily the causal SNP, but rather may be SNPs in high linkage disequilibrium with the causal SNP. This can lead to failure to replicate findings in other ancestry populations where the strength of the linkage between the tag SNP and the causal SNP is not as strong.[224] Gene-gene interactions and gene-environment interactions may also vary across ancestries, leading to differences in genotype-phenotype associations.[224] While our findings provide important insight into the genetic relationships between COPD, lung cancer, and comorbid traits, the generalizability of these findings across populations is unknown.

A second limitation of this project is the reliance on EHR data. Inconsistencies in documentation, data missingness, and inaccuracies in the clinical record pose challenges for EHR research.[305–313] ICD codes were designed for billing purposes, so they are not always the most accurate tools for secondary research use.[316–318] In addition, tertiary care centers such as Vanderbilt University Medical Center often contain denser clinical data for individuals with more severe disease, which can bias findings.[313] To address these issues, our code-only COPD phenotyping algorithm required the presence of multiple ICD

codes, and our code+keyword algorithm included information beyond ICD codes, which has been shown to improve phenotyping accuracy.[316,318] For both algorithms, we demonstrated that the association between COPD case status and epidemiologic and genetic risk factors was consistent with previous research. Furthermore, we implemented a medical home definition to help mitigate differences in documentation density between our case and control sets. In our PRS and PheWAS analyses, we relied on phecodes for phenotype definition, which have been shown to be more accurate than ICD codes alone.[449] For our study of irAEs in NSCLC, all charts underwent clinical review to ensure accuracy of diagnoses and treatment dates. Overall, we implemented several strategies to minimize the impact of EHR misclassification on our findings.

5.3 Future Directions

The research presented in this dissertation expands our understanding of the genetic relationships between COPD, lung cancer, and comorbid conditions and presents opportunities for future research. The phenotyping algorithm we developed in the first aim enables a wide range of research applications, including epidemiologic, genetic, and clinical research opportunities for COPD studies. EHRs have been identified as a valuable tool for addressing gaps in COPD research,[219,220] and our phenotyping algorithm will empower continued research on COPD. Findings in our second and third aim provide a foundation for additional research on the relationship between COPD and MDD, such as possible role of systemic inflammation, hypoxemia or oxidative stress, smoking, and other environmental risk factors.[238,248–250] Our finding that a hypothyroidism PRS predicts thyroid irAEs but not improve survival in NSCLC warrants further study to identify other factors that may be involved in the survival benefit associated with thyroid irAEs. Finally, replication of our findings in diverse populations is necessary to identify potential differences in the genetic relationships between COPD and MDD and genetic risk for thyroid irAEs in non-European ancestry populations.

Overall, this project contributes important knowledge to the fields of COPD and lung cancer research. COPD and lung cancer are major contributors to morbidity and mortality worldwide. By leveraging genetic tools, we were able to improve our understanding of these diseases, which will ultimately lead to better diagnostic, treatment, and prevention strategies.

CHAPTER 6

Appendices

6.1 Appendix 1

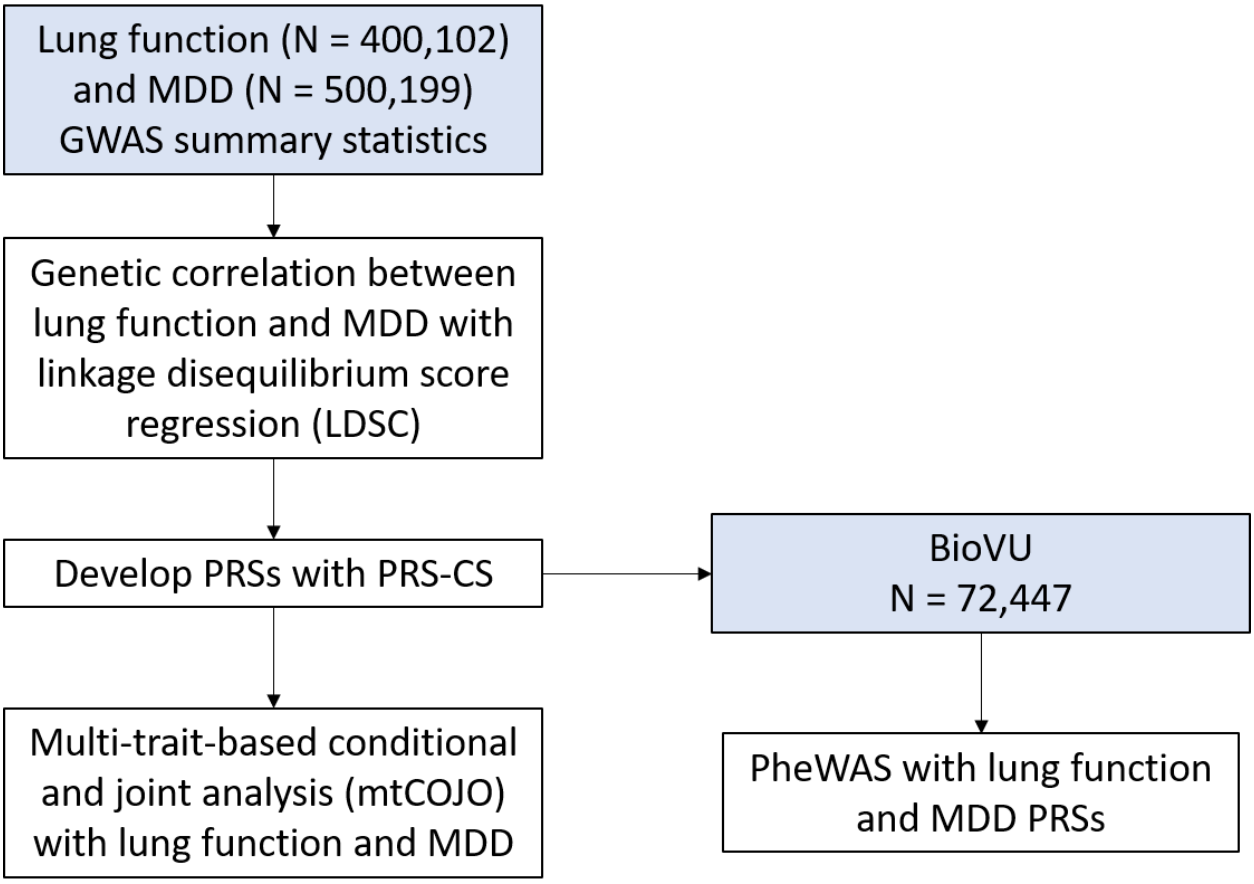

Figure 6-1. Overall study design.

MDD: major depressive disorder, GWAS: genome-wide association study, PRSs: polygenic risk scores, PRS-CS: polygenic risk score-continuous shrinkage, PheWAS: phenome-wide association study

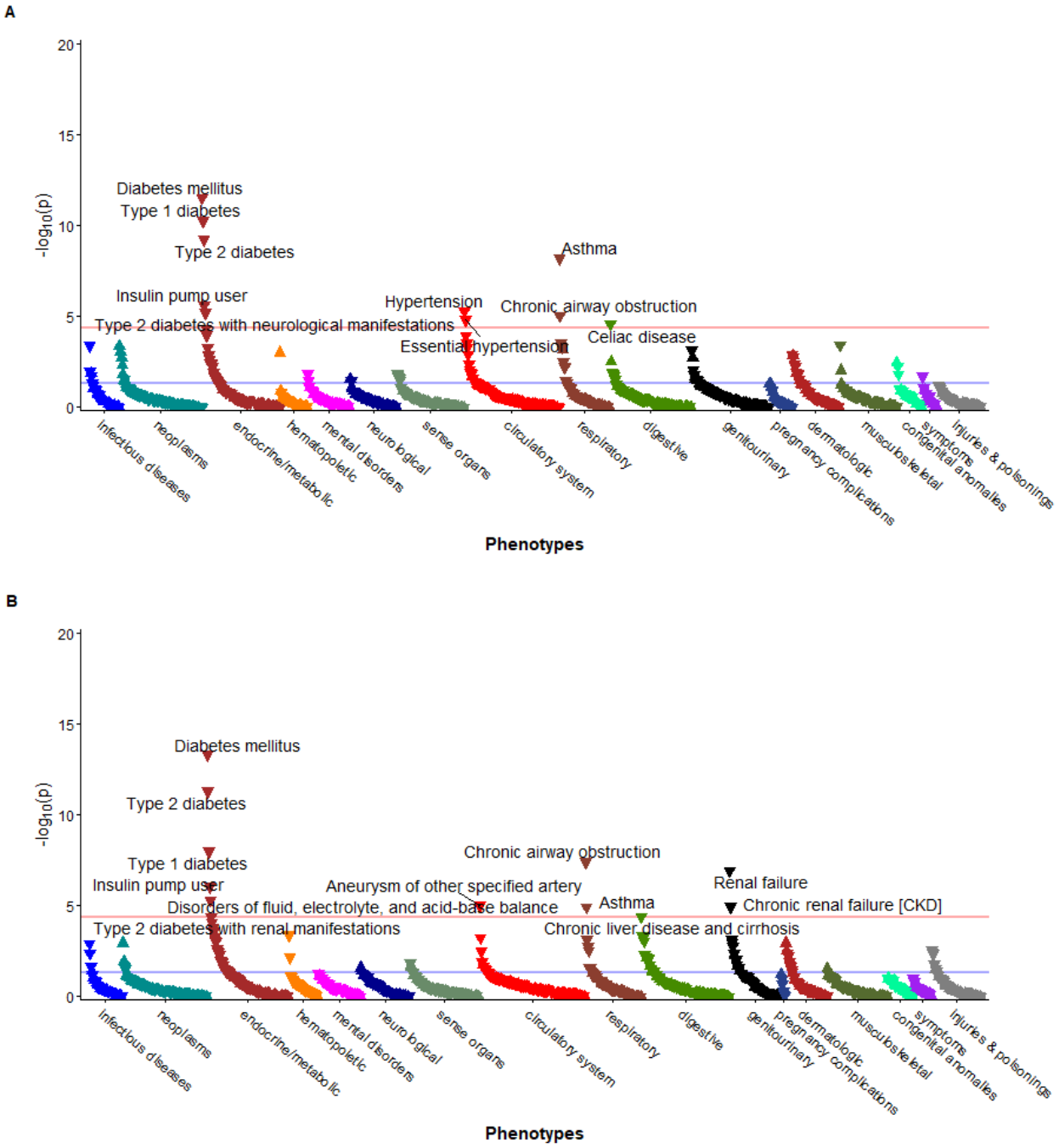

Figure 6-2. Sex-stratified phenome-wide association study results for FEV_1 PRS in A) women and B) men in BioVU, adjusted for age, first 3 principal components, and ever smoking.

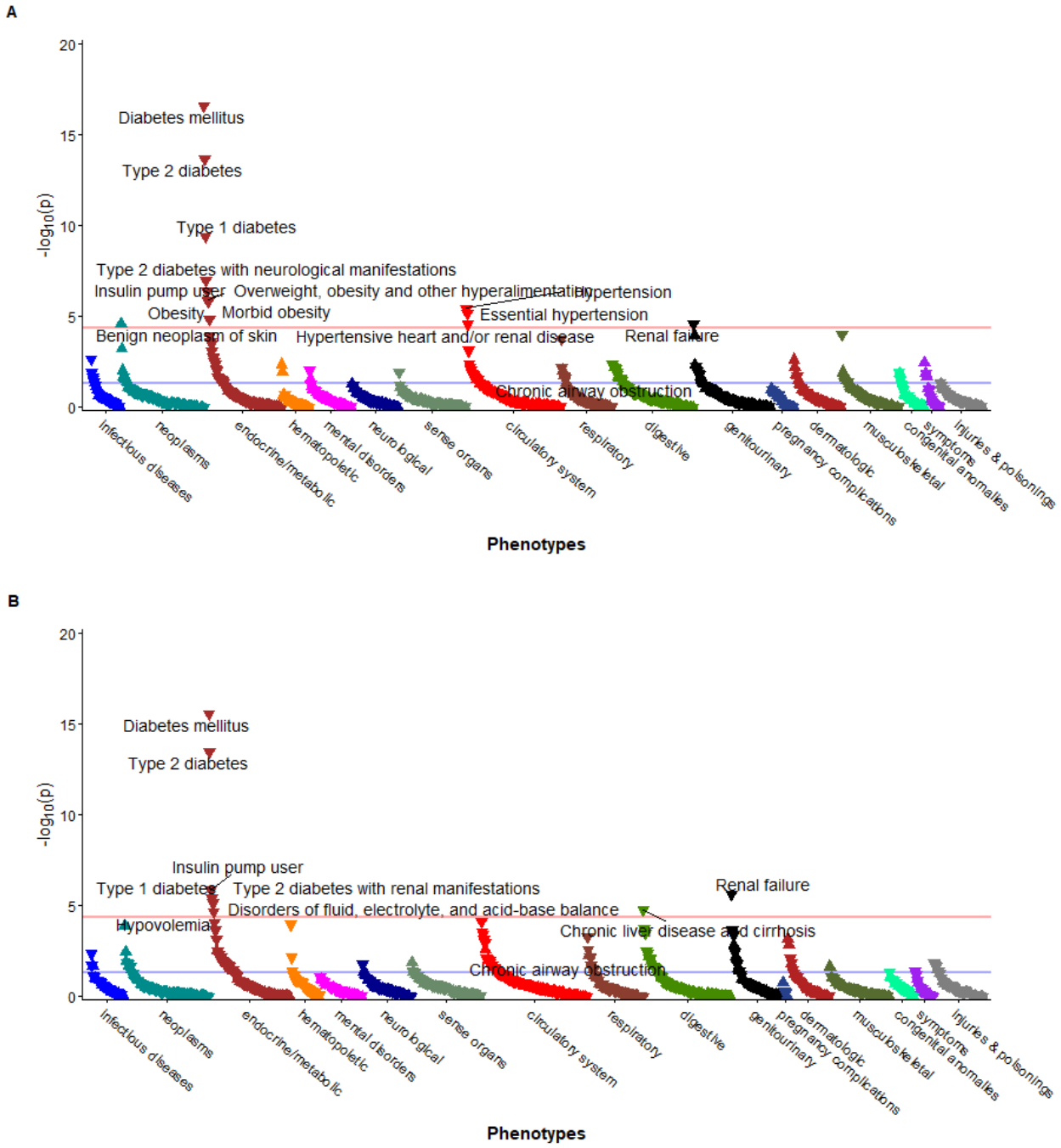

Figure 6-3. Sex-stratified phenome-wide association study results for FVC PRS in A) women and B) men in BioVU, adjusted for age, first 3 principal components, and ever smoking.

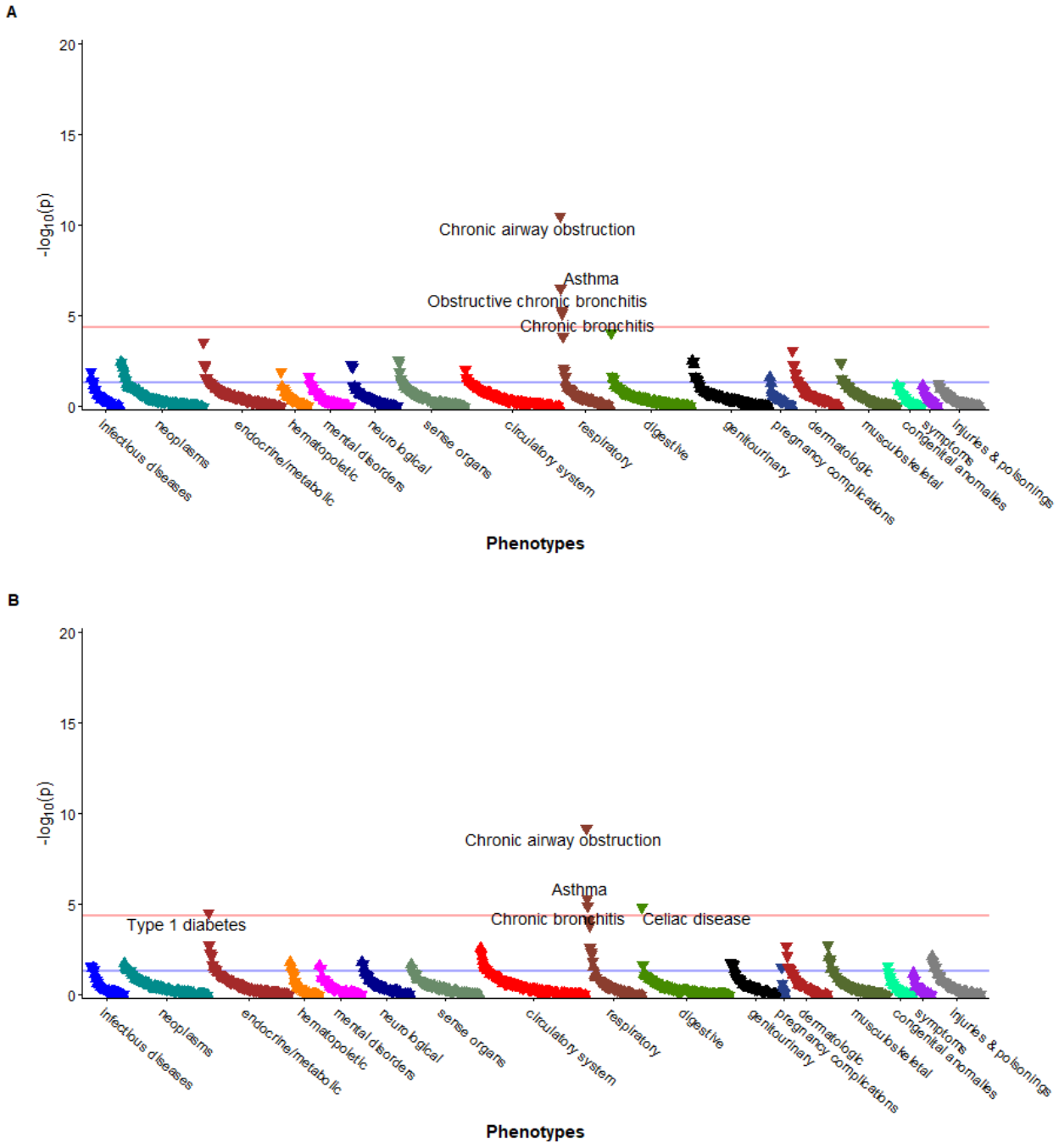

Figure 6-4. Sex-stratified phenome-wide association study results for FEV1/FVC PRS in A) women and B) men in BioVU, adjusted for age, first 3 principal components, and ever smoking.

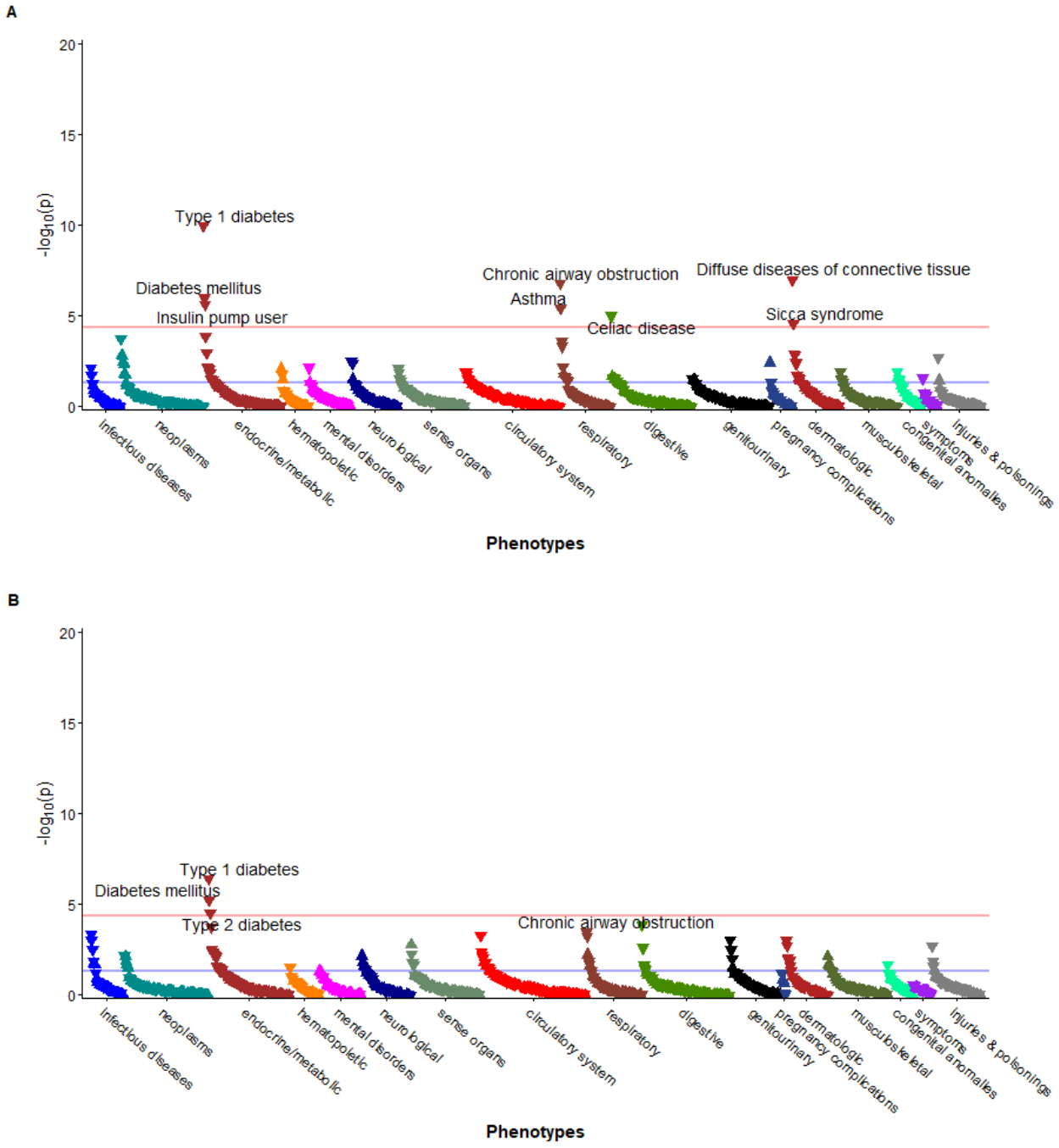

Figure 6-5. Sex-stratified phenome-wide association study results for PEF PRS in A) women and B) men in BioVU, adjusted for age, first 3 principal components, and ever smoking.

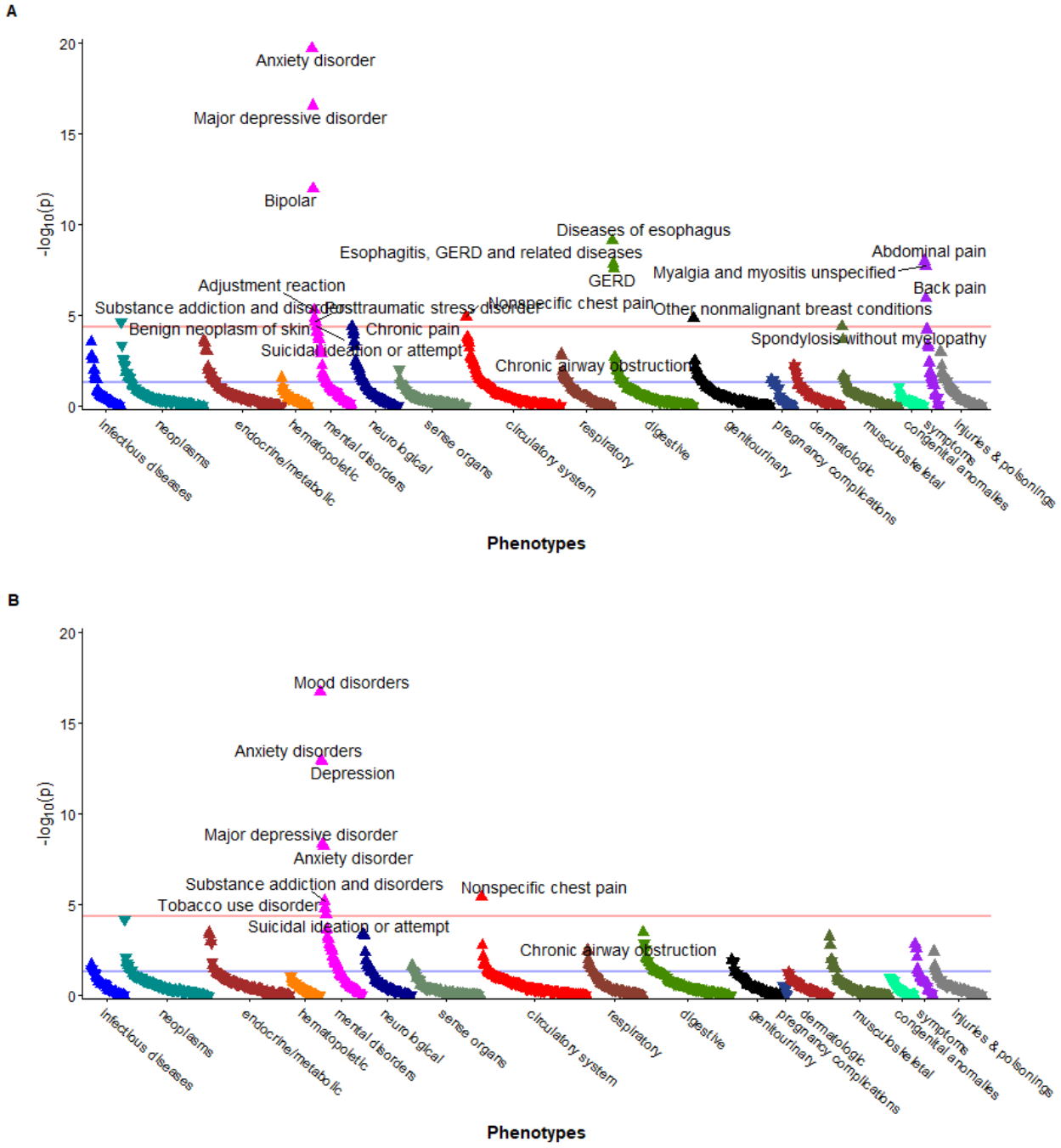

Figure 6-6. Sex-stratified phenome-wide association study results for MDD PRS in A) women and B) men in BioVU, adjusted for age, first 3 principal components, and ever smoking.

Table 6-1. Association of polygenic risk scores with pulmonary function measures.

Phenotype	Pre-bronchodilator PFT			Post-bronchodilator PFT		
	Beta	SE	P	Beta	SE	P
FEV_1	0.10	0.01	$7.71x10^{-15}$	0.08	0.02	$1.08x10^{-4}$
FVC	0.10	0.02	$2.55x10^{-10}$	0.08	0.02	$3.21x10^{-4}$
FEV_1/FVC	2.58	0.19	$<2x10^{-16}$	2.64	0.36	$5.52x10^{-13}$
PEF	0.25	0.04	$1.37x10^{-11}$	0.19	0.06	$2.23x10^{-3}$

Association tests were performed with the PFT corresponding to the PRS phenotype (eg, FEV_1-PRS was tested for associations with pre- and post-bronchodilator FEV_1). Analyses were adjusted for age at last visit, sex, ever smoking, and the first 3 PCs.

Table 6-2. Sex-stratified phenome-wide association between MDD and lung function PRS and phenotypes of interest among BioVU participants (2007-2019).

	COPD				MDD			
PRS	**OR[1]**	**95% CI[1]**	**OR[2]**	**95% CI[2]**	**OR[1]**	**95% CI[1]**	**OR[2]**	**95% CI[2]**
Women								
FEV_1	0.88	0.83-0.92	0.88	0.83-0.93	1.01	0.97-1.06	1.01	0.96-1.06
FVC	0.96	0.92-1.01	0.98	0.92-1.04	1.01	0.97-1.06	1.01	0.96-1.06
FEV_1/FVC	0.83	0.79-0.87	0.82	0.78-0.87	1.00	0.95-1.05	1.01	0.96-1.06
PEF	0.87	0.83-0.92	0.86	0.81-0.91	1.03	0.99-1.06	1.04	0.98-1.09
MDD	1.13	1.08-1.19	1.09	1.02-1.15	1.27	1.21-1.33	1.27	1.19-1.35
Men								
FEV_1	0.86	0.82-0.90	0.86	0.82-0.91	0.97	0.91-1.04	0.95	0.89-1.02
FVC	0.93	0.88-0.97	0.93	0.88-0.98	0.97	0.91-1.04	0.95	0.89-1.02
FEV_1/FVC	0.84	0.80-0.88	0.84	0.80-0.89	1.00	0.94-1.07	1.01	0.94-1.08
PEF	0.91	0.87-0.95	0.91	0.86-0.96	1.02	0.96-1.09	1.01	0.95-1.09
MDD	1.13	1.08-1.19	1.08	1.02-1.15	1.26	1.18-1.35	1.21	1.12-1.32

[1]Model adjusted for age, first 3 principal components (Women, N = 40,584; Men, N = 31,861)
[2]Model adjusted for age, first 3 principal components, and ever smoking (Women, N = 30,515; COPD and MDD men, N = 22,988)

COPD: chronic obstructive pulmonary disease, FEV_1: forced expiratory volume in one second, FVC: forced vital capacity, MDD: major depressive disorder, PEF: peak expiratory flow; OR = odds ratio; CI = confidence interval

Table 6-3. Phenome-wide association results passing Bonferroni significance for MDD-PRS, adjusted for age at last visit, sex, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
296	1.28	1.25	1.32	6.42E-76	6591	44955	Mood disorders	mental disorders
300	1.25	1.22	1.29	1.06E-57	5909	44955	Anxiety disorders	mental disorders
296.2	1.28	1.24	1.32	1.32E-55	4771	44955	Depression	mental disorders
300.1	1.23	1.19	1.27	3.04E-35	4158	44955	Anxiety disorder	mental disorders
296.22	1.27	1.22	1.32	1.41E-31	2690	44955	Major depressive disorder	mental disorders
296.1	1.32	1.25	1.40	3.83E-22	1280	44955	Bipolar	mental disorders
316	1.27	1.21	1.34	1.17E-20	1563	56257	Substance addiction and disorders	mental disorders
418	1.11	1.09	1.14	1.13E-18	9077	47707	Nonspecific chest pain	circulatory system
318	1.19	1.14	1.23	2.23E-17	2613	56257	Tobacco use disorder	mental disorders
785	1.09	1.06	1.11	2.28E-14	11096	45613	Abdominal pain	symptoms
760	1.11	1.08	1.14	5.12E-13	6119	53631	Back pain	symptoms
496	1.13	1.09	1.17	3.72E-12	3466	56992	Chronic airway obstruction	respiratory
297	1.39	1.26	1.52	1.52E-11	447	44955	Suicidal ideation or attempt	mental disorders
317	1.22	1.15	1.29	5.15E-11	1141	56257	Alcohol-related disorders	mental disorders
530	1.08	1.06	1.11	7.89E-11	7747	45696	Diseases of esophagus	digestive
338	1.11	1.07	1.14	1.60E-10	4174	56656	Pain	neurological
530.1	1.09	1.06	1.12	4.12E-10	6329	45696	Esophagitis, GERD and related diseases	digestive
304	1.25	1.16	1.34	6.70E-10	822	44955	Adjustment reaction	mental disorders
338.2	1.20	1.13	1.27	7.67E-10	1171	56656	Chronic pain	neurological
433	1.10	1.07	1.14	3.07E-09	4261	61020	Cerebrovascular disease	circulatory system
295	1.39	1.24	1.55	4.20E-09	334	44955	Schizophrenia and other psychotic disorders	mental disorders
411	1.08	1.05	1.10	4.65E-09	9907	54084	Ischemic Heart Disease	circulatory system
770	1.18	1.11	1.25	8.93E-09	1285	65166	Myalgia and myositis unspecified	symptoms
300.9	1.41	1.25	1.58	1.02E-08	293	44955	Posttraumatic stress disorder	mental disorders
721.1	1.15	1.10	1.21	1.05E-08	1749	57927	Spondylosis without myelopathy	musculoskeletal
278	1.08	1.05	1.11	1.05E-08	6063	55663	Overweight, obesity and other hyperalimentation	endocrine/metabolic

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
278.1	1.09	1.06	1.12	1.49E-08	4936	55663	Obesity	endocrine/metabolic
278.11	1.13	1.08	1.18	1.77E-08	2313	55663	Morbid obesity	endocrine/metabolic
721	1.14	1.09	1.19	2.56E-08	2057	57927	Spondylosis and allied disorders	musculoskeletal
411.4	1.08	1.05	1.10	5.19E-08	8505	54084	Coronary atherosclerosis	circulatory system
301	1.52	1.30	1.78	1.29E-07	164	44955	Personality disorders	mental disorders
530.11	1.09	1.05	1.12	1.55E-07	4683	45696	GERD	digestive
327.3	1.09	1.05	1.12	2.79E-07	4008	55592	Sleep apnea	neurological
512	1.05	1.03	1.07	3.38E-07	15823	38159	Other symptoms of respiratory system	respiratory
317.1	1.22	1.13	1.32	5.61E-07	647	56257	Alcoholism	mental disorders
216	0.88	0.84	0.93	6.85E-07	1668	64544	Benign neoplasm of skin	neoplasms
313	1.17	1.10	1.24	7.52E-07	1126	67018	Pervasive developmental disorders	mental disorders
295.1	1.46	1.26	1.70	9.19E-07	171	44955	Schizophrenia	mental disorders
250.2	1.06	1.04	1.09	1.50E-06	7590	55837	Type 2 diabetes	endocrine/metabolic
300.11	1.24	1.13	1.37	4.77E-06	451	44955	Generalized anxiety disorder	mental disorders
798	1.07	1.04	1.10	4.97E-06	5940	43662	Malaise and fatigue	symptoms
428	1.07	1.04	1.10	6.58E-06	5562	58987	Congestive heart failure; nonhypertensive	circulatory system
496.2	1.19	1.10	1.29	8.05E-06	672	56992	Chronic bronchitis	respiratory
495	1.09	1.05	1.13	1.50E-05	2817	56992	Asthma	respiratory
297.1	1.32	1.16	1.50	1.59E-05	247	44955	Suicidal ideation	mental disorders
250	1.05	1.03	1.08	1.61E-05	8658	55837	Diabetes mellitus	endocrine/metabolic
532	1.08	1.04	1.13	1.97E-05	3015	45696	Dysphagia	digestive
496.21	1.21	1.11	1.33	2.01E-05	496	56992	Obstructive chronic bronchitis	respiratory
557.1	0.73	0.63	0.85	2.30E-05	185	49703	Celiac disease	digestive
070	1.14	1.07	1.21	2.42E-05	1092	58969	Viral hepatitis	infectious diseases
313.1	1.18	1.09	1.28	2.56E-05	663	67018	Attention deficit hyperactivity disorder	mental disorders

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-4. Phenome-wide association results passing Bonferroni significance for MDD-PRS, adjusted for age at last visit, sex, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
296	1.26	1.22	1.29	1.09E-54	5952	30132	Mood disorders	mental disorders
300	1.24	1.20	1.27	1.27E-43	5410	30132	Anxiety disorders	mental disorders
296.2	1.25	1.21	1.29	8.68E-42	4406	30132	Depression	mental disorders
300.1	1.21	1.17	1.25	4.38E-28	3927	30132	Anxiety disorder	mental disorders
296.22	1.24	1.19	1.30	5.21E-25	2529	30132	Major depressive disorder	mental disorders
296.1	1.26	1.19	1.34	1.16E-13	1134	30132	Bipolar	mental disorders
418	1.09	1.06	1.12	1.52E-10	7839	32907	Nonspecific chest pain	circulatory system
316	1.20	1.13	1.27	3.02E-10	1375	39784	Substance addiction and disorders	mental disorders
530	1.08	1.06	1.11	1.62E-09	7075	31386	Diseases of esophagus	digestive
530.1	1.09	1.06	1.12	1.65E-09	5923	31386	Esophagitis, GERD and related diseases	digestive
785	1.07	1.05	1.10	5.10E-09	9460	31689	Abdominal pain	symptoms
297	1.36	1.22	1.50	6.34E-09	395	30132	Suicidal ideation or attempt	mental disorders
770	1.18	1.12	1.25	8.13E-09	1277	46992	Myalgia and myositis unspecified	symptoms
304	1.24	1.15	1.33	8.65E-09	769	30132	Adjustment reaction	mental disorders
760	1.09	1.06	1.12	9.09E-09	5687	36955	Back pain	symptoms
530.11	1.09	1.06	1.13	2.26E-08	4556	31386	GERD	digestive
338.2	1.18	1.11	1.25	5.22E-08	1146	39419	Chronic pain	neurological
318	1.13	1.08	1.18	6.04E-08	2578	39784	Tobacco use disorder	mental disorders
338	1.09	1.06	1.13	1.15E-07	3998	39419	Pain	neurological
721.1	1.14	1.09	1.20	2.02E-07	1666	40733	Spondylosis without myelopathy	musculoskeletal
721	1.13	1.08	1.18	3.52E-07	1932	40733	Spondylosis and allied disorders	musculoskeletal
300.9	1.36	1.21	1.53	3.93E-07	289	30132	Posttraumatic stress disorder	mental disorders
295	1.36	1.20	1.53	5.86E-07	280	30132	Schizophrenia and other psychotic disorders	mental disorders
278.11	1.12	1.07	1.17	6.36E-07	2110	38824	Morbid obesity	endocrine/metabolic
327.3	1.09	1.05	1.13	6.49E-07	3706	38728	Sleep apnea	neurological
278.1	1.08	1.05	1.11	1.22E-06	4496	38824	Obesity	endocrine/metabolic

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
278	1.07	1.04	1.10	1.32E-06	5529	38824	Overweight, obesity and other hyperalimentation	endocrine/metabolic
433	1.09	1.05	1.13	4.03E-06	3556	44206	Cerebrovascular disease	circulatory system
301	1.48	1.25	1.75	4.54E-06	144	30132	Personality disorders	mental disorders
532	1.10	1.06	1.14	5.10E-06	2638	31386	Dysphagia	digestive
317	1.16	1.09	1.24	8.05E-06	963	39784	Alcohol-related disorders	mental disorders
300.11	1.24	1.13	1.36	9.79E-06	445	30132	Generalized anxiety disorder	mental disorders
798	1.07	1.04	1.10	1.11E-05	5849	29059	Malaise and fatigue	symptoms
313	1.17	1.09	1.25	1.22E-05	886	49476	Pervasive developmental disorders	mental disorders
613	1.17	1.09	1.26	1.92E-05	740	49722	Other nonmalignant breast conditions	genitourinary
345.3	1.17	1.09	1.25	2.12E-05	820	40411	Convulsions	neurological
216	0.90	0.85	0.94	2.56E-05	1625	46209	Benign neoplasm of skin	neoplasms

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-5. Phenome-wide association results passing Bonferroni significance for FEV_1-PRS, adjusted for age at last visit, sex, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250	0.88	0.85	0.90	8.89E-25	7474	39344	Diabetes mellitus	endocrine/metabolic
250.2	0.88	0.86	0.91	1.61E-20	6662	39344	Type 2 diabetes	endocrine/metabolic
250.1	0.80	0.76	0.84	3.55E-18	1456	39344	Type 1 diabetes	endocrine/metabolic
495	0.86	0.83	0.90	3.35E-13	2522	40953	Asthma	respiratory
496	0.87	0.84	0.90	1.35E-12	2901	40953	Chronic airway obstruction	respiratory
250.3	0.83	0.79	0.88	9.81E-12	1350	39344	Insulin pump user	endocrine/metabolic
585	0.91	0.89	0.94	5.16E-10	5635	39043	Renal failure	genitourinary
250.24	0.85	0.81	0.90	7.47E-09	1344	39344	Type 2 diabetes with neurological manifestations	endocrine/metabolic
557.1	0.66	0.57	0.77	4.55E-08	163	34315	Celiac disease	digestive
250.22	0.84	0.79	0.90	4.89E-07	891	39344	Type 2 diabetes with renal manifestations	endocrine/metabolic
585.3	0.91	0.88	0.95	8.54E-07	3457	39043	Chronic renal failure [CKD]	genitourinary
278	0.93	0.91	0.96	1.84E-06	5529	38824	Overweight, obesity and other hyperalimentation	endocrine/metabolic
250.14	0.76	0.67	0.85	2.23E-06	271	39344	Type 1 diabetes with neurological manifestations	endocrine/metabolic
276	0.94	0.92	0.97	3.24E-06	8647	33867	Disorders of fluid, electrolyte, and acid-base balance	endocrine/metabolic
512	0.95	0.93	0.97	6.56E-06	13541	25437	Other symptoms of respiratory system	respiratory
278.1	0.93	0.90	0.96	1.02E-05	4496	38824	Obesity	endocrine/metabolic
250.6	0.85	0.79	0.91	1.22E-05	710	39344	Polyneuropathy in diabetes	endocrine/metabolic
250.7	0.83	0.76	0.90	1.35E-05	541	47391	Diabetic retinopathy	endocrine/metabolic
394.7	0.74	0.65	0.85	1.36E-05	202	42442	Disease of tricuspid valve	circulatory system
401	0.95	0.93	0.97	1.56E-05	17683	25393	Hypertension	circulatory system
070.4	0.72	0.62	0.84	2.15E-05	160	42117	Chronic hepatitis	infectious diseases
250.13	0.73	0.64	0.85	2.56E-05	172	39344	Type 1 diabetes with ophthalmic manifestations	endocrine/metabolic

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-6. Phenome-wide association results passing Bonferroni significance for FVC-PRS, adjusted for age at last visit, sex, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250	0.86	0.84	0.88	3.71E-32	7474	39344	Diabetes mellitus	endocrine/metabolic
250.2	0.86	0.84	0.89	3.29E-27	6662	39344	Type 2 diabetes	endocrine/metabolic
250.1	0.81	0.77	0.86	1.17E-14	1456	39344	Type 1 diabetes	endocrine/metabolic
250.3	0.82	0.78	0.87	2.44E-12	1350	39344	Insulin pump user	endocrine/metabolic
250.24	0.83	0.79	0.88	6.18E-11	1344	39344	Type 2 diabetes with neurological manifestations	endocrine/metabolic
585	0.91	0.88	0.94	1.85E-10	5635	39043	Renal failure	genitourinary
250.22	0.82	0.76	0.87	2.95E-09	891	39344	Type 2 diabetes with renal manifestations	endocrine/metabolic
278	0.92	0.90	0.95	2.42E-08	5529	38824	Overweight, obesity and other hyperalimentation	endocrine/metabolic
278.1	0.92	0.89	0.95	1.17E-07	4496	38824	Obesity	endocrine/metabolic
401	0.94	0.92	0.97	6.05E-07	17683	25393	Hypertension	circulatory system
216	1.13	1.08	1.19	8.03E-07	1625	46209	Benign neoplasm of skin	neoplasms
250.23	0.78	0.71	0.86	8.67E-07	410	39344	Type 2 diabetes with ophthalmic manifestations	endocrine/metabolic
250.7	0.81	0.75	0.88	1.48E-06	541	47391	Diabetic retinopathy	endocrine/metabolic
571	0.91	0.87	0.94	1.68E-06	2687	41807	Chronic liver disease and cirrhosis	digestive
401.1	0.95	0.92	0.97	2.90E-06	16060	25393	Essential hypertension	circulatory system
585.3	0.92	0.89	0.95	3.66E-06	3457	39043	Chronic renal failure [CKD]	genitourinary
571.5	0.90	0.87	0.94	4.63E-06	2187	41807	Other chronic nonalcoholic liver disease	digestive
416	0.88	0.84	0.93	5.47E-06	1430	42948	Cardiomegaly	circulatory system
250.6	0.85	0.78	0.91	9.67E-06	710	39344	Polyneuropathy in diabetes	endocrine/metabolic
276.5	0.91	0.87	0.95	1.28E-05	2279	33867	Hypovolemia	endocrine/metabolic
495	0.92	0.88	0.95	1.56E-05	2522	40953	Asthma	respiratory
250.14	0.77	0.68	0.87	1.57E-05	271	39344	Type 1 diabetes with neurological manifestations	endocrine/metabolic
276	0.95	0.93	0.97	2.40E-05	8647	33867	Disorders of fluid, electrolyte, and acid-base balance	endocrine/metabolic
401.2	0.91	0.88	0.95	2.42E-05	3200	25393	Hypertensive heart and/or renal disease	circulatory system

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-7. Phenome-wide association results passing Bonferroni significance for FEV_1/FVC-PRS, adjusted for age at last visit, sex, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
496	0.83	0.80	0.87	1.54E-19	2901	40953	Chronic airway obstruction	respiratory
495	0.87	0.83	0.90	1.16E-11	2522	40953	Asthma	respiratory
496.2	0.76	0.70	0.83	3.73E-10	570	40953	Chronic bronchitis	respiratory
496.21	0.74	0.67	0.82	1.60E-09	425	40953	Obstructive chronic bronchitis	respiratory
557.1	0.63	0.54	0.74	1.15E-08	163	34315	Celiac disease	digestive
250.1	0.86	0.82	0.91	3.86E-08	1456	39344	Type 1 diabetes	endocrine/metabolic
496.1	0.77	0.70	0.85	7.71E-08	447	40953	Emphysema	respiratory

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-8. Phenome-wide association results passing Bonferroni significance for PEF-PRS, adjusted for age at last visit, sex, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250.1	0.81	0.77	0.85	2.86E-16	1456	39344	Type 1 diabetes	endocrine/metabolic
250	0.92	0.90	0.94	2.64E-11	7474	39344	Diabetes mellitus	endocrine/metabolic
496	0.88	0.85	0.92	4.40E-10	2901	40953	Chronic airway obstruction	respiratory
557.1	0.66	0.57	0.76	5.69E-09	163	34315	Celiac disease	digestive
709	0.80	0.75	0.87	7.42E-09	660	43401	Diffuse diseases of connective tissue	dermatologic
495	0.89	0.86	0.93	1.15E-08	2522	40953	Asthma	respiratory
250.2	0.93	0.90	0.95	1.41E-08	6662	39344	Type 2 diabetes	endocrine/metabolic
250.3	0.86	0.82	0.91	5.76E-08	1350	39344	Insulin pump user	endocrine/metabolic
709.2	0.72	0.63	0.82	8.12E-07	201	43401	Sicca syndrome	dermatologic
250.14	0.78	0.70	0.88	2.29E-05	271	39344	Type 1 diabetes with neurological manifestations	endocrine/metabolic

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-9. Phenome-wide association results passing Bonferroni significance for MDD-PRS in women, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
296	1.28	1.23	1.32	8.52E-39	3815	16045	Mood disorders	mental disorders
300	1.25	1.20	1.29	3.33E-31	3619	16045	Anxiety disorders	mental disorders
296.2	1.26	1.21	1.32	1.03E-29	2936	16045	Depression	mental disorders
300.1	1.22	1.17	1.27	1.82E-20	2648	16045	Anxiety disorder	mental disorders
296.22	1.25	1.19	1.31	2.54E-17	1688	16045	Major depressive disorder	mental disorders
296.1	1.33	1.23	1.44	9.36E-13	681	16045	Bipolar	mental disorders
530	1.12	1.08	1.16	6.63E-10	3885	18431	Diseases of esophagus	digestive
785	1.09	1.06	1.12	9.57E-09	6119	17020	Abdominal pain	symptoms
530.1	1.12	1.07	1.16	1.27E-08	3321	18431	Esophagitis, GERD and related diseases	digestive
770	1.19	1.12	1.26	1.80E-08	1111	25716	Myalgia and myositis unspecified	symptoms
530.11	1.13	1.08	1.17	2.59E-08	2660	18431	GERD	digestive
760	1.09	1.06	1.13	1.10E-06	3617	20345	Back pain	symptoms
304	1.23	1.13	1.35	4.93E-06	507	16045	Adjustment reaction	mental disorders
418	1.08	1.04	1.12	1.13E-05	4266	19221	Nonspecific chest pain	circulatory system
316	1.19	1.10	1.28	1.35E-05	710	23875	Substance addiction and disorders	mental disorders
613	1.18	1.09	1.27	1.41E-05	730	26872	Other nonmalignant breast conditions	genitourinary
300.9	1.37	1.18	1.58	1.92E-05	198	16045	Posttraumatic stress disorder	mental disorders
216	0.87	0.82	0.93	2.11E-05	1055	25996	Benign neoplasm of skin	neoplasms

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-10. Phenome-wide association results passing Bonferroni significance for MDD-PRS in men, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
296	1.23	1.17	1.28	1.76E-17	2137	14087	Mood disorders	mental disorders
300	1.21	1.15	1.27	1.06E-13	1791	14087	Anxiety disorders	mental disorders
296.2	1.23	1.16	1.30	1.22E-13	1470	14087	Depression	mental disorders
296.22	1.24	1.15	1.33	4.03E-09	841	14087	Major depressive disorder	mental disorders
300.1	1.19	1.12	1.26	5.61E-09	1279	14087	Anxiety disorder	mental disorders
418	1.09	1.05	1.14	3.48E-06	3573	13686	Nonspecific chest pain	circulatory system
316	1.21	1.11	1.31	5.91E-06	665	15909	Substance addiction and disorders	mental disorders
318	1.14	1.08	1.22	1.48E-05	1307	15909	Tobacco use disorder	mental disorders

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-11. Phenome-wide association results passing Bonferroni significance for FEV_1-PRS in women, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250	0.88	0.85	0.91	2.93E-12	3574	23498	Diabetes mellitus	endocrine/metabolic
250.1	0.79	0.74	0.85	5.19E-11	793	23498	Type 1 diabetes	endocrine/metabolic
250.2	0.89	0.86	0.92	5.83E-10	3136	23498	Type 2 diabetes	endocrine/metabolic
495	0.86	0.82	0.91	6.22E-09	1660	23492	Asthma	respiratory
250.3	0.83	0.77	0.90	2.23E-06	678	23498	Insulin pump user	endocrine/metabolic
401	0.93	0.90	0.96	5.00E-06	8733	16298	Hypertension	circulatory system
250.24	0.83	0.77	0.90	6.26E-06	606	23498	Type 2 diabetes with neurological manifestations	endocrine/metabolic
496	0.88	0.83	0.93	9.24E-06	1352	23492	Chronic airway obstruction	respiratory
401.1	0.93	0.90	0.96	1.49E-05	8022	16298	Essential hypertension	circulatory system
557.1	0.68	0.57	0.81	2.67E-05	110	19752	Celiac disease	digestive

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-12. Phenome-wide association results passing Bonferroni significance for FEV_1-PRS in men, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250	0.87	0.84	0.90	4.72E-14	3900	15846	Diabetes mellitus	endocrine/metabolic
250.2	0.88	0.84	0.91	4.76E-12	3526	15846	Type 2 diabetes	endocrine/metabolic
250.1	0.80	0.74	0.86	9.91E-09	663	15846	Type 1 diabetes	endocrine/metabolic
496	0.86	0.82	0.91	3.73E-08	1549	17461	Chronic airway obstruction	respiratory
585	0.90	0.87	0.94	1.20E-07	3283	15262	Renal failure	genitourinary
250.3	0.83	0.77	0.89	8.77E-07	672	15846	Insulin pump user	endocrine/metabolic
276	0.92	0.89	0.95	4.96E-06	4395	13550	Disorders of fluid, electrolyte, and acid-base balance	endocrine/metabolic
442.8	0.42	0.28	0.62	9.69E-06	21	17459	Aneurysm of other specified artery	circulatory system
585.3	0.90	0.86	0.94	1.13E-05	2055	15262	Chronic renal failure [CKD]	genitourinary
495	0.86	0.80	0.92	1.25E-05	862	17461	Asthma	respiratory

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-13. Phenome-wide association results passing Bonferroni significance for FVC-PRS in women, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250	0.86	0.83	0.89	2.39E-17	3574	23498	Diabetes mellitus	endocrine/metabolic
250.2	0.86	0.83	0.90	2.06E-14	3136	23498	Type 2 diabetes	endocrine/metabolic
250.1	0.80	0.74	0.86	4.10E-10	793	23498	Type 1 diabetes	endocrine/metabolic
250.24	0.80	0.74	0.87	1.02E-07	606	23498	Type 2 diabetes with neurological manifestations	endocrine/metabolic
250.3	0.82	0.76	0.89	3.84E-07	678	23498	Insulin pump user	endocrine/metabolic
278	0.92	0.88	0.95	1.21E-06	3516	21540	Overweight, obesity and other hyperalimentation	endocrine/metabolic
278.1	0.91	0.87	0.94	1.52E-06	2830	21540	Obesity	endocrine/metabolic
401	0.93	0.90	0.96	3.49E-06	8733	16298	Hypertension	circulatory system
401.1	0.93	0.90	0.96	6.91E-06	8022	16298	Essential hypertension	circulatory system
278.11	0.89	0.84	0.94	1.44E-05	1471	21540	Morbid obesity	endocrine/metabolic
585	0.91	0.87	0.95	2.49E-05	2352	23781	Renal failure	genitourinary
216	1.14	1.07	1.22	2.56E-05	1055	25996	Benign neoplasm of skin	neoplasms
401.2	0.88	0.82	0.93	2.69E-05	1324	16298	Hypertensive heart and/or renal disease	circulatory system

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-14. Phenome-wide association results passing Bonferroni significance for FVC-PRS in men, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250	0.86	0.83	0.89	2.70E-16	3900	15846	Diabetes mellitus	endocrine/metabolic
250.2	0.86	0.83	0.90	3.14E-14	3526	15846	Type 2 diabetes	endocrine/metabolic
250.3	0.83	0.77	0.89	1.32E-06	672	15846	Insulin pump user	endocrine/metabolic
585	0.91	0.88	0.95	2.30E-06	3283	15262	Renal failure	genitourinary
250.1	0.83	0.77	0.90	3.64E-06	663	15846	Type 1 diabetes	endocrine/metabolic
250.22	0.82	0.76	0.89	3.74E-06	583	15846	Type 2 diabetes with renal manifestations	endocrine/metabolic
276	0.92	0.89	0.96	7.03E-06	4395	13550	Disorders of fluid, electrolyte, and acid-base balance	endocrine/metabolic
571	0.89	0.84	0.94	1.68E-05	1402	17306	Chronic liver disease and cirrhosis	digestive
276.5	0.87	0.82	0.93	2.23E-05	1082	13550	Hypovolemia	endocrine/metabolic

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-15. Phenome-wide association results passing Bonferroni significance for FEV_1/FVC-PRS in women, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
496	0.82	0.78	0.87	3.27E-11	1352	23492	Chronic airway obstruction	respiratory
495	0.88	0.83	0.92	2.93E-07	1660	23492	Asthma	respiratory
496.21	0.73	0.64	0.84	5.00E-06	231	23492	Obstructive chronic bronchitis	respiratory
496.2	0.77	0.68	0.86	7.19E-06	307	23492	Chronic bronchitis	respiratory

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-16. Phenome-wide association results passing Bonferroni significance for FEV_1/FVC-PRS in men, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
496	0.84	0.80	0.89	6.07E-10	1549	17461	Chronic airway obstruction	respiratory
495	0.85	0.79	0.91	5.05E-06	862	17461	Asthma	respiratory
496.2	0.76	0.67	0.86	1.22E-05	263	17461	Chronic bronchitis	respiratory
557.1	0.54	0.41	0.71	1.38E-05	53	14563	Celiac disease	digestive

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-17. Phenome-wide association results passing Bonferroni significance for PEF-PRS in women, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250.1	0.80	0.75	0.86	1.07E-10	793	23498	Type 1 diabetes	endocrine/metabolic
709	0.81	0.74	0.87	9.71E-08	575	24051	Diffuse diseases of connective tissue	dermatologic
496	0.86	0.81	0.91	1.64E-07	1352	23492	Chronic airway obstruction	respiratory
250	0.92	0.88	0.95	1.00E-06	3574	23498	Diabetes mellitus	endocrine/metabolic
250.3	0.84	0.78	0.90	2.36E-06	678	23498	Insulin pump user	endocrine/metabolic
495	0.89	0.85	0.94	3.75E-06	1660	23492	Asthma	respiratory
557.1	0.67	0.57	0.80	9.81E-06	110	19752	Celiac disease	digestive
709.2	0.75	0.65	0.86	2.49E-05	190	24051	Sicca syndrome	dermatologic

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-18. Phenome-wide association results passing Bonferroni significance for PEF-PRS in men, adjusted for age at last visit, ever smoking, and first three principal components.

Phecode	OR	L95	U95	P	Ncases	Ncontrols	Phecode description	Phecode group
250.1	0.82	0.77	0.89	3.90E-07	663	15846	Type 1 diabetes	endocrine/metabolic
250	0.92	0.89	0.95	5.76E-06	3900	15846	Diabetes mellitus	endocrine/metabolic

OR: odds ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval, Ncases: number of cases, Ncontrols: number of controls

Table 6-19. Single nucleotide polymorphisms identified as potentially pleiotropic between FEV_1/FVC and major depressive disorder.

Chromosome	Position	SNP	Nearest gene	Phenotype	PMID	Reference
1	176,030,977	rs12040241	*COP1*	Body mass index	20935630	[450]
				Waist-hip ratio	20935629	[451]
3	49,173,299	rs7617480	*KLHDC8B*	Adiponectin levels	22479202	[452]
				Advanced age-related macular degeneration	23455636	[453]
				Age at menarche	21102462	[454]
				Bone mineral density (hip and spine)	18445777	[455]
				Cholesterol (HDL, LDL, and total)	20339536, 19060906	[456,457]
				College completion	23722424	[458]
				Late-onset Alzheimer's disease	21390209	[459]
				Parkinson's disease	22451204	[460]
				Paternal transmission distortion	22377632	[461]
				Serum creatinine	20383146	[462]
				Smoking cessation	30643251	[370]
				Sporadic Creutzfeldt-Jakob disease	22210626	[463]
				Subjective well-being	29292387	[464]
				Triglycerides	20686565	[465]
				Years of education	23722424	[458]
18	37,558,282	rs12967855	*CELF4*	College completion	23722424	[458]
				Depressed affect	29942085	[375]
				Height	20881960	[466]
				Household income	31844048	[467]
				Infant head circumference	22504419	[468]
				Lifetime smoking index	31689377	[372]
				Mitral annular calcium	23388002	[469]

				Unipolar depression	32231276, 30718901	373,374
				Years of education	23722424, 30595370	458,470

Table 6-20. Comparison of BioVU participants with and without smoking data.

Characteristic	**Missing Smoking Data (N = 18,944)**	**Complete Cases (N = 53,503)**
Median age (IQR)	47 (18-65)	58 (41-69)
Gender (N, %)		
Female	10,069 (53.2)	30,515 (57.0)
Male	8,873 (46.8)	22,988 (43.0)
Missing	2	0
Chronic obstructive pulmonary disease (N, %)	565 (3.0)	2,901 (5.4)
Major depressive disorder (N, %)	161 (0.8)	2,529 (4.7)

6.2 Appendix 2

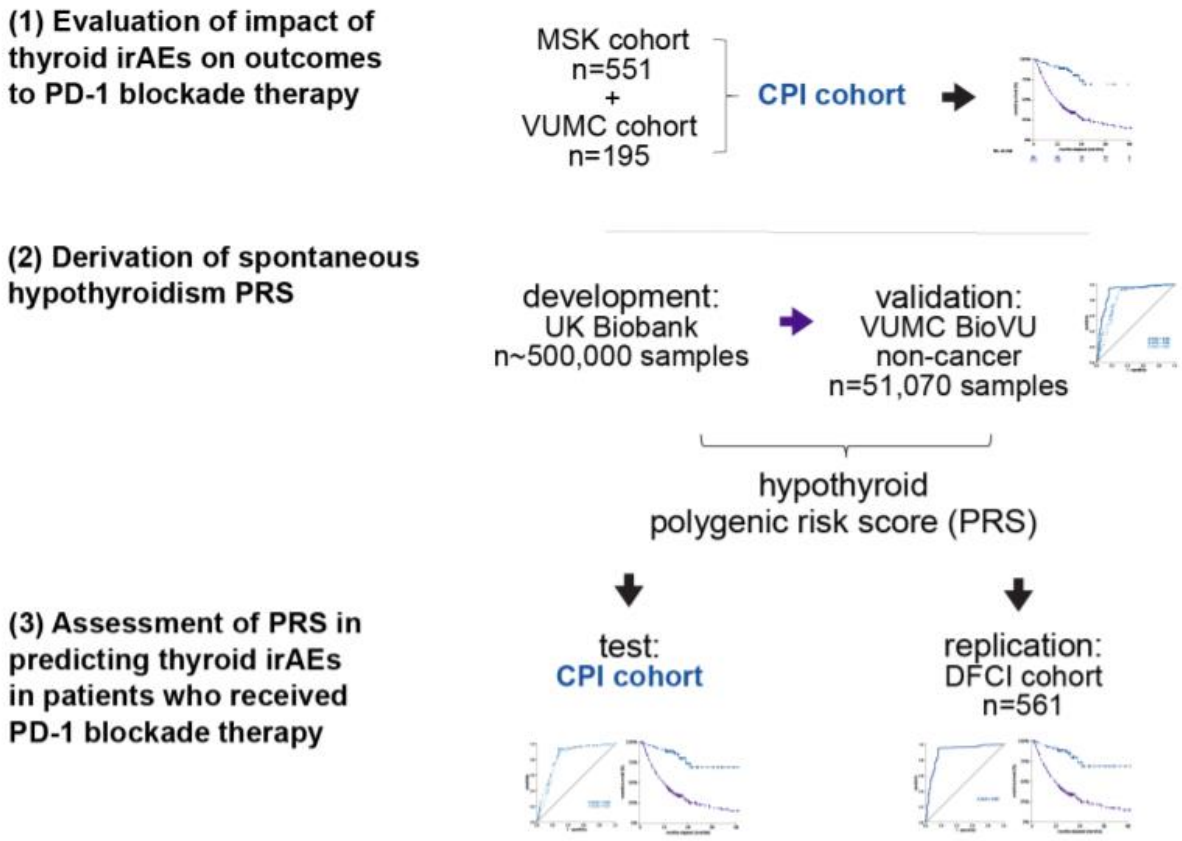

Figure 6-7. Schema of the methods for examining the relationships between spontaneous hypothyroidism, thyroid irAEs, and response to anti-PD-(L)1 therapy.

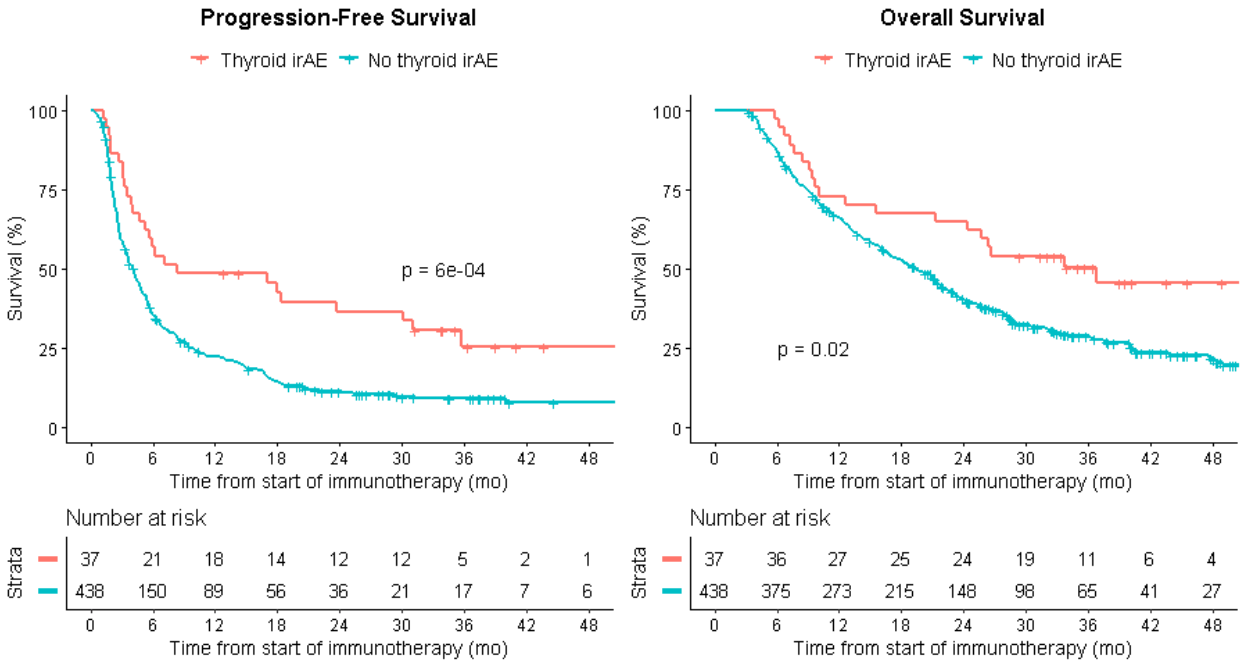

Figure 6-8. Thyroid irAEs as a predictor of PFS and OS in individuals with OS > 90 days in the MSK cohort. Kaplan-Meier survival curves are unadjusted and compare those who had a thyroid irAE to those who did not have a thyroid irAE. The xaxis reflects time from start of CPI therapy.

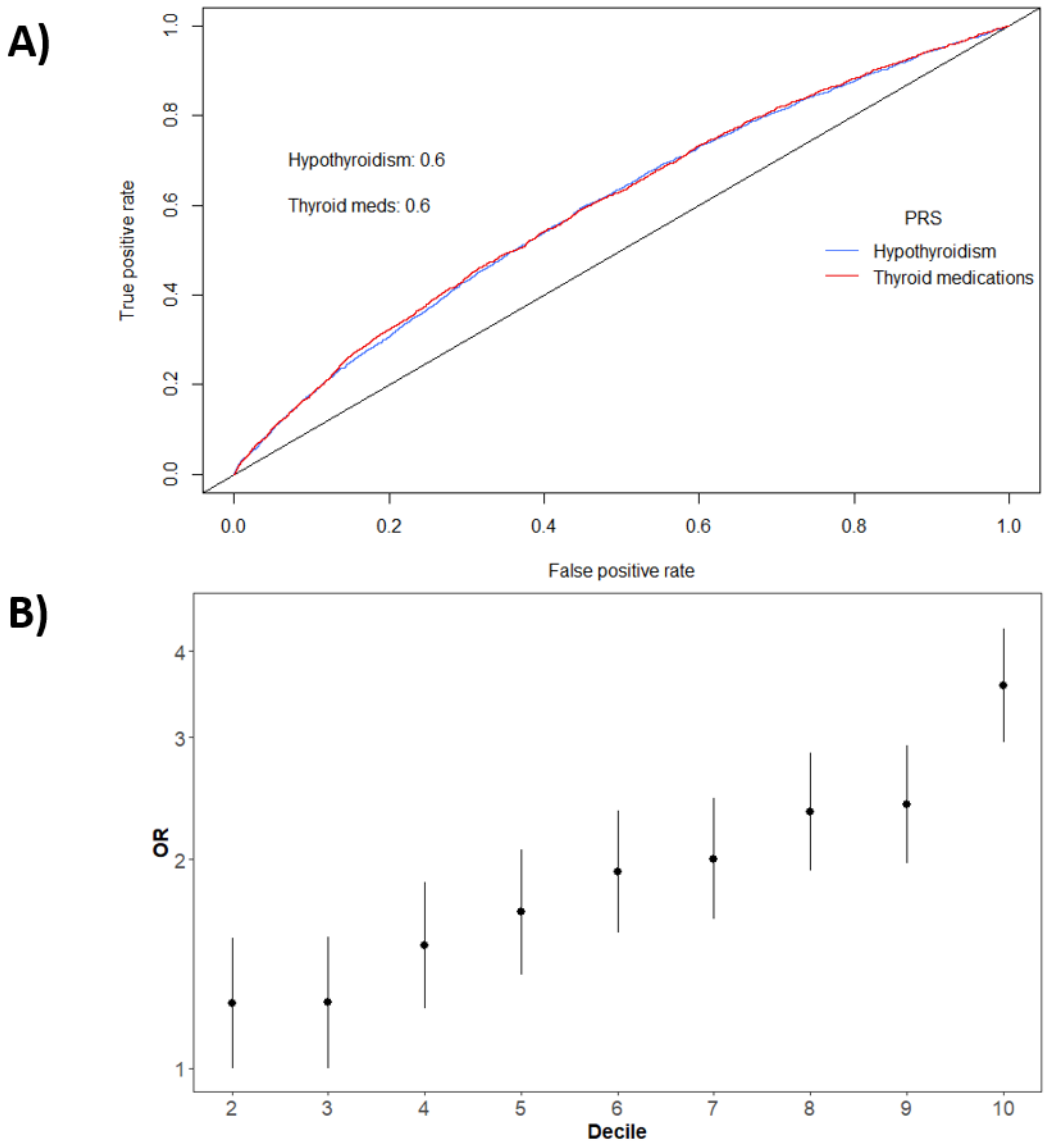

Figure 6-9. Validation of PRS models for thyroid disease developed in UK Biobank in BioVU. A) ROC curves for PRS for self-reported hypothyroidism and taking thyroid medications were developed in UK Biobank using LDpred and tested in BioVU among participants who were not known to have lung cancer and received CPI. B) Relative risk of PRS by decile in the non-cancer VUMC BioVU cohort.

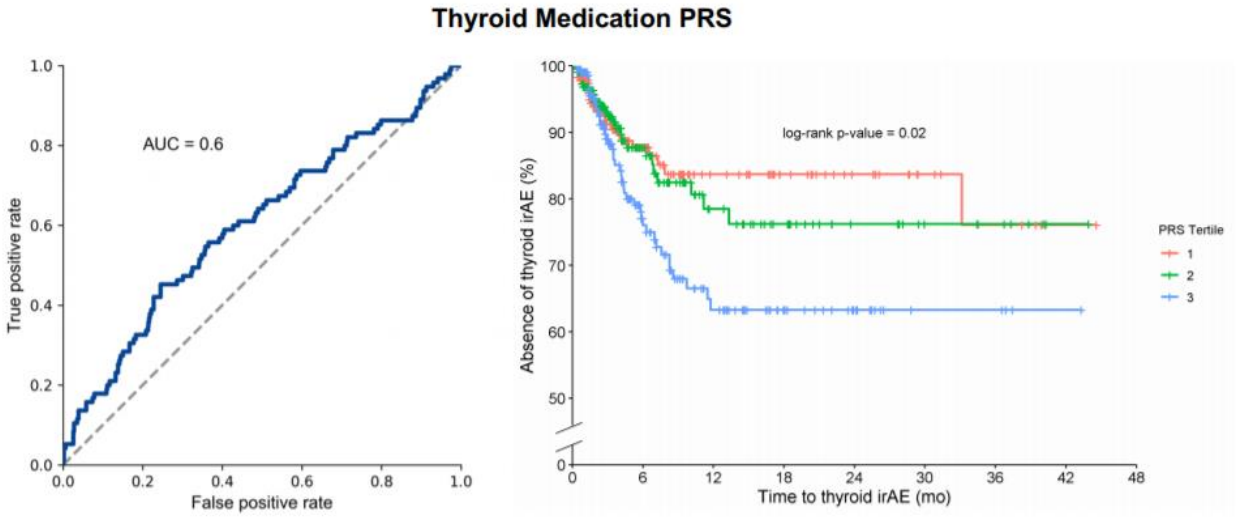

Figure 6-10. Thyroid medication PRS as a predictor of CPI-induced hypothyroidism events in the VUMC and MSK cohort. The left panel shows the ROC curve and right panel shows time to event by PRS tertile. P-values for the three curves are calculated using a log-rank test.

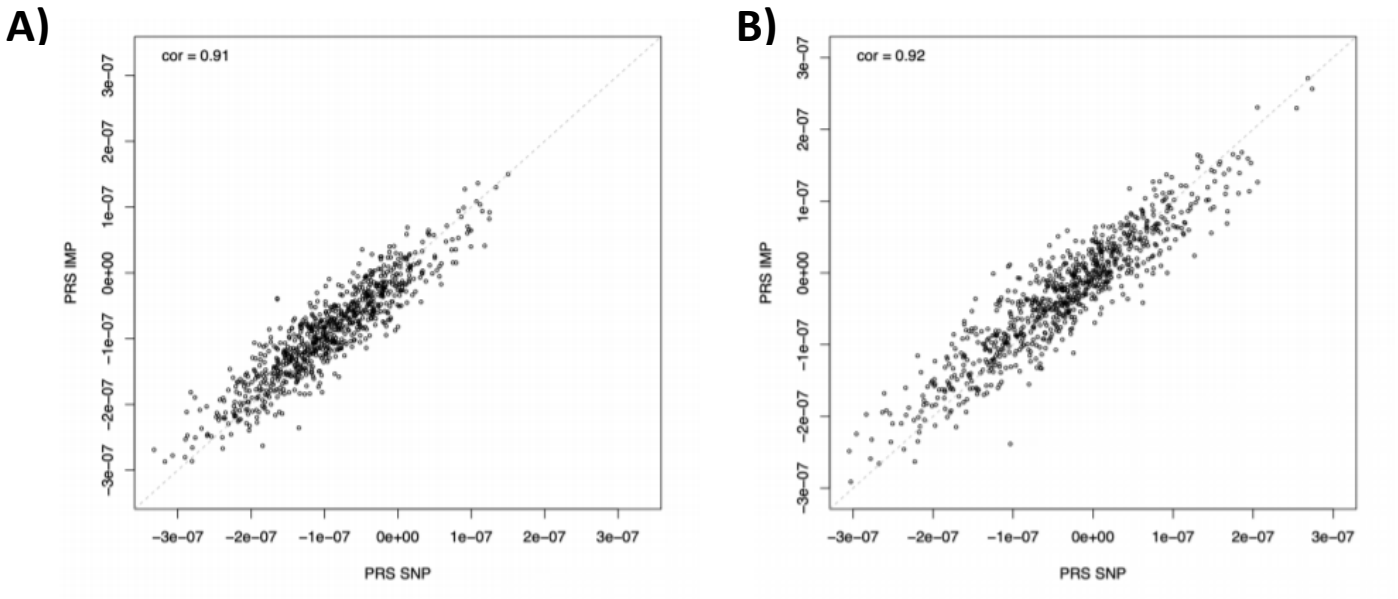

Figure 6-11. PRS values for DFCI tumor imputed and germline genotyped samples. PRSs shown are: A) hypothyroidism and B) thyroid medication. Each point represents an individual with genotyped (x-axis) and imputed (y-axis) PRS values. Pearson correlation between the scores is listed in each panel.

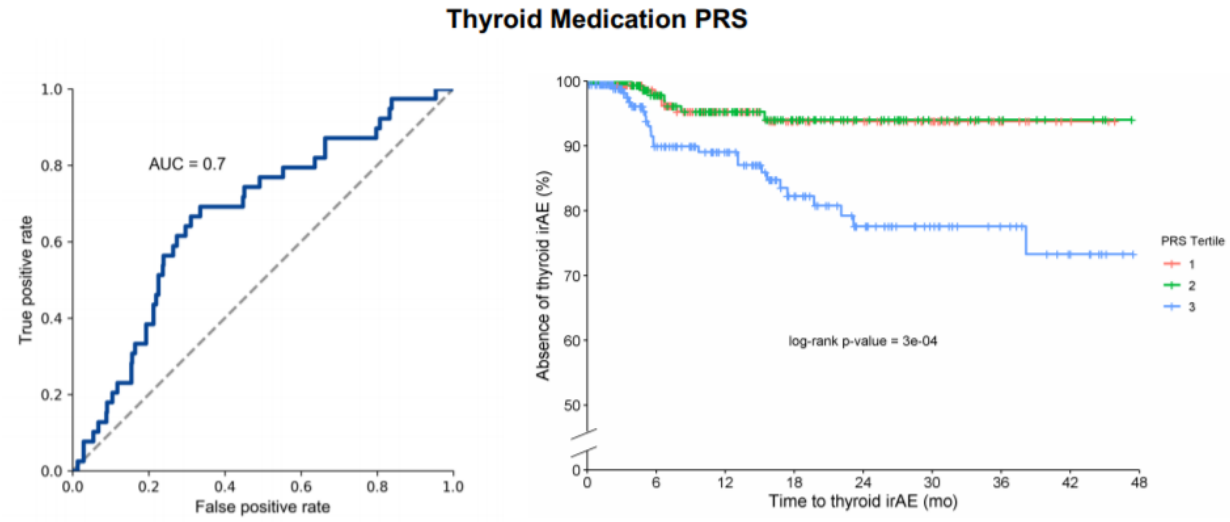

Figure 6-12. Hypothyroidism PRS (using thyroid medication PRS) as a predictor of CPI-induced hypothyroidism events in the DFCI cohort. The left panel shows the ROC curve and right panel shows time to event by PRS tertile. P-values for the three curves are calculated using a log-rank test.

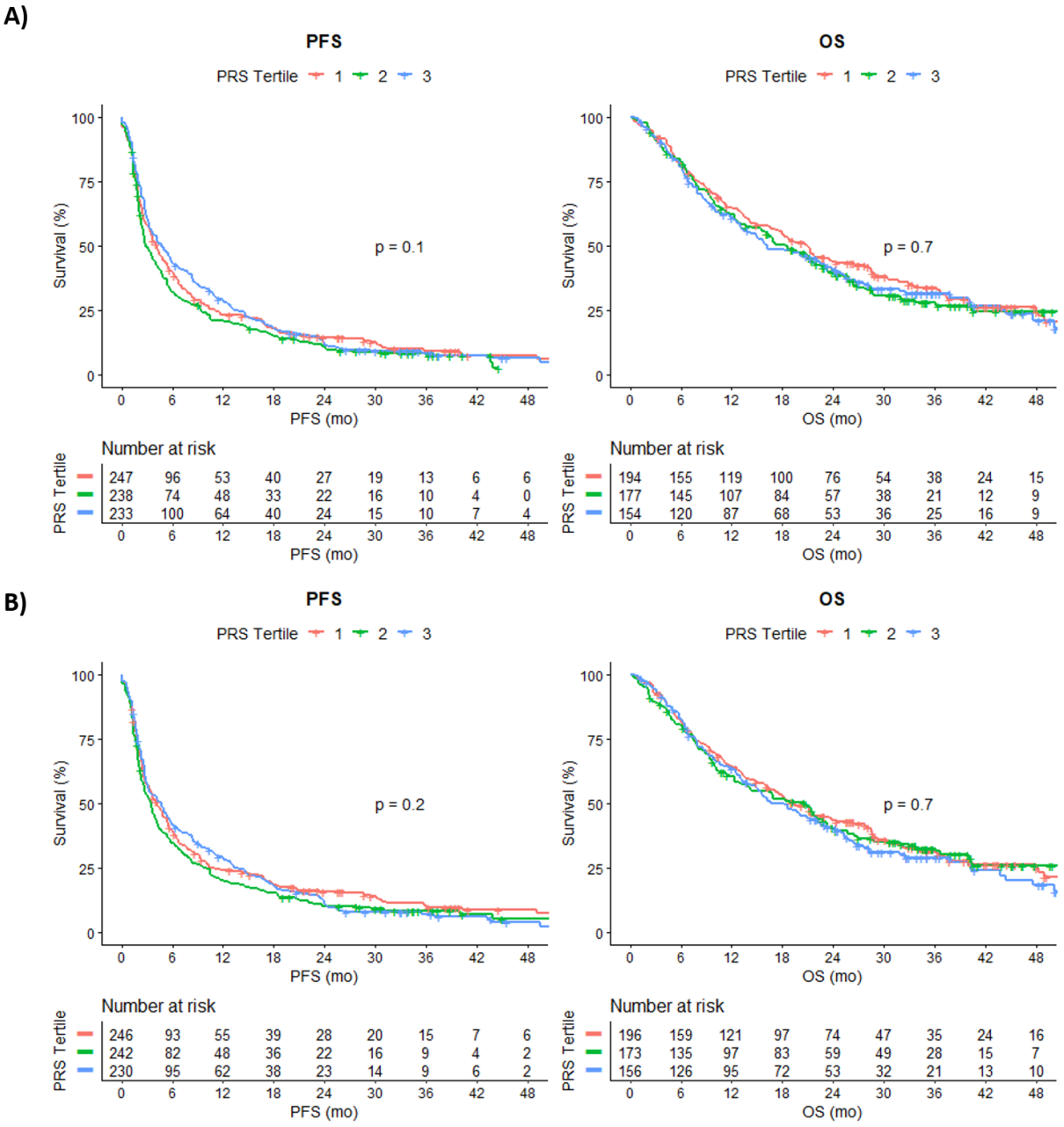

Figure 6-13. PFS in the combined MSK + VUMC cohort and OS in the MSK cohort starting from the time of CPI therapy start by PRS tertile. (A) hypothyroidism PRS and (B) thyroid medication PRS.

Table 6-21. Individual SNPs from the UK Biobank hypothyroidism GWAS that were associated with CPI-induced thyroid irAEs in the MSK + VUMC cohort at p < 0.05. Bold italics indicates variant passed multiple hypothesis testing. Effect and alternate alleles in the MSK + VUMC cohort were aligned to the effect and alternate alleles in the original UK Biobank GWAS. Models in the MSK + VUMC cohort were adjusted for age at diagnosis, sex, and the first ten principal components.

Variant Information					**MSK + VUMC Cohort Results**					**UK Biobank GWAS Results**			
Chr	**Position (GRCh37)**	**SNP**	**Effect allele**	**Alt allele**	**Effect allele freq**	**HR**	**L95**	**U95**	**p-value**	**Effect allele freq**	**β**	**Standard error**	**p-value**
6	32379295	rs9268515	G	C	0.86	0.5	0.35	0.71	1.12E-04	0.81	-0.00519	0.00883	2.80E-09
16	67973953	rs5923	G	A	0.93	0.45	0.28	0.72	8.80E-04	0.95	-0.00574	0.001039	2.60E-08
16	67476572	rs8056260	A	G	0.91	0.43	0.25	0.72	1.37E-03	0.96	-0.00704	0.001075	3.80E-11
8	141639262	rs11783023	C	T	0.26	1.62	1.19	2.21	2.09E-03	0.28	0.003325	0.000491	1.20E-11
16	68361498	rs7206600	C	T	0.96	0.41	0.23	0.73	2.41E-03	0.98	-0.00807	0.001454	2.60E-08
2	204769395	rs28386480	C	T	0.52	1.55	1.15	2.10	3.72E-03	0.54	0.002749	0.000442	6.60E-10
6	32608931	rs9272679	T	C	0.76	1.76	1.20	2.57	3.77E-03	0.43	-0.00992	0.000665	5.80E-50
6	32587350	rs1281932	G	A	0.81	0.62	0.45	0.86	4.68E-03	0.80	-0.00974	0.000792	1.40E-34
20	17860022	rs6111715	G	C	0.81	0.61	0.43	0.86	4.87E-03	0.82	0.003711	0.000576	7.80E-11
2	204698111	rs2882970	T	C	0.74	1.7	1.17	2.48	5.53E-03	0.75	0.005722	0.000518	8.70E-29
1	200840467	rs12756886	T	C	0.89	0.56	0.37	0.85	6.01E-03	0.88	-0.00437	0.00068	9.40E-11
6	32633282	rs9274447	T	C	0.79	1.83	1.19	2.81	6.15E-03	0.66	-0.00775	0.00069	1.80E-28
4	10716939	rs4293777	G	C	0.53	0.66	0.49	0.89	6.23E-03	0.53	-0.0037	0.000442	7.80E-17

Variant Information					MSK + VUMC Cohort Results					UK Biobank GWAS Results			
Chr	**Position (GRCh37)**	**SNP**	**Effect allele**	**Alt allele**	**Effect allele freq**	**HR**	**L95**	**U95**	**p-value**	**Effect allele freq**	**β**	**Standard error**	**p-value**
16	68390697	rs61733486	C	T	0.94	0.5	0.3	0.83	7.28E-03	0.94	-0.00521	0.000946	2.30E-08
6	29367498	rs4713220	A	T	0.56	1.52	1.12	2.06	7.31E-03	0.57	0.003625	0.000654	2.60E-10
12	111582630	rs73413596	T	C	0.87	0.56	0.37	0.86	7.86E-03	0.92	0.00514	0.000839	3.10E-10
11	116979911	rs12271161	G	A	0.78	1.83	1.17	2.86	7.88E-03	0.81	0.003652	0.000559	3.10E-11
6	31437566	rs9404989	G	T	0.98	0.32	0.14	0.74	8.22E-03	0.98	0.013439	0.001705	2.50E-14
7	37382465	rs60600003	T	G	0.9	0.56	0.36	0.86	8.24E-03	0.9	-0.0044	0.000739	9.30E-10
6	31312058	rs2394977	C	G	0.58	0.68	0.51	0.91	1.04E-02	0.56	0.004264	0.000653	1.00E-15
6	31463128	rs9267352	G	A	0.74	0.66	0.47	0.91	1.14E-02	0.74	0.005833	0.000676	7.20E-16
6	31616174	rs569347663	A	T	0.97	0.44	0.24	0.84	1.20E-02	1	-0.0354	0.005556	1.60E-10
2	1378060	rs4927602	G	A	0.49	0.69	0.51	0.93	1.36E-02	0.5	-0.00264	0.000467	2.40E-08
3	188070964	rs2103022	A	G	0.23	1.53	1.09	2.15	1.41E-02	0.24	0.002904	0.000527	1.40E-08
14	98692996	rs1257926	G	A	0.53	0.7	0.53	0.94	1.59E-02	0.52	-0.00278	0.000443	1.10E-10
6	32605295	rs1129735	C	T	0.66	0.7	0.52	0.94	1.67E-02	0.63	-0.00923	0.0007	8.00E-41
6	31435869	rs4713466	C	T	0.89	0.62	0.42	0.92	1.71E-02	0.86	0.005798	0.000943	3.10E-09
6	31436074	rs2518028	T	C	0.32	0.66	0.47	0.93	1.73E-02	0.35	0.005452	0.000708	8.90E-19
14	68749927	rs3784099	G	A	0.67	1.5	1.07	2.11	1.77E-02	0.72	0.003107	0.000491	2.70E-11
1	108379684	rs2125748	G	A	0.77	1.57	1.08	2.28	1.89E-02	0.75	0.002831	0.000512	1.80E-08

Variant Information					MSK + VUMC Cohort Results					UK Biobank GWAS Results			
Chr	**Position (GRCh37)**	**SNP**	**Effect allele**	**Alt allele**	**Effect allele freq**	**HR**	**L95**	**U95**	**p-value**	**Effect allele freq**	**β**	**Standard error**	**p-value**
6	32587588	rs9271406	A	G	0.56	0.71	0.54	0.95	2.08E-02	0.5	-0.00606	0.000686	4.60E-23
17	7240391	rs61759532	C	T	0.83	0.69	0.5	0.94	2.11E-02	0.75	-0.00399	0.000527	1.00E-14
6	32580591	rs7449585	G	T	0.74	0.7	0.52	0.95	2.12E-02	0.95	0.010027	0.001139	1.60E-20
6	32624377	rs567302488	G	A	0.93	3.19	1.19	8.59	2.15E-02	0.62	-0.00888	0.000693	1.70E-37
6	31247441	rs2844607	C	T	0.78	0.69	0.5	0.95	2.27E-02	0.72	0.005255	0.000684	1.30E-10
9	21637351	rs71504798	C	G	0.92	3.11	1.17	8.26	2.28E-02	0.91	-0.00499	0.000767	8.20E-11
11	117030633	rs200545029	T	C	0.94	2.75	1.15	6.59	2.34E-02	0.94	0.006043	0.000953	8.60E-11
6	32078373	rs3807039	A	C	0.92	0.61	0.39	0.94	2.36E-02	0.89	-0.00487	0.000805	2.20E-08
3	105911539	rs7633167	C	A	0.44	0.71	0.53	0.96	2.37E-02	0.43	0.002601	0.000448	5.70E-09
1	108337108	rs17484960	G	A	0.55	0.72	0.54	0.96	2.38E-02	0.5	-0.00312	0.000441	9.00E-13
1	236629134	rs12117927	C	A	0.54	0.71	0.53	0.96	2.52E-02	0.51	-0.00267	0.000452	1.10E-09
6	32402889	rs9268615	G	A	0.55	0.72	0.55	0.96	2.55E-02	0.64	-0.0069	0.000848	1.10E-15
19	7240776	rs4804433	G	T	0.25	1.42	1.04	1.93	2.56E-02	0.21	0.00325	0.00054	5.80E-09
6	32603321	rs62404084	C	T	0.84	0.66	0.45	0.95	2.71E-02	0.81	0.005054	0.000717	6.40E-10
3	12195622	rs308952	A	G	0.14	0.52	0.3	0.93	2.75E-02	0.13	0.004684	0.000648	1.40E-13
3	188080043	rs7640386	T	C	0.75	0.68	0.48	0.96	2.78E-02	0.78	-0.00432	0.000532	2.00E-16
9	100506414	rs7862400	C	A	0.44	0.71	0.52	0.96	2.80E-02	0.53	-0.00331	0.000447	3.90E-14

Variant Information					MSK + VUMC Cohort Results					UK Biobank GWAS Results			
Chr	**Position (GRCh37)**	**SNP**	**Effect allele**	**Alt allele**	**Effect allele freq**	**HR**	**L95**	**U95**	**p-value**	**Effect allele freq**	**β**	**Standard error**	**p-value**
9	100737755	rs10984601	A	G	0.74	0.71	0.52	0.96	2.88E-02	0.7	-0.00388	0.000491	1.50E-15
16	67349478	rs138453996	G	A	0.97	0.46	0.23	0.93	2.97E-02	0.98	-0.01041	0.001578	3.10E-11
6	29566369	rs3095273	A	G	0.26	0.67	0.47	0.96	3.00E-02	0.29	-0.00376	0.000667	3.00E-08
6	32781776	rs2856997	C	A	0.6	0.71	0.52	0.97	3.00E-02	0.61	0.003736	0.000577	1.60E-08
6	91024294	rs927297	G	C	0.35	1.39	1.03	1.86	3.04E-02	0.4	0.003524	0.000456	4.70E-15
9	5425847	rs10815220	A	G	0.75	1.47	1.04	2.09	3.09E-02	0.72	-0.00308	0.000491	4.80E-10
6	31850308	rs74434374	C	A	0.95	0.55	0.32	0.95	3.17E-02	0.95	0.008211	0.001318	2.30E-10
6	31249127	rs2253487	G	A	0.6	0.73	0.55	0.97	3.18E-02	0.59	0.004741	0.000632	2.00E-09
9	21585265	rs970987	C	A	0.37	0.69	0.49	0.97	3.33E-02	0.34	0.003916	0.000468	3.80E-17
2	204573392	rs35988305	C	G	0.62	1.38	1.02	1.87	3.59E-02	0.64	0.003279	0.000459	2.90E-13
6	31397689	rs143015185	A	T	0.96	0.53	0.3	0.96	3.61E-02	0.95	-0.01705	0.001288	5.30E-25
4	149648035	rs17583283	G	A	0.52	0.74	0.55	0.98	3.62E-02	0.56	0.00319	0.000446	3.50E-13
5	133455153	rs244693	A	C	0.07	1.66	1.03	2.67	3.64E-02	0.08	0.005357	0.000824	9.60E-11
10	63809624	rs10821948	C	A	0.62	0.73	0.55	0.98	3.77E-02	0.68	-0.00397	0.000474	1.20E-17
6	32667595	rs1794279	G	T	0.92	2.13	1.04	4.37	3.85E-02	0.87	-0.0141	0.001508	2.80E-19
6	32582603	rs13204736	G	T	0.65	0.74	0.55	0.99	3.94E-02	0.7	-0.00642	0.00762	3.00E-18
4	40307564	rs13136820	C	T	0.36	1.36	1.02	1.83	3.95E-02	0.32	0.002945	0.000475	8.90E-10

Variant Information					MSK + VUMC Cohort Results					UK Biobank GWAS Results			
Chr	**Position (GRCh37)**	**SNP**	**Effect allele**	**Alt allele**	**Effect allele freq**	**HR**	**L95**	**U95**	**p-value**	**Effect allele freq**	**β**	**Standard error**	**p-value**
19	50197406	rs12981033	A	G	0.58	0.73	0.54	0.99	3.98E-02	0.61	0.00303	0.000452	7.00E-12
1	108321313	rs4914960	A	G	0.22	0.66	0.44	0.98	4.08E-02	0.19	-0.00421	0.00056	3.70E-14
4	149720635	rs13143096	C	T	0.5	0.74	0.55	0.99	4.14E-02	0.5	0.002441	0.000444	1.90E-08
8	8317887	rs2921059	G	T	0.95	0.56	0.32	0.98	4.19E-02	0.56	0.005286	0.000867	9.60E-10
16	79337033	rs2881665	C	T	0.16	1.5	1.01	2.22	4.22E-02	0.12	0.003754	0.000682	3.90E-08
12	112179471	rs4766897	T	C	0.29	1.36	1.01	1.83	4.22E-02	0.34	0.008059	0.000466	2.20E-68
2	204740866	rs231726	C	T	0.7	0.72	0.53	0.99	4.26E-02	0.67	-0.0087	0.00047	7.00E-77
6	31080859	rs2233966	A	G	0.46	1.35	1.01	1.81	4.26E-02	0.5	-0.00759	0.000764	7.30E-16
6	31403625	rs7759215	C	T	0.97	0.56	0.32	0.98	4.35E-02	0.87	-0.00919	0.000835	1.50E-21
6	32594470	rs114309058	G	A	0.49	1.35	1.01	1.8	4.54E-02	0.75	0.005841	0.000705	4.30E-16
17	7252148	rs2292067	G	T	0.65	0.73	0.54	0.99	4.57E-02	0.63	-0.00264	0.000461	5.70E-09
13	111206226	rs9521838	G	A	0.8	1.48	1.01	2.19	4.65E-02	0.77	0.002838	0.000525	4.80E-08
6	33047898	rs2567281	C	A	0.89	1.99	1.01	3.95	4.77E-02	0.92	0.006656	0.000861	6.70E-15
13	24789706	rs1220604	G	A	0.53	1.35	1	1.81	4.83E-02	0.55	-0.00267	0.000449	1.90E-09
9	21582326	rs7046475	T	C	0.22	0.66	0.43	1	4.87E-02	0.22	0.004	0.000537	7.60E-14
6	33048937	rs7770501	C	G	0.86	1.81	1	3.27	4.91E-02	0.88	0.006474	0.000764	8.20E-17
13	43063831	rs66749983	A	T	0.7	1.4	1	1.94	4.91E-02	0.69	-0.00361	0.000478	3.90E-14

Variant Information					MSK + VUMC Cohort Results					UK Biobank GWAS Results			
Chr	**Position (GRCh37)**	**SNP**	**Effect allele**	**Alt allele**	**Effect allele freq**	**HR**	**L95**	**U95**	**p-value**	**Effect allele freq**	**β**	**Standard error**	**p-value**
6	31917291	rs2072634	C	T	0.98	0.49	0.24	1	4.98E-02	0.98	0.01083	0.001717	1.10E-09

Chr: chromosome, SNP: single nucleotide polymorphism, Alt allele: alternate allele, freq: frequency, HR: hazard ratio, L95: lower 95% confidence interval, U95: upper 95% confidence interval

Table 6-22. PRS as a predictor of PFS in the combined MSK + VUMC cohort and OS in the MSK cohort. Time-dependent models were adjusted for age, sex, combined anti-PD-(L)1 + anti-CTLA-4 therapy, and first 10 principal components. Adjusted hazard ratios are per standard deviation of the PRS.

	PFS			**OS**		
PRS phenotype	**aHR**	**95% CI**	**p-value**	**aHR**	**95% CI**	**p-value**
Hypothyroidism	1.00	0.92-1.08	0.96	1.05	0.94-1.18	0.4
Thyroid medications	1.00	0.92-1.08	0.97	1.07	0.96-1.20	0.2

PFS: progression-free survival, OS: overall survival, aHR: adjusted hazard ratio, CI: confidence interval

REFERENCES

1. López-Campos, J. L., Tan, W. & Soriano, J. B. Global burden of COPD. *Respirology* **21**, 14–23 (2016).
2. Adeloye, D. *et al.* Global and regional estimates of COPD prevalence: Systematic review and meta–analysis. *J. Glob. Health* **5**, (2015).
3. Centers for Disease Control and Prevention. Chronic Obstructive Pulmonary Disease Among Adults — United States, 2011. *Morb. Mortal. Wkly. Rep.* **61**, 938–943 (2012).
4. Murphy, S. L., Xu, J., Kochanek, K. D. & Arias, E. Mortality in the United States, 2017. *NCHS Data Brief* 8 (2018).
5. Xu, J., Murphy, S. L., Kochanek, K. D., Bastian, B. & Arias, E. National Vital Statistics Reports Volume 67, Number 5 July 26, 2018, Deaths: Final Data for 2016. 76.
6. WHO | Burden of COPD. *WHO* https://www.who.int/respiratory/copd/burden/en/.
7. Lopez, A. D. *et al.* Chronic obstructive pulmonary disease: current burden and future projections. *Eur. Respir. J.* **27**, 397–412 (2006).
8. Mathers, C. D. & Loncar, D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.* **3**, e442 (2006).
9. Lozano, R. *et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet Lond. Engl.* **380**, 2095–2128 (2012).
10. WHO | Projections of mortality and causes of death,
2016 to 2060. *WHO* http://www.who.int/healthinfo/global_burden_disease/projections/en/.
11. Weakley, J. *et al.* Agreement between obstructive airways disease diagnoses from self-report questionnaires and medical records. *Prev. Med.* **57**, 38–42 (2013).

12. Muggah, E., Graves, E., Bennett, C. & Manuel, D. G. Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report. *BMC Public Health* **13**, 16 (2013).

13. Koller, K. R., Wilson, A. S., Asay, E. D., Metzger, J. S. & Neal, D. E. Agreement Between Self-Report and Medical Record Prevalence of 16 Chronic Conditions in the Alaska EARTH Study. *J. Prim. Care Community Health* **5**, 160–165 (2014).

14. Almagro, P. *et al.* Underdiagnosis and prognosis of chronic obstructive pulmonary disease after percutaneous coronary intervention: a prospective study. *Int. J. Chron. Obstruct. Pulmon. Dis.* **10**, 1353–1361 (2015).

15. Soriano, J. B. *et al.* High prevalence of undiagnosed airflow limitation in patients with cardiovascular disease. *Chest* **137**, 333–340 (2010).

16. Kamimura, T. *et al.* Prevalence of Previously Undiagnosed Airflow Limitation in Patients Who Underwent Preoperative Pulmonary Function Test. *Kurume Med. J.* **53**, 53–57 (2006).

17. Bérard, E. *et al.* Undiagnosed airflow limitation in patients at cardiovascular risk. *Arch. Cardiovasc. Dis.* **104**, 619–626 (2011).

18. Martinez, C. H. *et al.* Undiagnosed Obstructive Lung Disease in the United States. Associated Factors and Long-term Mortality. *Ann. Am. Thorac. Soc.* **12**, 1788–1795 (2015).

19. Zhang, J. *et al.* Prevalence of undiagnosed and undertreated chronic obstructive pulmonary disease in lung cancer population. *Respirology* **18**, 297–302 (2013).

20. Bednarek, M., Maciejewski, J., Wozniak, M., Kuca, P. & Zielinski, J. Prevalence, severity and underdiagnosis of COPD in the primary care setting. *Thorax* **63**, 402–407 (2008).

21. Miravitlles, M. *et al.* Prevalence of COPD in Spain: impact of undiagnosed COPD on quality of life and daily life activities. *Thorax* **64**, 863–868 (2009).

22. Quach, A. *et al.* Prevalence and underdiagnosis of airway obstruction among middle-aged adults in northern France: The ELISABET study 2011–2013. *Respir. Med.* **109**, 1553–1561 (2015).

23. Schirnhofer, L. *et al.* Using targeted spirometry to reduce non-diagnosed chronic obstructive pulmonary disease. *Respir. Int. Rev. Thorac. Dis.* **81**, 476–482 (2011).

24. Jensen, H. H., Godtfredsen, N. S., Lange, P. & Vestbo, J. Potential misclassification of causes of death from COPD. *Eur. Respir. J.* **28**, 781–785 (2006).

25. Drummond, M. B., Wise, R. A., John, M., Zvarich, M. T. & McGarvey, L. P. Accuracy of Death Certificates in COPD: Analysis from the TORCH Trial. *COPD J. Chronic Obstr. Pulm. Dis.* **7**, 179–185 (2010).

26. Berry, C. E. & Wise, R. A. Mortality in COPD: causes, risk factors, and prevention. *COPD* **7**, 375–382 (2010).

27. Obi, J., Mehari, A. & Gillum, R. Mortality Related to Chronic Obstructive Pulmonary Disease and Co-morbidities in the United States, A Multiple Causes of Death Analysis. *COPD J. Chronic Obstr. Pulm. Dis.* **15**, 200–205 (2018).

28. Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System Survey Data*. (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2019).

29. Tilert, T., Dillon, C., Paulose-Ram, R., Hnizdo, E. & Doney, B. Estimating the U.S. prevalence of chronic obstructive pulmonary disease using pre- and post-bronchodilator spirometry: the National Health and Nutrition Examination Survey (NHANES) 2007–2010. *Respir. Res.* **14**, 103 (2013).

30. Ford, E. S. *et al.* Trends in the Prevalence of Obstructive and Restrictive Lung Function Among Adults in the United States: Findings From the National Health and Nutrition Examination Surveys From 1988-1994 to 2007-2010. *Chest* **143**, 1395–1406 (2013).

31. Ntritsos, G. *et al.* Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 1507–1514 (2018).

32. Eisner, M. D. *et al.* An Official American Thoracic Society Public Policy Statement: Novel Risk Factors and the Global Burden of Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **182**, 693–718 (2010).

33. Forey, B. A., Thornton, A. J. & Lee, P. N. Systematic review with meta-analysis of the epidemiological evidence relating smoking to COPD, chronic bronchitis and emphysema. *BMC Pulm. Med.* **11**, 36 (2011).

34. Liu, Y. *et al.* Smoking duration, respiratory symptoms, and COPD in adults aged ≥45 years with a smoking history. *Int. J. Chron. Obstruct. Pulmon. Dis.* (2015).

35. Hooper, R. *et al.* Risk factors for COPD spirometrically defined from the lower limit of normal in the BOLD project. *Eur. Respir. J.* **39**, 1343–1353 (2012).

36. Doney, B. *et al.* Occupational Risk Factors for COPD Phenotypes in the Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study. *COPD J. Chronic Obstr. Pulm. Dis.* **11**, 368–380 (2014).

37. Blanc, P. D. *et al.* Occupational exposures and the risk of COPD: dusty trades revisited. *Thorax* **64**, 6–12 (2009).

38. Waked, M., Salameh, Khayat & Salameh, P. Correlates of COPD and chronic bronchitis in nonsmokers: data from a cross-sectional study. *Int. J. Chron. Obstruct. Pulmon. Dis.* (2012).

39. Yeh, J.-J., Wang, Y.-C., Sung, F.-C., Chou, C. Y.-T. & Kao, C.-H. Nontuberculosis Mycobacterium Disease is a Risk Factor for Chronic Obstructive Pulmonary Disease: A Nationwide Cohort Study. *Lung* **192**, 403–411 (2014).

40. de Marco, R. *et al.* Risk Factors for Chronic Obstructive Pulmonary Disease in a European Cohort of Young Adults. *Am. J. Respir. Crit. Care Med.* **183**, 891–897 (2011).

41. Sexton, P. *et al.* Chronic Obstructive Pulmonary Disease in Non-smokers: A Case-Comparison Study. *COPD J. Chronic Obstr. Pulm. Dis.* **11**, 2–9 (2014).

42. Scanlon, P. D. Respiratory Testing and Function. in *Goldman-Cecil Medicine* 518-523.e9 (Elsevier, 2020).

43. Han, M. K. & Lazarus, S. C. COPD: Clinical Diagnosis and Management. in *Murray and Nadel's Textbook of Respiratory Medicine* 767-785.e7 (Elsevier/Saunders, 2016).

44. Niewoehner, D. E. Chronic Obstructive Pulmonary Disease. in *Goldman-Cecil Medicine* 555-562.e3 (Elsevier/Saunders, 2016).

45. Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: 2020 Report.

46. Krishnan, J. & Martinez, F. Lung function trajectories and chronic obstructive pulmonary disease: current understanding and knowledge gaps. *Curr. Opin. Pulm. Med.* **24**, 124–129 (2018).

47. Lange, P. *et al.* Lung-Function Trajectories Leading to Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* **373**, 111–122 (2015).

48. Berry, C. E. *et al.* A Distinct Low Lung Function Trajectory from Childhood to the Fourth Decade of Life. *Am. J. Respir. Crit. Care Med.* **194**, 607–612 (2016).

49. Gauderman, W. J. *et al.* The Effect of Air Pollution on Lung Development from 10 to 18 Years of Age. *N. Engl. J. Med.* **351**, 1057–1067 (2004).

50. Sears, M. R. *et al.* A Longitudinal, Population-Based, Cohort Study of Childhood Asthma Followed to Adulthood. *N. Engl. J. Med.* **349**, 1414–1422 (2003).

51. Allinson, J. P. *et al.* Combined Impact of Smoking and Early-Life Exposures on Adult Lung Function Trajectories. *Am. J. Respir. Crit. Care Med.* **196**, 1021–1030 (2017).

52. Agustí, A., Noell, G., Brugada, J. & Faner, R. Lung function in early adulthood and health in later life: a transgenerational cohort analysis. *Lancet Respir. Med.* **5**, 935–945 (2017).

53. Silverman, E. K. Genetics of COPD. *Annu. Rev. Physiol.* **82**, 413–431 (2020).

54. Köhnlein, T. & Welte, T. Alpha-1 Antitrypsin Deficiency: Pathogenesis, Clinical Presentation, Diagnosis, and Treatment. *Am. J. Med.* **121**, 3–9 (2008).

55. Kalfopoulos, M., Wetmore, K. & K. ElMallah, M. Pathophysiology of Alpha-1 Antitrypsin Lung Disease. in *Alpha-1 Antitrypsin Deficiency : Methods and Protocols* (eds. Borel, F. & Mueller, C.) 9–19 (Springer, 2017). doi:10.1007/978-1-4939-7163-3_2.

56. Seixas, S. & Marques, P. I. Known Mutations at the Cause of Alpha-1 Antitrypsin Deficiency an Updated Overview of SERPINA1 Variation Spectrum. *Appl. Clin. Genet.* **14**, 173–194 (2021).

57. de Serres, F. J. & Blanco, I. Prevalence of α1-antitrypsin deficiency alleles PI*S and PI*Z worldwide and effective screening for each of the five phenotypic classes PI*MS, PI*MZ, PI*SS, PI*SZ, and PI*ZZ: a comprehensive review. *Ther. Adv. Respir. Dis.* **6**, 277–295 (2012).

58. Silverman, E. K. & Sandhaus, R. A. Alpha1-Antitrypsin Deficiency. *N. Engl. J. Med.* **360**, 2749–2757 (2009).

59. Silverman, E. K. *et al.* Genetic Epidemiology of Severe, Early-onset Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **157**, 1770–1778 (1998).

60. Hersh, C. P. *et al.* Family History Is a Risk Factor for COPD. *Chest* **140**, 343–350 (2011).

61. McCloskey, S. C. *et al.* Siblings of Patients With Severe Chronic Obstructive Pulmonary Disease Have a Significant Risk of Airflow Obstruction. *Am. J. Respir. Crit. Care Med.* **164**, 1419–1424 (2001).

62. Zöller, B., Li, X., Sundquist, J. & Sundquist, K. Familial transmission of chronic obstructive pulmonary disease in adoptees: a Swedish nationwide family study. *BMJ Open* **5**, (2015).

63. Ingebrigtsen, T. *et al.* Genetic influences on chronic obstructive pulmonary disease – A twin study. *Respir. Med.* **104**, 1890–1895 (2010).

64. Gim, J. *et al.* A Between Ethnicities Comparison of Chronic Obstructive Pulmonary Disease Genetic Risk. *Front. Genet.* **11**, (2020).

65. Zhou, J. J. *et al.* Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am. J. Respir. Crit. Care Med.* **188**, 941–947 (2013).

66. Regan, E. A. *et al.* Genetic Epidemiology of COPD (COPDGene) Study Design. *COPD J. Chronic Obstr. Pulm. Dis.* **7**, 32–43 (2011).

67. Couper, D. *et al.* Design of the Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS). *Thorax* **69**, 492–495 (2014).

68. Vestbo, J. *et al.* Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur. Respir. J.* **31**, 869–873 (2008).

69. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am. J. Epidemiol.* **129**, 687–702 (1989).

70. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: Objectives and Design. *Am. J. Epidemiol.* **156**, 871–881 (2002).

71. Friedman, G. D. *et al.* Cardia: study design, recruitment, and some characteristics of the examined subjects. *J. Clin. Epidemiol.* **41**, 1105–1116 (1988).

72. Sakornsakolpat, P. *et al.* Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.* **51**, 494–505 (2019).

73. Pillai, S. G. *et al.* A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.* **5**, e1000421 (2009).

74. Cho, M. H. *et al.* Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat. Genet.* **42**, 200–202 (2010).

75. Cho, M. H. *et al.* Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir. Med.* **2**, 214–225 (2014).

76. Hobbs, B. D. *et al.* Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.* **49**, 426–432 (2017).

77. van Durme, Y. M. T. A. *et al.* Prevalence, incidence, and lifetime risk for the development of copd in the elderly: The rotterdam study. *Chest* **135**, 368–377 (2009).

78. Zhou, X. *et al.* Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Hum. Mol. Genet.* **21**, 1325–1335 (2012).

79. 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. *Lancet Respir. Med.* **3**, 769–781 (2015).

80. Xie, J. *et al.* Gene susceptibility identification in a longitudinal study confirms new loci in the development of chronic obstructive pulmonary disease and influences lung function decline. *Respir. Res.* **16**, (2015).

81. Ziółkowska-Suchanek, I. *et al.* Susceptibility loci in lung cancer and COPD: association of IREB2 and FAM13A with pulmonary diseases. *Sci. Rep.* **5**, 13502 (2015).

82. Wilk, J. B. *et al.* A Genome-Wide Association Study of Pulmonary Function Measures in the Framingham Heart Study. *PLOS Genet.* **5**, e1000429 (2009).

83. Hancock, D. B. *et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* **42**, 45–52 (2010).

84. Repapi, E. *et al.* Genome-wide association study identifies five loci associated with lung function. *Nat. Genet.* **42**, 36–44 (2010).

85. Wain, L. V. *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat. Genet.* **49**, 416–425 (2017).

86. Wyss, A. B. *et al.* Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. *Nat. Commun.* **9**, 2976 (2018).

87. Shrine, N. *et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493 (2019).

88. Seo, M. *et al.* Genomics and response to long-term oxygen therapy in chronic obstructive pulmonary disease. *J. Mol. Med. Berl. Ger.* **96**, 1375–1385 (2018).

89. Condreay, L. D., Gao, C., Bradford, E., Yancey, S. W. & Ghosh, S. No genetic associations with mepolizumab efficacy in COPD with peripheral blood eosinophilia. *Respir. Med.* **155**, 26–28 (2019).

90. Condreay, L. D., Qu, X. A., Anderson, J., Compton, C. & Ghosh, S. Genetic effects on efficacy to fluticasone propionate/salmeterol treatment in COPD. *Respir. Med.* **155**, 51–53 (2019).

91. Hardin, M. *et al.* A genome-wide analysis of the response to inhaled β2-agonists in chronic obstructive pulmonary disease. *Pharmacogenomics J.* **16**, 326–335 (2016).

92. Obeidat, M. *et al.* The pharmacogenomics of inhaled corticosteroids and lung function decline in COPD. *Eur. Respir. J.* **54**, (2019).

93. Hansel, N. N. *et al.* Genome-Wide Association Study Identification of Novel Loci Associated with Airway Responsiveness in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Cell Mol. Biol.* **53**, 226–234 (2015).

94. Lee, J. H. *et al.* IREB2 and GALC are associated with pulmonary artery enlargement in chronic obstructive pulmonary disease. *Am. J. Respir. Cell Mol. Biol.* **52**, 365–376 (2015).

95. Manichaikul, A. *et al.* Genome-wide study of percent emphysema on computed tomography in the general population. The Multi-Ethnic Study of Atherosclerosis Lung/SNP Health Association Resource Study. *Am. J. Respir. Crit. Care Med.* **189**, 408–418 (2014).

96. Kong, X. *et al.* Genome-wide Association Study Identifies BICD1 as a Susceptibility Gene for Emphysema. *Am. J. Respir. Crit. Care Med.* **183**, 43–49 (2011).

97. Ejike, C. O. *et al.* Chronic Obstructive Pulmonary Disease in America's Black Population. *Am. J. Respir. Crit. Care Med.* **200**, 423–430 (2019).

98. Lutz, S. M. *et al.* A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet.* **16**, 138 (2015).

99. Moll, M. *et al.* Chronic obstructive pulmonary disease and related phenotypes: polygenic risk scores in population-based and case-control cohorts. *Lancet Respir. Med.* **8**, 696–708 (2020).

100. Putcha, N., Drummond, M. B., Wise, R. A. & Hansel, N. N. Comorbidities and Chronic Obstructive Pulmonary Disease: Prevalence, Influence on Outcomes, and Management. *Semin. Respir. Crit. Care Med.* **36**, 575–591 (2015).

101. Baty, F., Putora, P. M., Isenring, B., Blum, T. & Brutsche, M. Comorbidities and Burden of COPD: A Population Based Case-Control Study. *PLOS ONE* **8**, e63285 (2013).

102. Yin, H.-L. *et al.* Prevalence of comorbidities in chronic obstructive pulmonary disease patients: A meta-analysis. *Medicine (Baltimore)* **96**, e6836 (2017).

103. Barnes, P. J. & Celli, B. R. Systemic manifestations and comorbidities of COPD. *Eur. Respir. J.* **33**, 1165–1185 (2009).

104. Cavaillès, A. *et al.* Comorbidities of COPD. *Eur. Respir. Rev.* **22**, 454–475 (2013).

105. Divo, M. J. *et al.* Chronic Obstructive Pulmonary Disease (COPD) as a disease of early aging: Evidence from the EpiChron Cohort. *PloS One* **13**, e0193143 (2018).

106. Laforest, L. *et al.* Frequency of comorbidities in chronic obstructive pulmonary disease, and impact on all-cause mortality: A population-based cohort study. *Respir. Med.* **117**, 33–39 (2016).

107. Wheaton, A. G., Ford, E. S., Cunningham, T. J. & Croft, J. B. Chronic Obstructive Pulmonary Disease, Hospital Visits, and Comorbidities: National Survey of Residential Care Facilities, 2010. *J. Aging Health* **27**, 480–499 (2015).

108. Divo, M. J. *et al.* COPD comorbidities network. *Eur. Respir. J.* **46**, 640–650 (2015).

109. Putcha, N., Puhan, M. A., Hansel, N. N., Drummond, M. B. & Boyd, C. M. Impact of co-morbidities on self-rated health in self-reported COPD: An analysis of NHANES 2001–2008. *COPD* **10**, 324–332 (2013).

110. López Varela, M. V. *et al.* Comorbidities and Health Status in Individuals With and Without COPD in Five Latin American Cities: The PLATINO Study. *Arch. Bronconeumol. Engl. Ed.* **49**, 468–474 (2013).

111. Koskela, J. *et al.* Co-morbidities are the key nominators of the health related quality of life in mild and moderate COPD. *BMC Pulm. Med.* **14**, 102 (2014).

112. Frei, A. *et al.* Five comorbidities reflected the health status in patients with chronic obstructive pulmonary disease: the newly developed COMCOLD index. *J. Clin. Epidemiol.* **67**, 904–911 (2014).

113. Maselli, D. J. *et al.* Clinical Epidemiology of COPD: Insights From 10 Years of the COPDGene Study. *Chest* **156**, 228–238 (2019).

114. Kahnert, K. *et al.* The revised GOLD 2017 COPD categorization in relation to comorbidities. *Respir. Med.* **134**, 79–85 (2018).

115. Miller, J. *et al.* Comorbidity, systemic inflammation and outcomes in the ECLIPSE cohort. *Respir. Med.* **107**, 1376–1384 (2013).

116. Rothnie, K. J. *et al.* Closing the mortality gap after a myocardial infarction in people with and without chronic obstructive pulmonary disease. *Heart* **101**, 1103–1110 (2015).

117. Axson, E. L. *et al.* Hospitalisation and mortality outcomes of patients with comorbid COPD and heart failure: a systematic review protocol. *BMJ Open* **8**, e023058 (2018).

118. Șerban, R. C. *et al.* Impact of chronic obstructive pulmonary disease on in-hospital morbidity and mortality in patients with ST-segment elevation myocardial infarction treated by primary percutaneous coronary intervention. *Int. J. Cardiol.* **243**, 437–442 (2017).

119. Zhang, M. *et al.* Impact of Chronic Obstructive Pulmonary Disease on Long-Term Outcome in Patients with Coronary Artery Disease Undergoing Percutaneous Coronary Intervention. *BioMed Res. Int.* **2016**, e8212459 (2016).

120. Navaneethan, S. D. *et al.* Mortality Outcomes of Patients with Chronic Kidney Disease and Chronic Obstructive Pulmonary Disease. *Am. J. Nephrol.* **43**, 39–46 (2016).

121. Du, W. *et al.* Obstructive sleep apnea, COPD, the overlap syndrome, and mortality: results from the 2005-2008 National Health and Nutrition Examination Survey. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 665–674 (2018).

122. Rothnie, K. J., Yan, R., Smeeth, L. & Quint, J. K. Risk of myocardial infarction (MI) and death following MI in people with chronic obstructive pulmonary disease (COPD): a systematic review and meta-analysis. *BMJ Open* **5**, e007824 (2015).

123. Liao, Y.-B. *et al.* The relationship between chronic obstructive pulmonary disease and transcatheter aortic valve implantation—A systematic review and meta-analysis. *Catheter. Cardiovasc. Interv.* **87**, 570–578 (2016).

124. Yoshida, Y. *et al.* Worse Prognosis for Stage IA Lung Cancer Patients with Smoking History and More Severe Chronic Obstructive Pulmonary Disease. *Ann. Thorac. Cardiovasc. Surg.* **21**, 194–200 (2015).

125. Tao, H., Onoda, H., Okabe, K. & Matsumoto, T. The impact of coexisting lung diseases on outcomes in patients with pathological Stage I non-small-cell lung cancer. *Interact. Cardiovasc. Thorac. Surg.* **26**, 1009–1015 (2018).

126. Zhai, R., Yu, X., Shafer, A., Wain, J. C. & Christiani, D. C. The impact of coexisting COPD on survival of patients with early-stage non-small cell lung cancer undergoing surgical resection. *Chest* **145**, 346–353 (2014).

127. Schroedl, C. & Kalhan, R. Incidence, treatment options, and outcomes of lung cancer in patients with chronic obstructive pulmonary disease. *Curr. Opin. Pulm. Med.* **18**, 131–137 (2012).

128. Lim, J. U. *et al.* Overall survival of driver mutation-negative non-small cell lung cancer patients with COPD under chemotherapy compared to non-COPD non-small cell lung cancer patients. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 2139–2146 (2018).

129. Newsome, B. R., McDonnell, K., Hucks, J. & Dawson Estrada, R. Chronic Obstructive Pulmonary Disease: Clinical Implications for Patients With Lung Cancer. *Clin. J. Oncol. Nurs.* **22**, 184–192 (2018).

130. Tan, L.-E., A M, R. & Lim, C.-S. Association of chronic obstructive pulmonary disease and postresection lung cancer survival: a systematic review and meta-analysis. *J. Investig. Med. Off. Publ. Am. Fed. Clin. Res.* **65**, 342–352 (2017).

131. Lin, H., Lu, Y., Lin, L., Meng, K. & Fan, J. Does chronic obstructive pulmonary disease relate to poor prognosis in patients with lung cancer?: A meta-analysis. *Medicine (Baltimore)* **98**, e14837 (2019).

132. Gao, Y. *et al.* Impact of COPD and emphysema on survival of patients with lung cancer: A meta-analysis of observational studies. *Respirology* **21**, 269–279 (2016).

133. Wang, W. *et al.* Impact of COPD on prognosis of lung cancer: from a perspective on disease heterogeneity. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 3767–3776 (2018).

134. Almagro, P. *et al.* Insights into Chronic Obstructive Pulmonary Disease as Critical Risk Factor for Cardiovascular Disease. *Int. J. Chron. Obstruct. Pulmon. Dis.* **15**, 755–764 (2020).

135. Sinden, N. J. & Stockley, R. A. Systemic inflammation and comorbidity in COPD: a result of 'overspill' of inflammatory mediators from the lungs? Review of the evidence. *Thorax* **65**, 930–936 (2010).

136. Agustí, A. & Faner, R. Systemic Inflammation and Comorbidities in Chronic Obstructive Pulmonary Disease. *Proc. Am. Thorac. Soc.* **9**, 43–46 (2012).

137. Thomsen, M., Dahl, M., Lange, P., Vestbo, J. & Nordestgaard, B. G. Inflammatory Biomarkers and Comorbidities in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **186**, 982–988 (2012).

138. Cunningham, T. J., Ford, E. S., Rolle, I. V., Wheaton, A. G. & Croft, J. B. Associations of Self-Reported Cigarette Smoking with Chronic Obstructive Pulmonary Disease and Co-Morbid Chronic Conditions in the United States. *COPD J. Chronic Obstr. Pulm. Dis.* **12**, 281–291 (2015).

139. Golpe, R. *et al.* Prevalence of Major Comorbidities in Chronic Obstructive Pulmonary Disease Caused by Biomass Smoke or Tobacco. *Respir. Int. Rev. Thorac. Dis.* **94**, 38–44 (2017).

140. Agustí, A. & Faner, R. COPD beyond smoking: new paradigm, novel opportunities. *Lancet Respir. Med.* **6**, 324–326 (2018).

141. Zhu, Z. *et al.* Genetic overlap of chronic obstructive pulmonary disease and cardiovascular disease-related traits: a large-scale genome-wide cross-trait analysis. *Respir. Res.* **20**, 64 (2019).

142. Young, R. P. *et al.* Genetic evidence linking lung cancer and COPD: a new perspective. *Appl. Clin. Genet.* **4**, 99–111 (2011).

143. Byun, J. *et al.* The Shared Genetic Architectures Between Lung Cancer and Multiple Polygenic Phenotypes in Genome-Wide Association Studies. *Cancer Epidemiol. Biomarkers Prev.* cebp;1055-9965.EPI-20-1635v2 (2021) doi:10.1158/1055-9965.EPI-20-1635.

144. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **n/a**,.

145. National Cancer Institute. SEER Cancer Stat Facts: Lung and Bronchus Cancer. *SEER* https://seer.cancer.gov/statfacts/html/lungb.html.

146. U.S. Cancer Statistics Working Group. U.S. Cancer Statistics Data Visualizations Tool. https://gis.cdc.gov/grasp/USCS/DataViz.html.

147. Centers for Disease Control and Prevention. Map of Cigarette Use Among Adults | STATE System | CDC. https://www.cdc.gov/statesystem/cigaretteuseadult.html (2019).

148. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA. Cancer J. Clin.* **69**, 7–34 (2019).

149. Lu, T. *et al.* Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. *Cancer Manag. Res.* **11**, 943–953 (2019).

150. Harris, J. E. Cigarette smoking among successive birth cohorts of men and women in the United States during 1900-80. *J. Natl. Cancer Inst.* **71**, 473–479 (1983).

151. Jones, C. C. *et al.* Racial Disparities in Lung Cancer Survival: The Contribution of Stage, Treatment, and Ancestry. *J. Thorac. Oncol.* **13**, 1464–1473 (2018).

152. Ganti, A. K. *et al.* Association Between Race and Survival of Patients With Non–Small-Cell Lung Cancer in the United States Veterans Affairs Population. *Clin. Lung Cancer* **15**, 152–158 (2014).

153. Zheng, L. *et al.* Lung Cancer Survival among Black and White Patients in an Equal Access Health System. *Cancer Epidemiol. Prev. Biomark.* **21**, 1841–1847 (2012).

154. Aldrich, M. C., Grogan, E. L., Munro, H. M., Signorello, L. B. & Blot, W. J. Stage-Adjusted Lung Cancer Survival Does Not Differ between Low-Income Blacks and Whites. *J. Thorac. Oncol.* **8**, 1248–1254 (2013).

155. Gillaspie, E. A., Cass, A. S., Lewis, J. & Horn, L. Lung Cancer. in *Conn's Current Therapy 2021* 878–889 (Elsevier, 2021).

156. Rodriguez-Canales, J., Parra-Cuentas, E. & Wistuba, I. I. Diagnosis and Molecular Classification of Lung Cancer. *Lung Cancer* 25–46 (2016) doi:10.1007/978-3-319-40389-2_2.

157. Zheng, M. Classification and Pathology of Lung Cancer. *Surg. Oncol. Clin. N. Am.* **25**, 447–468 (2016).

158. Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553**, 446–454 (2018).

159. Meza, R., Meernik, C., Jeon, J. & Cote, M. L. Lung Cancer Incidence Trends by Gender, Race and Histology in the United States, 1973–2010. *PLoS ONE* **10**, (2015).

160. Howlader, N. *et al.* The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *N. Engl. J. Med.* **383**, 640–649 (2020).

161. Onoi, K. *et al.* Immune Checkpoint Inhibitors for Lung Cancer Treatment: A Review. *J. Clin. Med.* **9**, 1362 (2020).

162. Qu, J. *et al.* The progress and challenge of anti-PD-1/PD-L1 immunotherapy in treating non-small cell lung cancer. *Ther. Adv. Med. Oncol.* **13**, 1758835921992968 (2021).

163. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).

164. Borghaei, H. *et al.* Nivolumab versus Docetaxel in Advanced Nonsquamous Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **373**, 1627–1639 (2015).

165. Brahmer, J. *et al.* Nivolumab versus Docetaxel in Advanced Squamous-Cell Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **373**, 123–135 (2015).

166. Hellmann, M. D. *et al.* Nivolumab plus Ipilimumab in Advanced Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* (2019) doi:10.1056/NEJMoa1910231.

167. Hellmann, M. D. *et al.* Nivolumab plus ipilimumab as first-line treatment for advanced non-small-cell lung cancer (CheckMate 012): results of an open-label, phase 1, multicohort study. *Lancet Oncol.* **18**, 31–41 (2017).

168. Ready, N. E. *et al.* Nivolumab Monotherapy and Nivolumab Plus Ipilimumab in Recurrent Small Cell Lung Cancer: Results From the CheckMate 032 Randomized Cohort. *J. Thorac. Oncol.* **15**, 426–435 (2020).

169. Reck, M. *et al.* Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **375**, 1823–1833 (2016).

170. Gandhi, L. *et al.* Pembrolizumab plus Chemotherapy in Metastatic Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **378**, 2078–2092 (2018).

171. Paz-Ares, L. *et al.* Pembrolizumab plus Chemotherapy for Squamous Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **379**, 2040–2051 (2018).

172. Rittmeyer, A. *et al.* Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *The Lancet* **389**, 255–265 (2017).

173. Socinski, M. A. *et al.* Atezolizumab for First-Line Treatment of Metastatic Nonsquamous NSCLC. *N. Engl. J. Med.* **378**, 2288–2301 (2018).

174. Ramos-Casals, M. *et al.* Immune-related adverse events of checkpoint inhibitors. *Nat. Rev. Dis. Primer* **6**, 1–21 (2020).

175. Darnell, E. P., Mooradian, M. J., Baruch, E. N., Yilmaz, M. & Reynolds, K. L. Immune-Related Adverse Events (irAEs): Diagnosis, Management, and Clinical Pearls. *Curr. Oncol. Rep.* **22**, 1–11 (2020).

176. Das, S. & Johnson, D. B. Immune-related adverse events and anti-tumor efficacy of immune checkpoint inhibitors. *J. Immunother. Cancer* **7**, 306 (2019).

177. Zhong, L., Wu, Q., Chen, F., Liu, J. & Xie, X. Immune-related adverse events: promising predictors for efficacy of immune checkpoint inhibitors. *Cancer Immunol. Immunother.* 1–18 (2021) doi:10.1007/s00262-020-02803-5.

178. Zhou, X. *et al.* Are immune-related adverse events associated with the efficacy of immune checkpoint inhibitors in patients with cancer? A systematic review and meta-analysis. *BMC Med.* **18**, 1–14 (2020).

179. Toi, Y. *et al.* Profiling Preexisting Antibodies in Patients Treated With Anti-PD-1 Therapy for Advanced Non-Small Cell Lung Cancer. *JAMA Oncol.* **5**, 376–383 (2019).

180. Peng, L. *et al.* Peripheral blood markers predictive of outcome and immune-related adverse events in advanced non-small cell lung cancer treated with PD-1 inhibitors. *Cancer Immunol. Immunother. CII* **69**, 1813–1822 (2020).

181. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, (2019).

182. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

183. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

184. Cooke Bailey, J. N. & Igo, R. P. Genetic Risk Scores. *Curr. Protoc. Hum. Genet.* **91**, 1.29.1-1.29.9 (2016).

185. Dudbridge, F. Polygenic Epidemiology. *Genet. Epidemiol.* **40**, 268–272 (2016).

186. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).

187. Mistry, S., Harrison, J. R., Smith, D. J., Escott-Price, V. & Zammit, S. The use of polygenic risk scores to identify phenotypes associated with genetic risk of bipolar disorder and depression: A systematic review. *J. Affect. Disord.* **234**, 148–155 (2018).

188. Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol. Psychiatry* **19**, 1017–1024 (2014).

189. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet Lond. Engl.* **381**, 1371–1379 (2013).

190. Whalley, H. C. *et al.* Dissection of major depressive disorder using polygenic risk scores for schizophrenia in two independent cohorts. *Transl. Psychiatry* **6**, e938 (2016).

191. Guo, W. *et al.* Polygenic risk score and heritability estimates reveals a genetic relationship between ASD and OCD. *Eur. Neuropsychopharmacol.* **27**, 657–666 (2017).

192. Takahashi, N. *et al.* Polygenic risk score analysis revealed shared genetic background in attention deficit hyperactivity disorder and narcolepsy. *Transl. Psychiatry* **10**, 284 (2020).

193. Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705-1715.e16 (2018).

194. Graff, R. E. *et al.* Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat. Commun.* **12**, 970 (2021).

195. Klarin, D. *et al.* Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nat. Genet.* **51**, 1574–1579 (2019).

196. Lareau, C. A. *et al.* Polygenic risk assessment reveals pleiotropy between sarcoidosis and inflammatory disorders in the context of genetic ancestry. *Genes Immun.* **18**, 88–94 (2017).

197. Stringer, S., Kahn, R. S., de Witte, L. D., Ophoff, R. A. & Derks, E. M. Genetic liability for schizophrenia predicts risk of immune disorders. *Schizophr. Res.* **159**, 347–352 (2014).

198. Wassertheil-Smoller, S. *et al.* Polygenic Risk for Depression Increases Risk of Ischemic Stroke: From the Stroke Genetics Network Study. *Stroke* **49**, 543–548 (2018).

199. Fanelli, G. *et al.* Polygenic risk scores for psychiatric, inflammatory, and cardio-metabolic traits and diseases highlight possible genetic overlaps with suicide attempt and treatment-emergent suicidal ideation. *medRxiv* 2021.03.08.21253145 (2021) doi:10.1101/2021.03.08.21253145.

200. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE* **12**, e0175508 (2017).

201. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).

202. Roden, D. M. Phenome-wide association studies: a new method for functional genomics in humans. *J. Physiol.* **595**, 4109–4115 (2017).

203. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu. Rev. Genomics Hum. Genet.* **17**, 353–373 (2016).

204. Pendergrass, S. A. & Ritchie, M. D. Phenome-Wide Association Studies: Leveraging Comprehensive Phenotypic and Genotypic Data for Discovery. *Curr. Genet. Med. Rep.* **3**, 92–100 (2015).

205. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17**, 129–145 (2016).

206. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

207. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).

208. Dashti, H. S., Redline, S. & Saxena, R. Polygenic risk score identifies associations between sleep duration and diseases determined from an electronic medical record biobank. *Sleep* **42**, (2019).

209. Docherty, A. R. *et al.* Polygenic prediction of the phenome, across ancestry, in emerging adulthood. *Psychol. Med.* **48**, 1814–1823 (2018).

210. Kember, R. L. *et al.* Polygenic Risk of Psychiatric Disorders Exhibits Cross-trait Associations in Electronic Health Record Data From European Ancestry Individuals. *Biol. Psychiatry* **89**, 236–245 (2021).

211. Fritsche, L. G. *et al.* Exploring various polygenic risk scores for skin cancer in the phenomes of the Michigan genomics initiative and the UK Biobank with a visual catalog: PRSWeb. *PLOS Genet.* **15**, e1008202 (2019).

212. Fritsche, L. G. *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* **102**, 1048–1061 (2018).

213. Fritsche, L. G. *et al.* Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. *Am. J. Hum. Genet.* **107**, 815–836 (2020).

214. Joo, Y. Y. *et al.* A Polygenic and Phenotypic Risk Prediction for Polycystic Ovary Syndrome Evaluated by Phenome-Wide Association Studies. *J. Clin. Endocrinol. Metab.* **105**, (2020).

215. Krapohl, E. *et al.* Phenome-wide analysis of genome-wide polygenic scores. *Mol. Psychiatry* **21**, 1188–1193 (2016).

216. Leppert, B. *et al.* A cross-disorder PRS-pheWAS of 5 major psychiatric disorders in UK Biobank. *PLoS Genet.* **16**, e1008185 (2020).

217. Zheutlin, A. B. *et al.* Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four Health Care Systems. *Am. J. Psychiatry* **176**, 846–855 (2019).

218. Criner, R. N. & Han, M. K. COPD Care in the 21st Century: A Public Health Priority. *Respir. Care* **63**, 591–600 (2018).

219. National Institutes of Health. COPD National Action Plan. 68 (2017).

220. Centers for Disease Control and Prevention. Public Health Strategic Framework for COPD Prevention. https://www.cdc.gov/copd/resources.htm.

221. Pikoula, M. *et al.* Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. *BMC Med. Inform. Decis. Mak.* **19**, 86 (2019).

222. Vazquez Guillamet, R., Ursu, O., Iwamoto, G., Moseley, P. L. & Oprea, T. Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records. *Health Informatics J.* 1460458216675661 (2016) doi:10.1177/1460458216675661.

223. Oh, S. S. *et al.* Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. *PLOS Med.* **12**, e1001918 (2015).

224. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

225. Dransfield, M. T., Davis, J. J., Gerald, L. B. & Bailey, W. C. Racial and gender differences in susceptibility to tobacco smoke among patients with chronic obstructive pulmonary disease. *Respir. Med.* **100**, 1110–1116 (2006).

226. Chatila, W. M., Wynkoop, W. A., Vance, G. & Criner, G. J. Smoking Patterns in African Americans and Whites With Advanced COPD. *Chest* **125**, 15–21 (2004).

227. Chatila, W. M., Hoffman, E. A., Gaughan, J., Robinswood, G. B. & Criner, G. J. Advanced Emphysema in African-American and White Patients: Do Differences Exist? *Chest* **130**, 108–118 (2006).

228. Hankinson, J. L., Odencrantz, J. R. & Fedan, K. B. Spirometric Reference Values from a Sample of the General U.S. Population. *Am. J. Respir. Crit. Care Med.* **159**, 179–187 (1999).

229. Harik-Khan, R. I., Fleg, J. L., Muller, D. C. & Wise, R. A. The effect of anthropometric and socioeconomic factors on the racial difference in lung function. *Am. J. Respir. Crit. Care Med.* **164**, 1647–1654 (2001).

230. Aldrich, M. C. *et al.* Genetic ancestry-smoking interactions and lung function in African Americans: a cohort study. *PloS One* **7**, e39541 (2012).

231. Kumar, R. *et al.* Genetic Ancestry in Lung-Function Predictions. *N. Engl. J. Med.* **363**, 321–330 (2010).

232. Tsai, C.-L. & Carmago, C. A. C. Racial and Ethnic Differences in Emergency Care for Acute Exacerbation of Chronic Obstructive Pulmonary Disease. *Acad. Emerg. Med.* **16**, 108–115 (2009).

233. Vaughan Sarrazin, M., Cannon, K. T., Rosenthal, G. E. & Kaldjian, L. C. Racial Differences in Mortality Among Veterans Hospitalized for Exacerbation of Chronic Obstructive Pulmonary Disease. *J. Natl. Med. Assoc.* **101**, 656–662 (2009).

234. Han, M. K. *et al.* Racial Differences in Quality of Life in Patients With COPD. *Chest* **140**, 1169–1176 (2011).

235. Putcha, N. *et al.* Comorbidities of COPD Have a Major Impact on Clinical Outcomes, Particularly in African Americans. *Chronic Obstr. Pulm. Dis. J. COPD Found.* **1**, 105–114 (2014).

236. Atlantis, E., Fahey, P., Cochrane, B. & Smith, S. Bidirectional Associations Between Clinically Relevant Depression or Anxiety and COPD: A Systematic Review and Meta-analysis. *Chest* **144**, 766–777 (2013).

237. Pumar, M. I. *et al.* Anxiety and depression—Important psychological comorbidities of COPD. *J. Thorac. Dis.* **6**, 1615–1631 (2014).

238. Pelgrim, C. E. *et al.* Psychological co-morbidities in COPD: Targeting systemic inflammation, a benefit for both? *Eur. J. Pharmacol.* **842**, 99–110 (2019).

239. Matte, D. L. *et al.* Prevalence of depression in COPD: A systematic review and meta-analysis of controlled studies. *Respir. Med.* **117**, 154–161 (2016).

240. Yohannes, A. M. & Alexopoulos, G. S. Depression and anxiety in patients with COPD. *Eur. Respir. Rev.* **23**, 345–349 (2014).

241. Yohannes, A. M., Willgoss, T. G., Baldwin, R. C. & Connolly, M. J. Depression and anxiety in chronic heart failure and chronic obstructive pulmonary disease: prevalence, relevance, clinical implications and management principles. *Int. J. Geriatr. Psychiatry* **25**, 1209–1221 (2010).

242. Montserrat-Capdevila, J. *et al.* Overview of the Impact of Depression and Anxiety in Chronic Obstructive Pulmonary Disease. *Lung* **195**, 77–85 (2017).

243. Alqahtani, J. S. *et al.* Risk factors for all-cause hospital readmission following exacerbation of COPD: a systematic review and meta-analysis. *Eur. Respir. Rev. Off. J. Eur. Respir. Soc.* **29**, (2020).

244. Paine, N. J. *et al.* Psychological distress is related to poor health behaviours in COPD and non-COPD patients: Evidence from the CanCOLD study. *Respir. Med.* **146**, 1–9 (2019).

245. Jang, S. M. *et al.* Depression is a major determinant of both disease-specific and generic health-related quality of life in people with severe COPD: *Chron. Respir. Dis.* (2018) doi:10.1177/1479972318775422.

246. Miravitlles, M. & Ribera, A. Understanding the impact of symptoms on the burden of COPD. *Respir. Res.* **18**, 67 (2017).

247. Montserrat-Capdevila, J. *et al.* Mental disorders in chronic obstructive pulmonary diseases. *Perspect. Psychiatr. Care* **54**, 398–404 (2018).

248. Lu, Y. *et al.* Systemic inflammation, depression and obstructive pulmonary function: a population-based study. *Respir. Res.* **14**, 1–8 (2013).

249. Al-shair, K. *et al.* Biomarkers of systemic inflammation and depression and fatigue in moderate clinically stable COPD. *Respir. Res.* **12**, 1–6 (2011).

250. Riblet, N. B., Gottlieb, D. J., Hoyt, J. E., Watts, B. V. & Shiner, B. An analysis of the relationship between chronic obstructive pulmonary disease, smoking and depression in an integrated healthcare system. *Gen. Hosp. Psychiatry* **64**, 72–79 (2020).

251. Fluharty, M., Taylor, A. E., Grabski, M. & Munafò, M. R. The Association of Cigarette Smoking With Depression and Anxiety: A Systematic Review. *Nicotine Tob. Res.* **19**, 3–13 (2017).

252. Audrain-McGovern, J., Leventhal, A. M. & Strong, D. R. Chapter Eight - The Role of Depression in the Uptake and Maintenance of Cigarette Smoking. in *International Review of Neurobiology* (ed. De Biasi, M.) vol. 124 209–243 (Academic Press, 2015).

253. Ishii, T., Wakabayashi, R., Kurosaki, H., Gemma, A. & Kida, K. Association of serotonin transporter gene variation with smoking, chronic obstructive pulmonary disease, and its depressive symptoms. *J. Hum. Genet.* **56**, 41–46 (2011).

254. Heinzman, J. T. *et al.* GWAS and systems biology analysis of depressive symptoms among smokers from the COPDGene cohort. *J. Affect. Disord.* **243**, 16–22 (2019).

255. Arnau-Soler, A. *et al.* Genome-wide by environment interaction studies of depressive symptoms and psychosocial stress in UK Biobank and Generation Scotland. *Transl. Psychiatry* **9**, 1–13 (2019).

256. Weinmann, S. C. & Pisetsky, D. S. Mechanisms of immune-related adverse events during the treatment of cancer with immune checkpoint inhibitors. *Rheumatol. Oxf. Engl.* **58**, vii59–vii67 (2019).

257. Mangan, B. L. *et al.* Evolving insights into the mechanisms of toxicity associated with immune checkpoint inhibitor therapy. *Br. J. Clin. Pharmacol.* **86**, 1778–1789 (2020).

258. Abdel-Wahab, N., Shah, M., Lopez-Olivo, M. A. & Suarez-Almazor, M. E. Use of Immune Checkpoint Inhibitors in the Treatment of Patients With Cancer and Preexisting Autoimmune Disease. *Ann. Intern. Med.* **168**, 121–130 (2018).

259. Johnson, D. B. *et al.* Ipilimumab Therapy in Patients With Advanced Melanoma and Preexisting Autoimmune Disorders. *JAMA Oncol.* **2**, 234 (2016).

260. Khan, Z., Hammer, C., Guardino, E., Chandler, G. S. & Albert, M. L. Mechanisms of immune-related adverse events associated with immune checkpoint blockade: using germline genetics to develop a personalized approach. *Genome Med.* **11**, 1–3 (2019).

261. Centers for Disease Control and Prevention (CDC). Chronic obstructive pulmonary disease among adults--United States, 2011. *MMWR Morb. Mortal. Wkly. Rep.* **61**, 938–943 (2012).

262. Siu, A. L. *et al.* Screening for Chronic Obstructive Pulmonary Disease: US Preventive Services Task Force Recommendation Statement. *JAMA* **315**, 1372–1377 (2016).

263. Han, M. K. *et al.* Spirometry Utilization for COPD: How Do We Measure Up? *Chest* **132**, 403–409 (2007).

264. Bourbeau, J. *et al.* Canadian Cohort Obstructive Lung Disease (CanCOLD): Fulfilling the Need for Longitudinal Observational Studies in COPD. *COPD J. Chronic Obstr. Pulm. Dis.* **11**, 125–132 (2014).

265. Eagan, T. M. L. *et al.* Systemic inflammatory markers in COPD: results from the Bergen COPD Cohort Study. *Eur. Respir. J.* **35**, 540–548 (2010).

266. Watz, H., Waschki, B., Meyer, T. & Magnussen, H. Physical activity in patients with COPD. *Eur. Respir. J.* **33**, 262–272 (2009).

267. Schwartz, D. & Lellouch, J. Explanatory and Pragmatic Attitudes in Therapeutical Trials. *J. Clin. Epidemiol.* **5**, 499–505 (2009).

268. Ford, I. & Norrie, J. Pragmatic Trials. *N. Engl. J. Med.* **375**, 454–463 (2016).

269. Staa, T.-P. van *et al.* Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ* **344**, e55 (2012).

270. Richesson, R. L. *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J. Am. Med. Inform. Assoc.* **20**, e226–e231 (2013).

271. NHLBI. COPD National Action Plan.

272. Celli, B. R. *et al.* An Official American Thoracic Society/European Respiratory Society Statement: Research Questions in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **191**, e4–e27 (2015).

273. Hopkinson, N. S. *et al.* Designing and implementing a COPD discharge care bundle. *Thorax* **67**, 90–92 (2012).

274. Brown, K. E., Johnson, K. J., DeRonne, B. M., Parenti, C. M. & Rice, K. L. Order Set to Improve the Care of Patients Hospitalized for an Exacerbation of Chronic Obstructive Pulmonary Disease. *Ann. Am. Thorac. Soc.* **13**, 811–815 (2016).

275. Li, A., Chan, Y.-H., Liew, M. F., Pandey, R. & Phua, J. Improving Influenza Vaccination Coverage Among Patients With COPD: A Pilot Project. *Int. J. Chron. Obstruct. Pulmon. Dis.* **14**, 2527–2533 (2019).

276. Williamson, T. *et al.* Validating the 8 CPCSSN Case Definitions for Chronic Disease Surveillance in a Primary Care Database of Electronic Health Records. *Ann. Fam. Med.* **12**, 367–372 (2014).

277. Soriano, J. B., Maier, W. C., Visick, G. & Pride, N. B. Validation of general practitioner-diagnosed COPD in the UK General Practice Research Database. *Eur. J. Epidemiol.* **17**, 1075–1080 (2001).

278. Lacasse, Y., Daigle, J.-M., Martin, S. & Maltais, F. Validity of Chronic Obstructive Pulmonary Disease Diagnoses in a Large Administrative Database. *Can. Respir. J.* **19**, e5-9 (2012).

279. Smidth, M., Sokolowski, I., Kærsvang, L. & Vedsted, P. Developing an algorithm to identify people with Chronic Obstructive Pulmonary Disease (COPD) using administrative data. *BMC Med. Inform. Decis. Mak.* **12**, 38 (2012).

280. Mapel, D. W., Dutro, M. P., Marton, J. P., Woodruff, K. & Make, B. Identifying and characterizing COPD patients in US managed care. A retrospective, cross-sectional analysis of administrative claims data. *BMC Health Serv. Res.* **11**, 43 (2011).

281. Coleman, N. *et al.* From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC Fam. Pract.* **16**, 11 (2015).

282. Quint, J. K. *et al.* Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open* **4**, e005540 (2014).

283. Lee, T. M., Tu, K., Wing, L. L. & Gershon, A. S. Identifying individuals with physician-diagnosed chronic obstructive pulmonary disease in primary care electronic medical records: a retrospective chart abstraction study. *NPJ Prim. Care Respir. Med.* **27**, 34 (2017).

284. Cooke, C. R. *et al.* The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC Health Serv. Res.* **11**, 37 (2011).

285. Gershon, A. S. *et al.* Identifying Individuals with Physcian Diagnosed COPD in Health Administrative Databases. *COPD J. Chronic Obstr. Pulm. Dis.* **6**, 388–394 (2009).

286. Himes, B. E., Dai, Y., Kohane, I. S., Weiss, S. T. & Ramoni, M. F. Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records. *J. Am. Med. Inform. Assoc.* **16**, 371–379 (2009).

287. Moretz, C. *et al.* Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database. *J. Manag. Care Spec. Pharm.* **21**, 1149–1159 (2015).

288. Ho, T.-W. *et al.* Validity of ICD9-CM codes to diagnose chronic obstructive pulmonary disease from National Health Insurance claim data in Taiwan. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 3055–3063 (2018).

289. Crothers, K. *et al.* Accuracy of electronic health record data for the diagnosis of chronic obstructive pulmonary disease in persons living with HIV and uninfected persons. *Pharmacoepidemiol. Drug Saf.* **28**, 140–147 (2019).

290. Leidy, N. K. *et al.* Insight into Best Variables for COPD Case Identification: A Random Forests Analysis. *Chronic Obstr. Pulm. Dis. Miami Fla* **3**, 406–418 (2016).

291. Roden, D. *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).

292. GOLD 2017 Global Strategy for the Diagnosis, Management and Prevention of COPD. *Global Initiative for Chronic Obstructive Lung Disease - GOLD* http://goldcopd.org/.

293. Gamer, M., Lemon, J., Fellows, I. & Singh, P. *irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1*.

294. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

295. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

296. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

297. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

298. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

299. American Thoracic Society & European Respiratory Society. American Thoracic Society/European Respiratory Society statement: standards for the diagnosis and

management of individuals with alpha-1 antitrypsin deficiency. *Am. J. Respir. Crit. Care Med.* **168**, 818–900 (2003).

300. Alpha 1-antitrypsin deficiency: memorandum from a WHO meeting. *Bull. World Health Organ.* **75**, 397–415 (1997).
301. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
302. Payne, N. R. & Puumala, S. E. Racial disparities in ordering laboratory and radiology tests for pediatric patients in the emergency department. *Pediatr. Emerg. Care* **29**, 598–606 (2013).
303. Heisler, M., Smith, D. M., Hayward, R. A., Krein, S. L. & Kerr, E. A. Racial Disparities in Diabetes Care Processes, Outcomes, and Treatment Intensity. *Med. Care* **41**, 1221–1232 (2003).
304. López, L., Wilper, A. P., Cervantes, M. C., Betancourt, J. R. & Green, A. R. Racial and Sex Differences in Emergency Department Triage Assessment and Test Ordering for Chest Pain, 1997–2006. *Acad. Emerg. Med.* **17**, 801–808 (2010).
305. Wysocki, T., Diaz, M. C. G., Crutchfield, J. H., Franciosi, J. P. & Werk, L. N. Electronic health record as a research tool: Frequency of exposure to targeted clinical problems and health care providers' clinical proficiency. *J. Biomed. Inform.* **70**, 14–26 (2017).
306. Tse, J. & You, W. How accurate is the electronic health record? - a pilot study evaluating information accuracy in a primary care setting. *Stud. Health Technol. Inform.* **168**, 158–164 (2011).
307. Song, Y. *et al.* Regional Variations in Diagnostic Practices. *N. Engl. J. Med.* **363**, 45–53 (2010).

308. Chan, K. S., Fowles, J. B. & Weiner, J. P. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med. Care Res. Rev. MCRR* **67**, 503–527 (2010).

309. Hersh, W. R. *et al.* Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Med. Care* **51**, S30–S37 (2013).

310. Hogan, W. R. & Wagner, M. M. Accuracy of data in computer-based patient records. *J. Am. Med. Inform. Assoc. JAMIA* **4**, 342–355 (1997).

311. Weiskopf, N. G. & Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc. JAMIA* **20**, 144–151 (2013).

312. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc. JAMIA* **20**, 117–121 (2013).

313. Wei, W.-Q. & Denny, J. C. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med.* **7**, 41 (2015).

314. Mandrekar, J. N. Simple Statistical Measures for Diagnostic Accuracy Assessment. *J. Thorac. Oncol.* **5**, 763–764 (2010).

315. Zhou, X.-H., McClish, D. K. & Obuchowski, N. A. *Statistical Methods in Diagnostic Medicine*. (John Wiley & Sons, 2009).

316. Wei, W.-Q. *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc.* **23**, e20–e27 (2016).

317. O'Malley, K. J. *et al.* Measuring diagnoses: ICD code accuracy. *Health Serv. Res.* **40**, 1620–1639 (2005).

318. Pendergrass, S. A. & Crawford, D. C. Using Electronic Health Records To Generate Phenotypes For Research. *Curr. Protoc. Hum. Genet.* **100**, e80 (2019).

319. Crothers, K. *et al.* Increased COPD Among HIV-Positive Compared to HIV-Negative Veterans. *Chest* **130**, 1326–1333 (2006).

320. Diab, N. *et al.* Underdiagnosis and Overdiagnosis of Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **198**, 1130–1139 (2018).

321. Oude Rengerink, K. *et al.* Series: Pragmatic trials and real world evidence: Paper 3. Patient selection challenges and consequences. *J. Clin. Epidemiol.* **89**, 173–180 (2017).

322. Chubak, J., Pocobelli, G. & Weiss, N. S. Tradeoffs between accuracy measures for electronic health care data algorithms. *J. Clin. Epidemiol.* **65**, 343-349.e2 (2012).

323. Yu, W. C. *et al.* Spirometry is underused in the diagnosis and monitoring of patients with chronic obstructive pulmonary disease (COPD). *Int. J. Chron. Obstruct. Pulmon. Dis.* **8**, 389–395 (2013).

324. Joo, M. J., Sharp, L. K., Au, D. H., Lee, T. A. & Fitzgibbon, M. L. Use of Spirometry in the Diagnosis of COPD: A Qualitative Study in Primary Care. *COPD* **10**, 444–449 (2013).

325. Damarla, M., Celli, B. R., Mullerova, H. X. & Pinto-Plata, V. M. Discrepancy in the Use of Confirmatory Tests in Patients Hospitalized With the Diagnosis of Chronic Obstructive Pulmonary Disease or Congestive Heart Failure. *Respir. Care* **51**, 1120–1124 (2006).

326. Arne, M. *et al.* How often is diagnosis of COPD confirmed with spirometry? *Respir. Med.* **104**, 550–556 (2010).

327. Gershon, A. S., Hwee, J., Croxford, R., Aaron, S. D. & To, T. Patient and Physician Factors Associated With Pulmonary Function Testing for COPD: A Population Study. *Chest* **145**, 272–281 (2014).

328. Lamprecht, B. *et al.* Determinants of Underdiagnosis of COPD in National and International Surveys. *Chest* **148**, 971–985 (2015).

329. Frank, T. L., Hazell, M. L., Linehan, M. F. & Frank, P. I. The diagnostic accuracies of chronic obstructive pulmonary disease (COPD) in general practice: The results of the MAGIC (Manchester Airways Group Identifying COPD) study. *Prim. Care Respir. J. J. Gen. Pract. Airw. Group* **15**, 286–293 (2006).

330. Moore, P. L. Practice Management and Chronic Obstructive Pulmonary Disease in Primary Care. *Am. J. Med.* **120**, S23–S27 (2007).

331. Wu, H., Wise, R. A. & Medinger, A. E. Do Patients Hospitalized With COPD Have Airflow Obstruction? *Chest* **151**, 1263–1271 (2017).

332. Camp, P. G. *et al.* The sex factor: epidemiology and management of chronic obstructive pulmonary disease in British Columbia. *Can. Respir. J.* **15**, 417–422 (2008).

333. *Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2016*. (World Health Organization, 2018).

334. Kendler, K. S., Ohlsson, H., Lichtenstein, P., Sundquist, J. & Sundquist, K. The Genetic Epidemiology of Treated Major Depression in Sweden. *Am. J. Psychiatry* **175**, 1137–1144 (2018).

335. Baselmans, B. M. L., Yengo, L., van Rheenen, W. & Wray, N. R. Risk in relatives, heritability, SNP-based heritability and genetic correlations in psychiatric disorders: a review. *Biol. Psychiatry* (2020) doi:10.1016/j.biopsych.2020.05.034.

336. Fernandez-Pujals, A. M. *et al.* Epidemiology and Heritability of Major Depressive Disorder, Stratified by Age of Onset, Sex, and Illness Course in Generation Scotland: Scottish Family Health Study (GS:SFHS). *PLoS ONE* **10**, (2015).

337. Hall, R., Hall, I. P. & Sayers, I. Genetic risk factors for the development of pulmonary disease identified by genome-wide association. *Respirology* **24**, 204–214 (2019).

338. Busch, R., Cho, M. H. & Silverman, E. K. Progress in disease progression genetics: dissecting the genetic origins of lung function decline in COPD. *Thorax* **72**, 389–390 (2017).

339. Klimentidis, Y. C. *et al.* Heritability of pulmonary function estimated from pedigree and whole-genome markers. *Front. Genet.* **4**, 174 (2013).

340. Eid, R. S., Gobinath, A. R. & Galea, L. A. M. Sex differences in depression: Insights from clinical and preclinical studies. *Prog. Neurobiol.* **176**, 86–102 (2019).

341. Labonté, B. *et al.* Sex-specific transcriptional signatures in human depression. *Nat. Med.* **23**, 1102–1111 (2017).

342. Kang, H.-J. *et al.* Sex differences in the genetic architecture of depression. *Sci. Rep.* **10**, 9927 (2020).

343. Sørheim, I.-C. *et al.* Gender differences in COPD: are women more susceptible to smoking effects than men? *Thorax* **65**, 480–485 (2010).

344. Gan, W. Q., Man, S. P., Postma, D. S., Camp, P. & Sin, D. D. Female smokers beyond the perimenopausal period are at increased risk of chronic obstructive pulmonary disease: a systematic review and meta-analysis. *Respir. Res.* **7**, 52 (2006).

345. Raghavan, D. & Jain, R. Increasing awareness of sex differences in airway diseases. *Respirology* **21**, 449–459 (2016).

346. Dashti, H. S. *et al.* Genetic determinants of daytime napping and effects on cardiometabolic health. *Nat. Commun.* **12**, 900 (2021).

347. Mosley, J. D. *et al.* The polygenic architecture of left ventricular mass mirrors the clinical epidemiology. *Sci. Rep.* **10**, 7561 (2020).

348. Salem, J.-E. *et al.* Association of Thyroid Function Genetic Predictors With Atrial Fibrillation: A Phenome-Wide Association Study and Inverse-Variance Weighted Average Meta-analysis. *JAMA Cardiol.* **4**, 136–143 (2019).

349. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLOS Genet.* **2**, e190 (2006).

350. Purcell, S. & Chang, C. *PLINK [v1.9]*.

351. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

352. Howard, D. M. *et al.* Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat. Commun.* **9**, 1470 (2018).

353. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

354. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

355. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).

356. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).

357. Fritsche, L. G. *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* (2018) doi:10.1016/j.ajhg.2018.04.001.

358. Kember, R. L. *et al.* Polygenic risk of psychiatric disorders exhibits cross-trait associations in electronic health record data. *bioRxiv* 858027 (2019) doi:10.1101/858027.

359. Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.* **21**, 493–502 (2020).

360. Robinson, J. R., Wei, W.-Q., Roden, D. M. & Denny, J. C. Defining Phenotypes from Clinical Data to Drive Genomic Research. *Annu. Rev. Biomed. Data Sci.* **1**, 69–92 (2018).

361. Shen, X. *et al.* A phenome-wide association and Mendelian Randomisation study of polygenic risk for depression in UK Biobank. *Nat. Commun.* **11**, 2301 (2020).

362. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinforma. Oxf. Engl.* **30**, 2375–2376 (2014).

363. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).

364. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

365. Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinforma. Oxf. Engl.* **30**, i185-194 (2014).

366. Caramori, G. *et al.* Autoimmunity and COPD: Clinical Implications. *Chest* **153**, 1424–1431 (2018).

367. Barnes, P. J. Cellular and molecular mechanisms of asthma and COPD. *Clin. Sci. Lond. Engl. 1979* **131**, 1541–1558 (2017).

368. Wohleb, E. S., Franklin, T., Iwata, M. & Duman, R. S. Integrating neuroimmune systems in the neurobiology of depression. *Nat. Rev. Neurosci.* **17**, 497–511 (2016).

369. Otte, C. *et al.* Major depressive disorder. *Nat. Rev. Dis. Primer* **2**, 1–20 (2016).

370. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).

371. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).

372. Wootton, R. E. *et al.* Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychol. Med.* 1–9 (2019) doi:10.1017/S0033291719002678.

373. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet.* **52**, 437–447 (2020).

374. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).

375. Nagel, M. *et al.* Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).

376. Goodwin, R. D. *et al.* Depression, Anxiety, and COPD: The Unexamined Role of Nicotine Dependence. *Nicotine Tob. Res.* **14**, 176–183 (2012).

377. Lou, P. *et al.* Effects of Smoking, Depression, and Anxiety on Mortality in COPD Patients: A Prospective Study. *Respir. Care* **59**, 54–61 (2014).

378. Mathew, A. R., Yount, S. E., Kalhan, R. & Hitsman, B. Psychological Functioning in Patients with Chronic Obstructive Pulmonary Disease: A Preliminary Study of Relations with Smoking Status and Disease Impact. *Nicotine Tob. Res. Off. J. Soc. Res. Nicotine Tob.* (2018) doi:10.1093/ntr/nty102.

379. Callaghan, R. C. *et al.* Patterns of tobacco-related mortality among individuals diagnosed with schizophrenia, bipolar disorder, or depression. *J. Psychiatr. Res.* **48**, 102–110 (2014).

380. Tidey, J. W. & Miller, M. E. Smoking cessation and reduction in people with chronic mental illness. *BMJ* **351**, (2015).

381. Tam, J., Warner, K. E. & Meza, R. Smoking and the Reduced Life Expectancy of Individuals With Serious Mental Illness. *Am. J. Prev. Med.* **51**, 958–966 (2016).

382. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

383. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).

384. Beesley, L. J., Fritsche, L. G. & Mukherjee, B. An analytic framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. *Stat. Med.* **39**, 1965–1979 (2020).

385. Funk, M. J. & Landi, S. N. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. *Curr. Epidemiol. Rep.* **1**, 175–185 (2014).

386. Polubriaginof, F., Salmasian, H., Albert, D. A. & Vawdrey, D. K. Challenges with Collecting Smoking Status in Electronic Health Records. *AMIA. Annu. Symp. Proc.* **2017**, 1392–1400 (2018).

387. Szatkowski, L., Lewis, S., McNeill, A. & Coleman, T. Is smoking status routinely recorded when patients register with a new GP? *Fam. Pract.* **27**, 673–675 (2010).

388. Marston, L. *et al.* Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study. *BMJ Open* **4**, e004958 (2014).

389. Wu, C.-Y. *et al.* Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PloS One* **8**, e74262 (2013).

390. Robert, C. *et al.* Pembrolizumab versus ipilimumab in advanced melanoma (KEYNOTE-006): post-hoc 5-year results from an open-label, multicentre, randomised, controlled, phase 3 study. *Lancet Oncol.* **20**, 1239–1251 (2019).

391. Larkin, J. *et al.* Five-Year Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *N. Engl. J. Med.* **381**, 1535–1546 (2019).

392. Garon, E. B. *et al.* Five-Year Overall Survival for Patients With Advanced Non–Small-Cell Lung Cancer Treated With Pembrolizumab: Results From the Phase I KEYNOTE-001 Study. *J. Clin. Oncol.* **37**, 2518–2527 (2019).

393. Garon, E. B. *et al.* Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **372**, 2018–2028 (2015).

394. Doroshow, D. B. *et al.* PD-L1 as a biomarker of response to immune-checkpoint inhibitors. *Nat. Rev. Clin. Oncol.* 1–18 (2021) doi:10.1038/s41571-021-00473-5.

395. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**, 124–128 (2015).

396. Peled, J. U. *et al.* Intestinal Microbiota and Relapse After Hematopoietic-Cell Transplantation. *J. Clin. Oncol.* **35**, 1650–1659 (2017).

397. Baruch, E. N. *et al.* Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. *Science* **371**, 602–609 (2021).

398. Davar, D. *et al.* Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. *Science* **371**, 595–602 (2021).

399. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

400. Orrù, V. *et al.* Genetic variants regulating immune cell levels in health and disease. *Cell* **155**, 242–256 (2013).

401. Lim, Y. W. *et al.* Germline genetic polymorphisms influence tumor gene expression and immune cell infiltration. *Proc. Natl. Acad. Sci.* **115**, E11701–E11710 (2018).

402. Shahamatdar, S. *et al.* Germline Features Associated with Immune Infiltration in Solid Tumors. *Cell Rep.* **30**, 2900-2908.e4 (2020).

403. Sayaman, R. W. *et al.* Germline genetic contribution to the immune landscape of cancer. *Immunity* **54**, 367-386.e8 (2021).

404. Postow, M. A., Sidlow, R. & Hellmann, M. D. Immune-Related Adverse Events Associated with Immune Checkpoint Blockade. *N. Engl. J. Med.* **378**, 158–168 (2018).

405. Ricciuti, B. *et al.* Impact of immune-related adverse events on survival in patients with advanced non-small cell lung cancer treated with nivolumab: long-term outcomes from a multi-institutional analysis. *J. Cancer Res. Clin. Oncol.* **145**, 479–485 (2019).

406. Fujimoto, D. *et al.* Efficacy and safety of nivolumab in previously treated patients with non-small cell lung cancer: A multicenter retrospective cohort study. *Lung Cancer Amst. Neth.* **119**, 14–20 (2018).

407. Street, S. *et al.* The positive effect of immune checkpoint inhibitor-induced thyroiditis on overall survival accounting for immortal time bias: a retrospective cohort study of 6596 patients. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* (2021) doi:10.1016/j.annonc.2021.05.357.

408. Shankar, B. *et al.* Multisystem Immune-Related Adverse Events Associated With Immune Checkpoint Inhibitors for Treatment of Non-Small Cell Lung Cancer. *JAMA Oncol.* **6**, 1952–1956 (2020).

409. Horvat, T. Z. *et al.* Immune-Related Adverse Events, Need for Systemic Immunosuppression, and Effects on Survival and Time to Treatment Failure in Patients With Melanoma Treated With Ipilimumab at Memorial Sloan Kettering Cancer Center. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **33**, 3193–3198 (2015).

410. Weber, J. S. *et al.* Safety Profile of Nivolumab Monotherapy: A Pooled Analysis of Patients With Advanced Melanoma. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **35**, 785–792 (2017).

411. Verheijden, R. J. *et al.* Association of Anti-TNF with Decreased Survival in Steroid Refractory Ipilimumab and Anti-PD1-Treated Patients in the Dutch Melanoma Treatment Registry. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **26**, 2268–2274 (2020).

412. Suresh, K. & Naidoo, J. Lower Survival in Patients Who Develop Pneumonitis Following Immunotherapy for Lung Cancer. *Clin. Lung Cancer* **21**, e169–e170 (2020).

413. Osorio, J. C. *et al.* Antibody-mediated thyroid dysfunction during T-cell checkpoint blockade in patients with non-small-cell lung cancer. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **28**, 583–589 (2017).

414. Lee, H. *et al.* Characterization of Thyroid Disorders in Patients Receiving Immune Checkpoint Inhibition Therapy. *Cancer Immunol. Res.* **5**, 1133–1140 (2017).

415. Quandt, Z., Young, A., Perdigoto, A. L., Herold, K. C. & Anderson, M. S. Autoimmune Endocrinopathies: An Emerging Complication of Immune Checkpoint Inhibitors. *Annu. Rev. Med.* **72**, 313–330 (2021).

416. Robert, C. *et al.* Association of immune-related thyroid disorders with pembrolizumab (pembro, MK-3475) in patients (pts) with advanced melanoma treated in KEYNOTE-001. *J. Clin. Oncol.* **33**, 9050–9050 (2015).

417. Hansen, P. S., Brix, T. H., Sørensen, T. I. A., Kyvik, K. O. & Hegedüs, L. Major genetic influence on the regulation of the pituitary-thyroid axis: a study of healthy Danish twins. *J. Clin. Endocrinol. Metab.* **89**, 1181–1187 (2004).

418. Panicker, V. *et al.* Heritability of serum TSH, free T4 and free T3 concentrations: a study of a large UK twin cohort. *Clin. Endocrinol. (Oxf.)* **68**, 652–659 (2008).

419. Teumer, A. *et al.* Genome-wide analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. *Nat. Commun.* **9**, 4455 (2018).

420. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

421. Xu, H. *et al.* MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc. JAMIA* **17**, 19–24 (2010).

422. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

423. Wu, Y. *et al.* Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat. Commun.* **10**, 1891 (2019).

424. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

425. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).

426. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

427. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).

428. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

429. Hamnvik, O.-P. R., Larsen, P. R. & Marqusee, E. Thyroid dysfunction from antineoplastic agents. *J. Natl. Cancer Inst.* **103**, 1572–1587 (2011).

430. Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* **51**, 1349–1355 (2019).

431. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. *bioRxiv* 588020 (2019) doi:10.1101/588020.

432. Beattie, J. *et al.* Success and failure of additional immune modulators in steroid-refractory/resistant pneumonitis related to immune checkpoint blockade. *J. Immunother. Cancer* **9**, (2021).

433. Moslehi, J. J., Salem, J.-E., Sosman, J. A., Lebrun-Vignes, B. & Johnson, D. B. Increased reporting of fatal immune checkpoint inhibitor-associated myocarditis. *Lancet Lond. Engl.* **391**, 933 (2018).

434. Kotwal, A., Kottschade, L. & Ryder, M. PD-L1 Inhibitor-Induced Thyroiditis Is Associated with Better Overall Survival in Cancer Patients. *Thyroid Off. J. Am. Thyroid Assoc.* **30**, 177–184 (2020).

435. Hasan Ali, O. *et al.* Human leukocyte antigen variation is associated with adverse events of checkpoint inhibitors. *Eur. J. Cancer Oxf. Engl. 1990* **107**, 8–14 (2019).

436. Cappelli, L. C., Dorak, M. T., Bettinotti, M. P., Bingham, C. O. & Shah, A. A. Association of HLA-DRB1 shared epitope alleles and immune checkpoint inhibitor-induced inflammatory arthritis. *Rheumatol. Oxf. Engl.* **58**, 476–480 (2019).

437. Heaney, A. P. *et al.* HLA Markers DQ8 and DR53 Are Associated With Lymphocytic Hypophysitis and May Aid in Differential Diagnosis. *J. Clin. Endocrinol. Metab.* **100**, 4092–4097 (2015).

438. Quandt, Z. *et al.* Investigation of the Predictive Utility of a Type 1 Diabetes Genetic Risk Score in Immune Checkpoint Inhibitor Induced Diabetes Mellitus. (2020).

439. Abdel-Wahab, N. *et al.* Genetic determinants of immune-related adverse events in patients with melanoma receiving immune checkpoint inhibitors. *Cancer Immunol. Immunother. CII* (2021) doi:10.1007/s00262-020-02797-0.

440. Concannon, P., Rich, S. S. & Nepom, G. T. Genetics of type 1A diabetes. *N. Engl. J. Med.* **360**, 1646–1654 (2009).

441. Tomer, Y. Genetic susceptibility to autoimmune thyroid disease: past, present, and future. *Thyroid Off. J. Am. Thyroid Assoc.* **20**, 715–725 (2010).

442. Khan, Z. *et al.* Polygenic risk for skin autoimmunity impacts immune checkpoint blockade in bladder cancer. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12288–12294 (2020).

443. Classification of Diseases (ICD). https://www.who.int/standards/classifications/classification-of-diseases.

444. Baxi, S. *et al.* Immune-related adverse events for anti-PD-1 and anti-PD-L1 drugs: systematic review and meta-analysis. *BMJ* **360**, k793 (2018).

445. Myers, G. Immune-related adverse events of immune checkpoint inhibitors: a brief review. *Curr. Oncol. Tor. Ont* **25**, 342–347 (2018).

446. Knepper, T. C. & McLeod, H. L. When will clinical trials finally reflect diversity? *Nature* **557**, 157–159 (2018).

447. Fiscella, K. & Sanders, M. R. Racial and Ethnic Disparities in the Quality of Health Care. *Annu. Rev. Public Health* **37**, 375–394 (2016).

448. Wheeler, S. M. & Bryant, A. S. Racial and Ethnic Disparities in Health and Health Care. *Obstet. Gynecol. Clin. North Am.* **44**, 1–11 (2017).

449. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE* **12**, e0175508 (2017).

450. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).

451. Heid, I. M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.* **42**, 949–960 (2010).

452. Dastani, Z. *et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, e1002607 (2012).

453. Fritsche, L. G. *et al.* Seven new loci associated with age-related macular degeneration. *Nat. Genet.* **45**, 433–439, 439e1-2 (2013).

454. Elks, C. E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat. Genet.* **42**, 1077–1085 (2010).

455. Styrkarsdottir, U. *et al.* Multiple genetic loci for bone mineral density and fractures. *N. Engl. J. Med.* **358**, 2355–2365 (2008).

456. Barber, M. J. *et al.* Genome-wide association of lipid-lowering response to statins in combined study populations. *PloS One* **5**, e9763 (2010).

457. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).

458. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).

459. Hu, X. *et al.* Meta-analysis for genome-wide association study identifies multiple variants at the BIN1 locus associated with late-onset Alzheimer's disease. *PloS One* **6**, e16616 (2011).

460. Pankratz, N. *et al.* Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2. *Ann. Neurol.* **71**, 370–384 (2012).

461. Meyer, W. K. *et al.* Evaluating the evidence for transmission distortion in human pedigrees. *Genetics* **191**, 215–232 (2012).

462. Köttgen, A. *et al.* New loci associated with kidney function and chronic kidney disease. *Nat. Genet.* **42**, 376–384 (2010).

463. Mead, S. *et al.* Genome-wide association study in multiple human prion diseases suggests genetic risk factors additional to PRNP. *Hum. Mol. Genet.* **21**, 1897–1906 (2012).

464. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).

465. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

466. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).

467. Hill, W. D. *et al.* Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nat. Commun.* **10**, 5741 (2019).

468. Taal, H. R. *et al.* Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat. Genet.* **44**, 532–538 (2012).

469. Thanassoulis, G. *et al.* Genetic associations with valvular calcification and aortic stenosis. *N. Engl. J. Med.* **368**, 503–512 (2013).

470. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).