Leveraging Computational Methods And Electronic Health Records-Linked Biobank Data In Oral And Craniofacial Health Research

Venkateswaran, Vidhya

2023

UNIVERSITY OF CALIFORNIA

Los Angeles

# Leveraging Computational Methods And Electronic Health Records-Linked Biobank Data In Oral And Craniofacial Health Research

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Oral Biology

by

Vidhya Venkateswaran

2023

# ABSTRACT OF THE DISSERTATION

## Leveraging Computational Methods And Electronic Health Records-Linked Biobank Data In Oral And Craniofacial Health Research

by

Vidhya Venkateswaran

Doctor of Philosophy in Oral Biology

University of California, Los Angeles, 2023

Professor Ichiro Nishimura, Co-Chair

Professor Bogdan Pasaniuc, Co-Chair

Bioinformatics and computational methods play an important role in advancing medical research with their ability to leverage large datasets, including data from electronic health records (EHR) linked biobanks. Precision medicine can benefit from leveraging a more comprehensive picture of a patient’s genotypes and phenotypic presentation for targeted interventions and treatment planning. In this work, I discuss the applications of bioinformatics methods in the UCLA ATLAS biobank, in evaluating craniofacial traits and their risk factors: specifically, head and neck cancer and tobacco use disorder. First, I describe phenome-wide and lab-wide association analysis pipelines that leverage the breadth of the available information in the biobank, and the results of preliminary investigations of the phenome-wide and laboratory-wide associations of a genetic predisposition to tobacco use disorder. Next, I present the results of an evaluation of the predictive performance of a tobacco use polygenic score across different genetic ancestry groups and further discuss the differences in disease presentations in tobacco use-predisposed individuals with and without a history of the associated tobacco use behavior. Next, I employ these pipelines and statistical methods in the examination of the interplay of serum bilirubin, tobacco use, head and neck, and lung cancer. I present the results of this project, examining the effect of environmental and genetic factors on serum bilirubin and associations with head and neck cancer and lung cancer. Lastly, I propose a research project to examine the germline risk factors for oropharyngeal cancer and discuss the future directions of this work.

The dissertation of Vidhya Venkateswaran is approved.

Sanjay M. Mallya

Yvonne L. Kapila

Maie A. R. St. John

Ichiro Nishimura, Committee Co-Chair

Bogdan Pasaniuc, Committee Co-Chair

University of California, Los Angeles

2023

# DEDICATION

I dedicate this work to

my family and friends

TABLE OF CONTENTS

## List of Figures

## List of Tables

## Acknowledgements

All the chapters included in this dissertation were written by V Venkateswaran. The projects were conceived by V Venkateswaran with guidance from B Pasaniuc who served as the PI in all projects. A Bhattacharya, K Boulier, R Johnson, Y Ding and E Petter were involved in data analysis and interpretation. V Venkateswaran is funded by NIH/NIDCR 5K12DE027830.

Chapter 3 is preprinted as “Polygenic scores for tobacco use provide insights into systemic health risks in a diverse EHR-linked biobank in Los Angeles”. Vidhya Venkateswaran, Kristin Boulier, Yi Ding, Ruth Johnson, Arjun Bhattacharya, Bogdan Pasaniuc. medRxiv 2023.02.05.23285493;

Chapters 2 and 4 are in preparation for publication.

We gratefully acknowledge the Institute for Precision Health, participating patients from the UCLA ATLAS Precision Health Biobank, UCLA David Geffen School of Medicine, UCLA Clinical and Translational Science Institute, and UCLA Health.

# VITA

## Education

| | |
|---|---|
| 2019 - Present | UCLA PhD in Oral Biology |
| 2019 - 2021 | UCLA Oral Radiology Residency |
| 2016 - 2018 | Harvard University, School Of Public Health, MPH Epidemiology |
| 2005 - 2009 | TN MGR University, Tamil Nadu, India, Bachelor of Dental Surgery |

## Publications

1. **Vidhya Venkateswaran**, Kristin Boulier, Yi Ding, Ruth Johnson, Arjun Bhattacharya, Bogdan Pasaniuc. Polygenic scores for tobacco use provide insights into systemic health risks in a diverse EHR-linked biobank in Los Angeles (Preprint under review at Translational Psychiatry)
2. **Vidhya Venkateswaran**, Tara Aghaloo, Reuben Kim, Sotirios Tetradis, Ritu Tiwari, Sanjay M. Mallya. A clinical trial to evaluate the effectiveness and safety of Dual Energy Cone Beam Computed Tomography (DE-CBCT) imaging for assessment of jaw bone density (Under review at OOOO)
3. Johnson, R., Ding, Y., **Venkateswaran, V**. et al. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. Genome Med 14, 104 (2022).
4. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet (Mentor: Dr. Emmanuela Gakidou, UW- IHME)

5. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study, The Lancet (Mentor: Dr. Emmanuela Gakidou, UW- IHME)
6. Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016 GBD 2016 Alcohol Collaborators - The Lancet. Published: August 23, 2018 (Mentor: Dr. Emmanuela Gakidou, UW- IHME)
7. “Radiographic and imaging assessment of facial asymmetry” – book chapter co-written with Dr. David Hatcher for a textbook entitled ‘Dental Asymmetry’ due to be published in 2023 by Wiley
8. Parth Patel, **Vidhya Venkateswaran**, Bogdan Pasaniuc, and Kymora Scotland . Genetic analysis of kidney stone disease in a multi-ethnic cohort: insights from genome-wide and phenome-wide association studies. Poster presentation at AUA 2022
9. **Vidhya Venkateswaran**, S.M. Mallya. The Biology Of Periosteal Reactions. doi:https://doi.org/10.1016/j.oooo.2021.04.033 (AAOMR Award For Best Scientific Poster, Presented At The 2020 AAOMR Annual Session)
10. **Vidhya Venkateswaran**, S.M. Mallya. Molecules, Morphogenesis and Malformations: Radiological Manifestations of Deregulated FGF signaling. IADMFR conference 2021

# Chapter 1

## Introduction

Bioinformatics is a subdiscipline of data science that uses information technology to collect, analyze, and disseminate biological data and information[1]. A large amount of data is generated in healthcare settings; including demographics, encounter information, laboratory results, prescription data, diagnostic, and procedure codes. A newer source of data that is now adopted in many healthcare settings is the electronic health records linked biobank[2]. These biobanks include almost all patients in a healthcare system who have consented to participate in research, with each patient's de-identified health records and linked genotypes. Consequently, these biobanks have large sample sizes, capturing a diverse range of ethnicities and races, across various socioeconomic groups[3].

Computational methods have yielded promising results in the study of many dental diseases and risk factors. A large genome-wide association meta-analysis of dental caries identified 47 novel genetic risk loci[4]. Mendelian randomization analyses have causally linked periodontal disease with diseases such as stroke[5] and obesity[6]. Genomic studies of oral cancer have identified the effect of germline mutations and genetic ancestry on somatic mutations and tumor characteristics[7-9]. Lastly, studies have identified genetic polymorphisms that interact with smoking pharmacotherapeutic outcomes including the *CHRNA5-A3-B4* and *CYP2A6* loci[10]. These examples demonstrate the potential of genomics and bioinformatics in expanding our understanding of the genetic basis of craniofacial diseases and risk factors, with clinically relevant results. There are some challenges to EHR-linked biobanks. While the data is rich, it is largely unstructured and unvalidated, requiring careful quality control and validation of any phenotypes to be examined. Building initial computational pipelines, and curating and preprocessing data are time-consuming and computationally intensive. However, after this initial investment, these pipelines and curated data are available for use by other researchers for many research questions.

My thesis focuses on building these pipelines and then utilizing them to study tobacco use and head and neck cancer (HNC) in the UCLA ATLAS biobank[10-14] - a diverse biobank embedded in the UCLA healthcare system with linked genotype information. Using these pipelines, I studied the genetic effects of a predisposition to tobacco use disorder and potential interplay of tobacco use with serum bilirubin on the risk of HNC and lung cancer in the UCLA biobank.

I found that a polygenic score (PGS)[15] for tobacco use disorder demonstrates inconsistent predictive performance in non-European ancestry populations in the UCLA biobank. The PGS was associated with a number of cardiometabolic, psychiatric and respiratory diseases across the phenome, capturing the effects of tobacco use. Interestingly, the PGS demonstrated associations with obesity and alcohol use disorder when tobacco use behavior was not present.

With further validation, these findings could have a significant impact on tobacco use management, suggesting that if an individual is genetically predisposed to tobacco use, early and comprehensive interventions to address underlying addictive tendencies might be warranted. Inconsistent predictive performance of the PGS across ancestry groups necessitates further research to improve this aspect before clinical translation to allow for equitable delivery of care.

Building on the results of the systemic effects of tobacco use, in an examination of the interplay between serum bilirubin, tobacco use, and head and neck, and lung cancer, I found that serum bilirubin had an inverse relationship with HNC and lung cancer risk. Tobacco use interacts with serum bilirubin on lung cancer risk and a polygenic score for serum bilirubin is associated with lung cancer. i.e. cigarette smokers with low serum bilirubin had a higher risk of lung cancer when compared to cigarette-smokers with high serum bilirubin. These findings indicate a potential role for serum bilirubin in the risk stratification of patients at risk of HNC and lung cancer.

Lastly in this thesis, I propose future studies to examine the germline genetic risk factors associated with oropharyngeal cancers using methods including genome-wide association studies, phenome-wide association studies and polygenic scores. These projects are organized into the following thesis chapters:

1. Introduction
2. Leveraging the Breadth of EHR-linked Biobank Data for Phenome and Lab-Wide Association Studies of Tobacco Use Genetic Variants
3. Polygenic Scores for Tobacco Use Provide Insights into Systemic Health Risks
4. EHR-Data And Polygenic Scores Reveal The Interplay Of Serum Bilirubin, Smoking, And Cancer
5. Conclusions and Future Directions

## References

1. NHGRI Talking Glossary of Genomic and Genetic Terms
2. Kinkorová J, Topolčan O. Biobanks in the era of big data: objectives, challenges, perspectives, and innovations for predictive, preventive, and personalised medicine. EPMA J. 2020;11(3):333-341. Published 2020 Jun 18. doi:10.1007/s13167-020-00213-2
3. Beesley LJ, Salvatore M, Fritsche LG, et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Stat Med. 2020;39(6):773-800. doi:10.1002/sim.8445
4. Shungin D, Haworth S, Divaris K, et al. Genome-wide analysis of dental caries and periodontitis combining clinical and self-reported data. *Nat Commun*. 2019;10(1):2773. Published 2019 Jun 24. doi:10.1038/s41467-019-10630-1
5. Ma C, Wu M, Gao J, et al. Periodontitis and stroke: A Mendelian randomization study. *Brain Behav*. 2023;13(2):e2888. doi:10.1002/brb3.2888
6. Dong J, Gong Y, Chu T, et al. Mendelian randomization highlights the causal association of obesity with periodontal diseases. J Clin Periodontol. 2022;49(7):662-671. doi:10.1111/jcpe.13640
7. Ferreiro-Iglesias A, McKay JD, Brenner N, et al. Germline determinants of humoral immune response to HPV-16 protect against oropharyngeal cancer. Nat Commun. 2021;12(1):5945. Published 2021 Oct 12. doi:10.1038/s41467-021-26151-9

8. Guo J, Liu X, Zeng Y, et al. Comprehensive Analysis of the Effects of Genetic Ancestry and Genetic Characteristics on the Clinical Evolution of Oral Squamous Cell Carcinoma. Front Cell Dev Biol. 2021;9:678464. Published 2021 Dec 7. doi:10.3389/fcell.2021.678464
9. Lesseur C, Diergaarde B, Olshan AF, et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. Nat Genet. 2016;48(12):1544-1550. doi:10.1038/ng.3685
10. Salloum NC, Buchalter ELF, Chanani S, et al. From genes to treatments: a systematic review of the pharmacogenetics in smoking cessation. Pharmacogenomics. 2018;19(10):861-871. doi:10.2217/pgs-2018-0023
11. Chang TS, Ding Y, Freund MK, et al. Pre-existing conditions in Hispanics/Latinxs that are COVID-19 risk factors. iScience. 2021;24(3):102188. doi:10.1016/j.isci.2021.102188
12. Lajonchere C, Naeim A, Dry S, et al. An Integrated, Scalable, Electronic Video Consent Process to Power Precision Health Research: Large, Population-Based, Cohort Implementation and Scalability Study. J Med Internet Res. 2021;23(12):e31121. Published 2021 Dec 8. doi:10.2196/31121
13. Johnson R, Ding Y, Venkateswaran V, et al. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative [published correction appears in Genome Med. 2022 Nov 16;14(1):128]. Genome Med. 2022;14(1):104. Published 2022 Sep 9. doi:10.1186/s13073-022-01106-x

14. Johnson R, Ding Y, Bhattacharya A, et al. The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. Cell Genom. 2023;3(1):100243. Published 2023 Jan 11. doi:10.1016/j.xgen.2022.100243

15. Sugrue LP, Desikan RS. What Are Polygenic Scores and Why Are They Important?. JAMA. 2019;321(18):1820-1821. doi:10.1001/jama.2019.3893

## Chapter 2

# Leveraging the Breadth of EHR-linked Biobank Data for Phenome and Lab-Wide Association Studies of Tobacco Use Genetic Variants

## 2.1 Abstract

EHR-linked biobanks offer a wealth of health information that can be utilized in research. We present the results of three pipelines, phenome-wide association analyses (PheWAS) using single nucleotide polymorphism (SNP) and a polygenic score (PGS) and a laboratory-wise association analysis (LabWAS) using a polygenic score for tobacco use disorder (TUD). We find that the SNP-PheWAS for rs6024489 is associated with 'tobacco use disorder' (*P-Value = 0.000011*), likely capturing the effect of this nicotine dependence risk loci. In the TUD-PGS-PheWAS, we observe top associations with 'Tobacco use disorder', 'Obesity', 'Diabetes Mellitus', 'Substance addiction and disorders', and 'Emphysema' (*P Values = 5.21E-36, 7.63E-13, 7.20E-12, 9.19E-11, 1.42E-10 respectively*). These results capture potential pleiotropic associations in addition to tobacco use driven comorbidities. Lastly, in the TUD-PGS-LabWAS, we observe associations with high glucose and HBA1c lab results (*P-Values: 4.30e-12 and 4.62e-08*) that validate the diseases noted above (diabetes mellitus). We also observe associations with HDL (*P Value: 1.30e-07*), a known effect of tobacco smoking. Lastly, a triad of decreased Vitamin D levels and calcium levels (*P Values: 1.71e-05 and 1.39e-05*) with increased parathyroid hormone (*P Value: 3.36e-06*) are suggestive of hyperparathyroidism, a new finding not captured in the PheWAS analysis. These findings demonstrate the potential of these pipelines in validation and discovery

of the system-wide effects of genetic variants associated with a trait of interest, in this case tobacco use disorder.

## 2.2 Introduction

Personalized medicine or precision medicine uses patient characteristics, including the genetic information of an individual, to design interventions personalized to each individual's genotype and disease presentation. Electronic health record (EHR)-linked biobanks provide a vast repository of information that can be leveraged in precision medicine research. EHR-linked biobanks can be used to identify disease biomarkers and diseases with shared genetic risk[1,2], thus identifying the potential symptoms, biomarkers, or laboratory measurements that may not directly relate to the classical presentation of a disease. Using methodological approaches developed for the complex data from EHRs, we can leverage genetic and health information to identify a network of comorbidities, and laboratory results to predict disease risk, treatment response, and adverse effects.

A novel way of harnessing the information available in EHR-linked biobanks is by testing for associations across the full spectrum of available data in order to generate hypotheses about the disease of interest. Two such methods that can evaluate associations across a wide spectrum of data include phenome-wide and lab-wide association studies. The phenome-wide association study (PheWAS), introduced by Denny et al in 2010[3], is an analysis designed to evaluate the associations between genetic variants and phecodes, which are meaningfully grouped ICD codes[3]. These phecodes span the full phenotypic spectrum, helping us identify patterns of disease presentation across different disease categories. PheWAS studies have been conducted in several biobanks including the UK Biobank and BioVu, providing us with novel disease insights across metabolic, psychiatric and cardiovascular phenotypes to name a few[4,5].

In a similar vein, laboratory test result data can be harnessed for lab-wide association studies (LabWAS), testing the effects of genetic variants across all laboratory tests available in the EHR to obtain a full picture of the potential impact of the genetic variants on a broad spectrum of lab results, including tests that might not directly be prescribed for the patient's existing medical history[6,7].

These association tests can utilize a polygenic score; a score assigned to each individual, summarizing the estimated effect of multiple genetic variants on a trait of interest, thus providing a more comprehensive estimate of genetic risk when compared to a single genetic variant[8]. In this chapter, we discuss computational pipelines for phenome-wide and lab-wide association testing within the UCLA ATLAS biobank for single nucleotide polymorphisms and polygenic scores. We highlight the differences between using individual variants and polygenic scores in these pipelines using tobacco-use genetic variants as a preliminary example and discuss the findings from each of these pipelines.

## 2.3 Methods

### 2.3.1 Study population

The PheWAS and LabWAS pipelines were built and evaluated in the UCLA ATLAS Biobank. The UCLA ATLAS biobank is an electronic health record-linked biobank embedded within the UCLA Health system, a comprehensive healthcare system serving the population in and around the greater Los Angeles area[9,10,11]. The UCLA Institute for Precision Health is home to the UCLA ATLAS biobank with approximately 60k participants genotyped, of which 25,463 participants were included in this study. ATLAS biobank includes a collection of genotyped biospecimens that are integrated with the UCLA Data Discovery Repository (DDR), which contains de-identified patient EHR that include clinical, procedural, laboratory, prescription, and demographic information. The participants included in this study were 18 years of age and above and provided informed consent to using their genotypes and EHR for research purposes.

We assigned the participants to genetically inferred ancestry (GIA) groups using their genotype information. The study population included in this study were individuals inferred to be in the European American continental ancestry group. Detailed descriptions of the workflow for inferring ancestry are discussed in previous publications[9,10,11]. Briefly, we computed the top 10 principal components of ATLAS participants using FlashPCA2 software[12]. We then grouped our study population into genetically inferred ancestry groups (GIAs) by using k-nearest neighbor (KNN) stratification of the principal components, using the continental ancestry populations from the 1000 Genomes Project as a reference[13].

### 2.3.2 ICD codes and Phecodes

The International Classification of Diseases (ICD) codes are a set of codes used by providers to record diseases, symptoms, and other elements of a patient's diagnosis in a patient’s health record. These ICD codes can be used to collect information about the patient's medical status and subsequently to study diseases, outcomes, and patterns in the population of interest. The UCLA ATLAS biobank includes versions 9 and 10 of the ICD codes. These ICD codes were mapped to ‘Phecodes’ to allow for meaningful clinical groupings. Phecodes are curated and grouped ICD codes that condense similar ICD codes into a single phecode[14], using the Phecode V1.2 mapping. This grouping reduces the dimensionality of ICD codes for research purposes, decreases multiple testing burden, and simplifies interpretation of results.  Phecode V1.2 contains 1864 phecodes in 18 categories that map to approximately 70,000 ICD codes. ICD 9 and 10 codes in each patient's record were mapped to the corresponding phecode, assigning them case or control status for each phecode. The presence of a phecode (i.e. ICD codes mapped to that particular phecode) in the individual’s EHR classifies them as a case, and the absence of all ICD codes mapped to a phecode classifies the individual as a control.

### 2.3.3 Lab tests and values

The de-identified data repository of the UCLA ATLAS biobank includes 1977 laboratory-base names that could be extracted from the electronic health records. 580 of these lab tests contained numeric lab values that could be used for the proposed analysis. After excluding lab tests with >80% missing data, 79 lab tests were included in the final analysis with minimal missing values and numeric results that could be used in statistical analysis. Since most individuals have multiple results of the same lab test, taken over their encounter span at UCLA health system, we extracted the maximum and minimum values in each patient’s record and computed their mean to get an average lab result. Final processing of the lab results included the following steps - We first replaced ‘999999’s with the code for missing data: ‘NA’, ensuring the lab tests still had <80% missing data. Next, we excluded outliers that were >0.9 percentile and <0.01 percentile for each lab test.

### 2.3.4 SNP and Polygenic Scores (Independent variables)

For the example SNP-PheWAS pipeline, we used a single nucleotide polymorphism (SNP) - rs6024489 (Chromosome 20, Base pair: 6465338) as the primary predictor. This SNP is located within 1 Megabase of genetic variants in the *CHRNA4* locus, a region known for its effect on nicotine metabolism and on the addictive response to nicotine[15].

For the PGS-PheWAS and PGS-LabWAS pipelines, we used a publicly available polygenic score (PGS) for tobacco use disorder from the PGS Catalog (PGS002037) as the primary predictor[16]; referred to as the TUD-PGS. This TUD-PGS was trained on 391,124 European individuals from the UK biobank. This trait, ‘tobacco use disorder’ was identified using phecode 318.0 which is also available for analysis within ATLAS. Additionally, 94.4% of the SNPs in the PGS demonstrate overlap with SNPs that were included in ATLAS. We computed the PGS for each ATLAS participant by multiplying the individual risk allele dosages by their corresponding weights that are provided by the PGS catalog. The PGS was mean-centered and standardized by the standard deviation within the EUR group to generate a standardized PGS Z-score.

### 2.3.5 Statistical analysis and association testing

All analysis was conducted in either Python 2.6.8[17] or R 4.2.1[18]. For the SNP-PheWAS and PGS-PheWAS, we used logistic regression models to evaluate the associations between the dependent variables which were the phecode case-control status and independent variables of either the SNP of interest or TUD-PGS. We adjusted for participant age, sex, and the first 5 principal components in all models. Odds ratios and confidence intervals were calculated, with P-values from Wald-type test statistics, using the following logistic regression model:

***1864 Phecodes ~ PGS/SNP + Age + Sex + PCs1-5***

For the PGS-LabWAS, we used linear regression models with the numeric laboratory test results as dependent variables and the independent variable as the PGS:

***79 Lab test results ~ PGS + Age + Sex + PCs1-5***

Similar to the logistic regression models, we adjusted for participant age, sex, and the first 5 principal components for the linear regression. Effect sizes and 95% confidence intervals were computed, with P-values from Wald-type test statistics.

### 2.3.6 Ethical Approval

Patient Recruitment and Sample Collection for Precision Health Activities at UCLA is an approved study by the UCLA Institutional Review Board (UCLA IRB) IRB#17-001013.

### 2.3.7 Data Sharing

All shareable data produced in the present work are contained in the manuscript.

## 2.4 Results

### 2.4.1 Study population characteristics

This study included 25,463 individuals of European American genetically inferred ancestry groups. All participants were aged 18 and above with an average age of the participants was 59.3 years and females constituted 52.6% of the study population.

Approximately 71,000 codes in ICD-10 and 4,000 codes in ICD-9 were available in patient records and these mapped to 1864 phecodes. Each participant had an average of 83 phecodes in their EHR. Phecode 272.0 - "Disorders of lipid metabolism", was the most represented with 14,237 patients containing this phecode at least once in their record.  Next, out of 580 numeric lab tests, 79 lab tests were available in ATLAS for analysis. Each patient had a mean of 48 lab test results in their record. The most commonly ordered lab test was platelet count with over 88% of included participants with a lab result for it in their EHR.

### 2.4.2 SNP-PheWAS for SNP near the *CHRNA4* region shows association with tobacco use disorder

First, we evaluated the phenome-wide effects of a single SNP, rs6024489 located in chromosome 20 using a SNP-PheWAS pipeline **(*Fig 2.1*)**. We examined the effects of this SNP on 1854 phecodes available within ATLAS. In a logistic regression-based SNP-PheWAS of individuals assigned to the European American genetically inferred ancestry group, we noted a significant association with the phecode 318.0 for 'tobacco use disorder' (*P-Value = 0.000011*) after adjusting for age, sex, and the first five principal components and correcting for multiple testing (0.05/1864). **(*Fig 2.2*)**

### 2.4.3 PGS-PheWAS of TUD-PGS shows associations with circulatory, endocrine, psychiatric, and respiratory phecodes

Next, we evaluated the phenome-wide associations of a TUD-PGS in European American GIA using a PGS-PheWAS pipeline **(*Fig 2.1*)**. While in the previous section, we evaluated the effect of a single SNP across the entire phenome, here we evaluate the effects of a polygenic score for tobacco use disorder, TUD-PGS (PGS002037) which captures the effect of multiple variants that genetically predispose an individual to tobacco use disorder. In a PGS-PheWAS of this TUD-PGS across 1864 phecodes in the European American GIA group, we observed 56 significant associations at Bonferroni-adjusted P < 0.05 after adjusting for age, sex, and the first 5 principal components. The top phecodes associated with the TUD-PGS were 'Tobacco use disorder', 'Obesity', 'Diabetes Mellitus', 'Substance addiction and disorders', and 'Emphysema' (*P Values = 5.21E-36, 7.63E-13, 7.20E-12, 9.19E-11, 1.42E-10 respectively*) **(*Fig 2.3*)**

### 2.4.4 PGS-LabWAS of TUD-PGS shows associations with lab tests that are not captured by phecodes

Lastly, we tested the association of the TUD-PGS across available numeric lab tests in ATLAS using linear regression models. In a PGS-LabWAS of TUD-PGS, we observed 17 significant associations with laboratory test results. Some lab tests followed a similar association pattern as noted in the PGS-PheWAS. For e.g., high glucose and HBA1c lab results (*P-Values: 4.30e-12 and 4.62e-08*). Other lab tests capture the effects of smoking on clinically observed and reported biomarkers. For e.g., low HDL (*P Value: 1.30e-07*) is a known effect of smoking behaviors and could also be secondary to insulin resistance. Lastly, some results generate interesting new avenues of research. For example, the triad of decreased Vitamin D levels and calcium levels (*P Values: 1.71e-05 and 1.39e-05*) with increased parathyroid hormone (PTHINT) (*P Value: 3.36e-06*) are suggestive of hyperparathyroidism.

## **2.5 Discussion**

In this paper, we leveraged the breadth of the rich data available in the UCLA ATLAS EHR-linked biobank by creating and utilizing phenome-wide and lab-wide pipelines. These pipelines enabled us to conduct association tests using both SNPs and PGS that are linked to tobacco use disorder or nicotine addiction. In our results, in a SNP-PheWAS of rs6024489, a variant <1 Megabase from *CHRNA4*, we observed a significant association with tobacco use disorder, capturing the reported effect of this loci on tobacco use. Next, in a TUD-PGS PheWAS, we observed several significant associations of the PGS with phecodes in various disease categories including cardiovascular, endocrine/metabolic, respiratory, and neuropsychiatric disorders. Lastly, we performed a TUD-PGS - LabWAS analysis where we found unique associations that were not captured by the previous PheWAS analysis.

The results of the SNP-PheWAS and PGS-PheWAS highlight some innate differences between the research questions answered by these two different methods. The former examines the effect of a single SNP across the entire phenome and these effects might be too small to capture and might not survive a very strict multiple-testing correction. The PGS-PheWAS tests the associations of an overall genetic predisposition to tobacco use disorder. The results of this analysis capture the comorbidities most noted in individuals with this predisposition, often likely secondary to the behavior they are predisposed to. Many of these TUD PGS-PheWAS correlations, including type 2 diabetes, cardiovascular and neuropsychiatric diseases, have been reported clinically and/or are reported in other genetic studies, validating our findings and the utility of this pipeline in identifying disease associations[19-23]. The results of the PGS-PheWAS can be interpreted in two different ways - 1) these diseases that are associated with a TUD-PGS could potentially share genetic architecture or 2) these diseases and tobacco use could be driven by common environmental factors including tobacco smoking behavior. Follow up studies are needed to disentangle the environmental vs the genetic drivers of these associations. In any case, identifying patterns of disease presentation is invaluable in designing precision health interventions and public policy around the treatment and prevention of tobacco use.

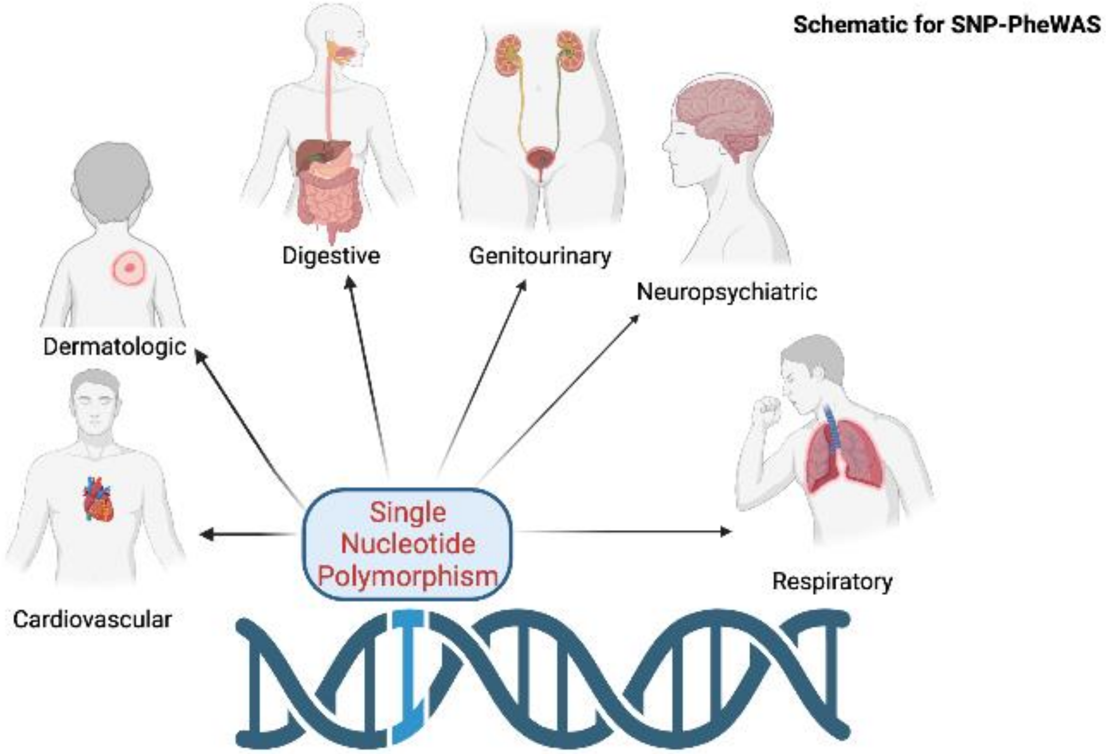
Lastly, we aimed to capture the effect of a genetic predisposition to tobacco use on common laboratory test results with the TUD-PGS-LabWAS pipeline. The results of this analysis validated several results of the PGS-PheWAS associations. For example, we found associations with increased glucose and HBA1C levels, capturing the Type 2 diabetes associations noted in the PGS-PheWAS. Next, we observed new associations with a triad of low vitamin D and low calcium levels with high parathyroid hormone levels, suggestive of an altered bone metabolism state, which could lead to osteoporosis and fractures[24]. This is a finding that was captured uniquely by the PGS-LabWAS analysis that we did not observe in the PGS-PheWAS analysis. Studies have linked tobacco use to altered Vitamin-D, calcium and parathyroid hormone levels[25], often a precursor to osteoporosis and the risk of increased fracture. These findings underscore the importance of assessing laboratory results in conjunction with phecodes, highlighting their combined value.

Our study has several strengths, we leverage the vast amount of information available in electronic health records to create flexible and reproducible pipelines for hypothesis-free association testing. In our preliminary analysis of tobacco use variants, we observe results that demonstrate the utility of these pipelines in validation and discovery. The results of our analysis are clinically relevant and could provide insights into the overall health and disease presentation of individuals with a genetic predisposition to tobacco use disorder. While we did not utilize longitudinal information in our analysis, the discovery and validation of potential biomarkers necessitate careful and well-designed follow-up studies that include longitudinal information to study such biomarkers in depth. The results of our pipelines and pilot analysis generate hypotheses between the genetic propensity to tobacco use disorder and several disease and lab associations that we have followed up with two individual studies, see chapters 3 and 4.

We conclude with some limitations of our study. We conducted our analyses solely on individuals of inferred European American Ancestry. This was intentional because studies have shown that polygenic scores perform best in the ancestry that the original PGS is trained in and do not generalize well to other ancestry groups[26]. However, this choice means that we cannot generalize the results of the PheWAS and LabWAS analyses to other ancestries as genetic effects may vary between ancestries[26]. Further studies are required to examine the predictive performance of TUD-PGS across ancestries and the cross-ancestry effects across the phenome. Next, we used ICD codes derived phecodes for our pipelines - ICD codes are billing codes and are not meant to record accurate or detailed diagnoses. Additionally, several phenotypes that are relevant to our trait of interest are not recorded by ICD codes, for e.g. quantitative measurements of tobacco and alcohol use. These limitations must be kept in mind while interpreting the results, noting that the ICD codes might not provide a fully complete picture of an individual's health. Lastly, several environmental factors have strong effects on tobacco use and the observed comorbidities including education level, income, and other socioeconomic factors. Unfortunately, as UCLA ATLAS is a de-identified database, these variables are unavailable for analysis and as such, the environmental effect on TUD and comorbidities cannot be quantified with this data alone. Further studies are required in other biobanks or in cohorts with these variables to further tease apart the effects of environment and genetics and study their interactive effects on tobacco use and associated comorbidities.

## 2.6 Figures

**Figure 2.1: Schematic of SNP-PheWAS and PGS-PheWAS/PGS-LabWAS: The disease status of the represented organ systems is captured through phecodes for the PheWAS and through lab test results for LabWAS**



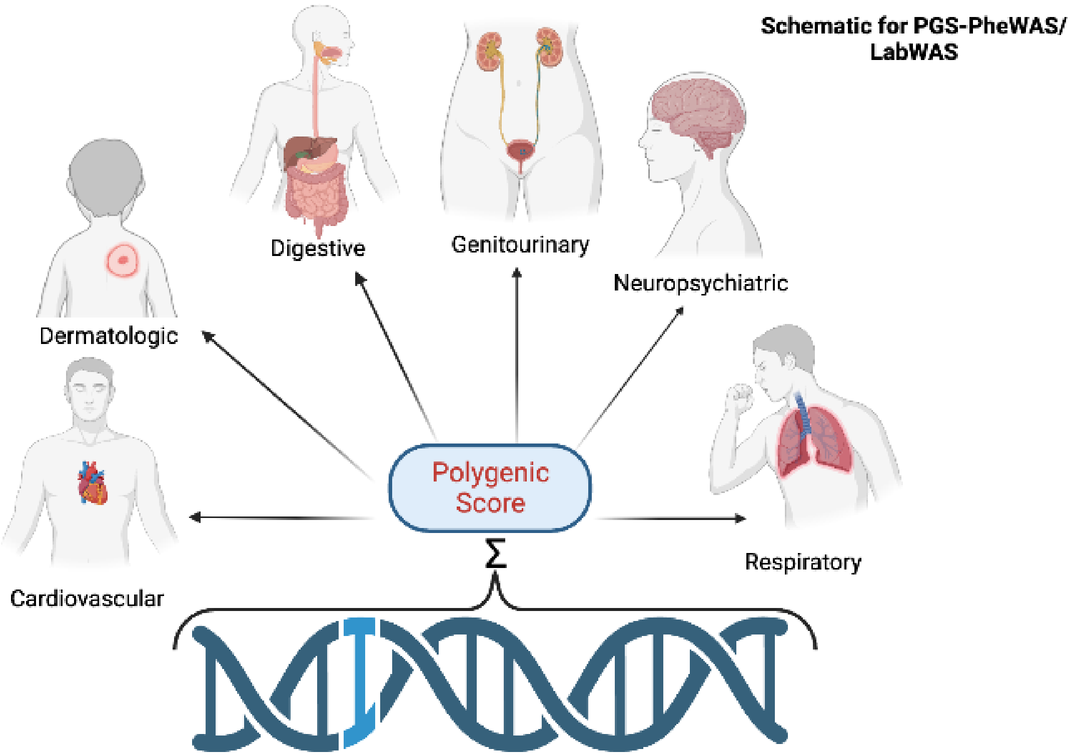

Schematic for PGS-PheWAS/
LabWAS
Digestive
Genitourinary
Neuropsychiatric
Dermatologic
Polygenic
Score
Cardiovascular
Respiratory
Σ

**Figure 2.2: SNP-PheWAS plot for rs6024489 in the European American GIA group showing top association with tobacco use disorder**

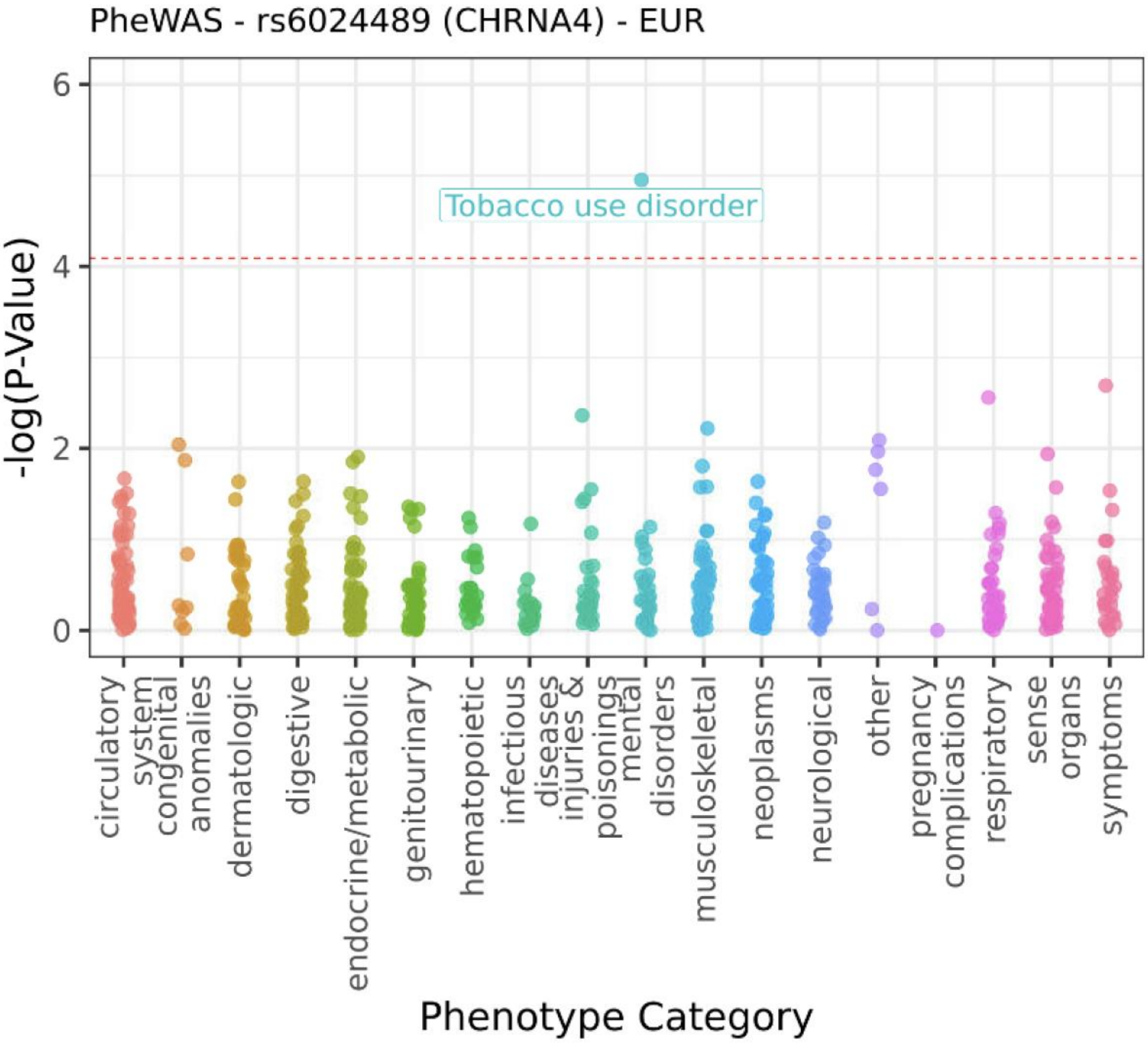


**Figure 2.3: PGS-PheWAS plot for TUD-PGS in the European American GIA group showing top associations with tobacco use disorder, obesity, diabetes mellitus, substance addiction disorders, and emphysema**

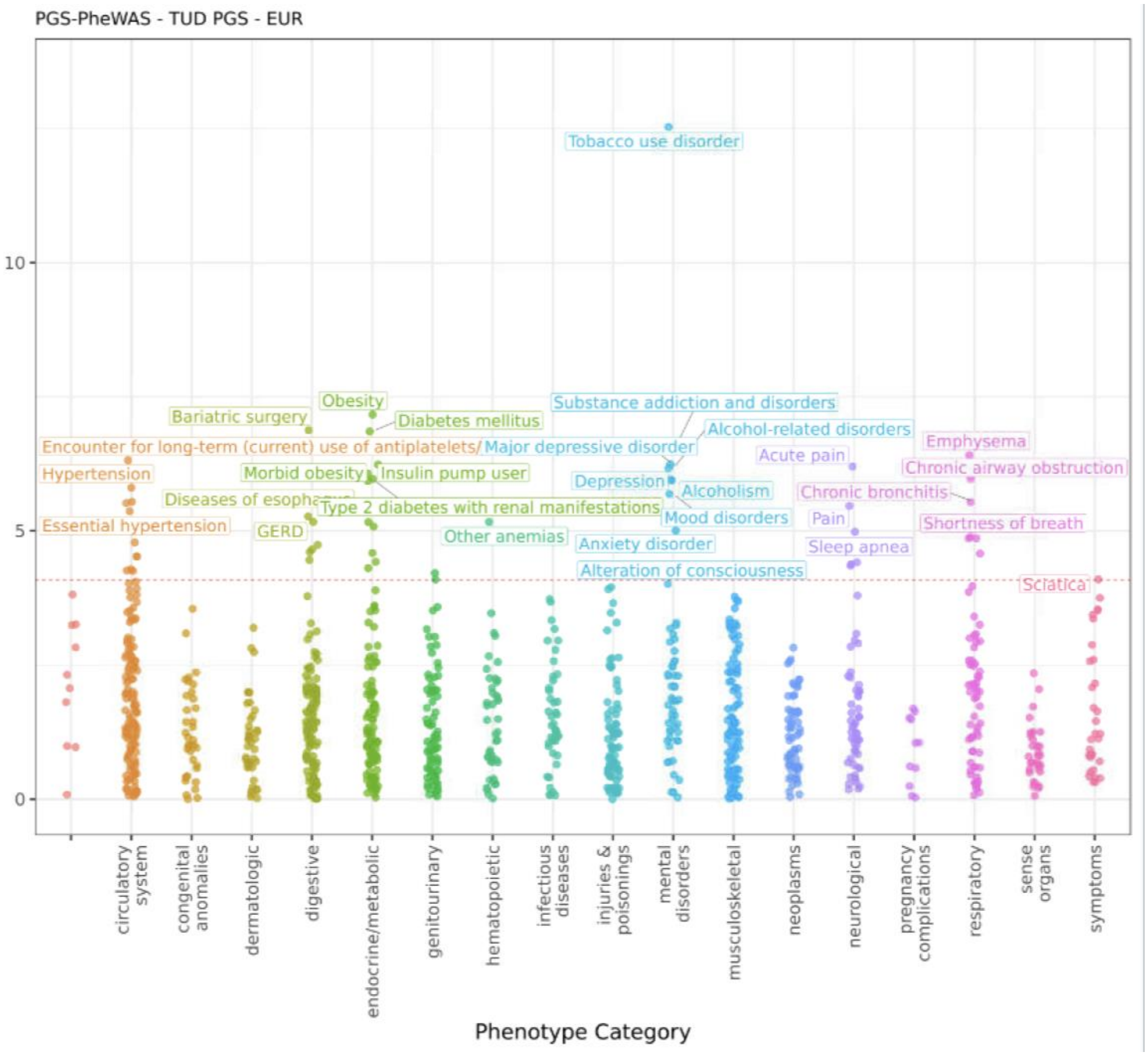
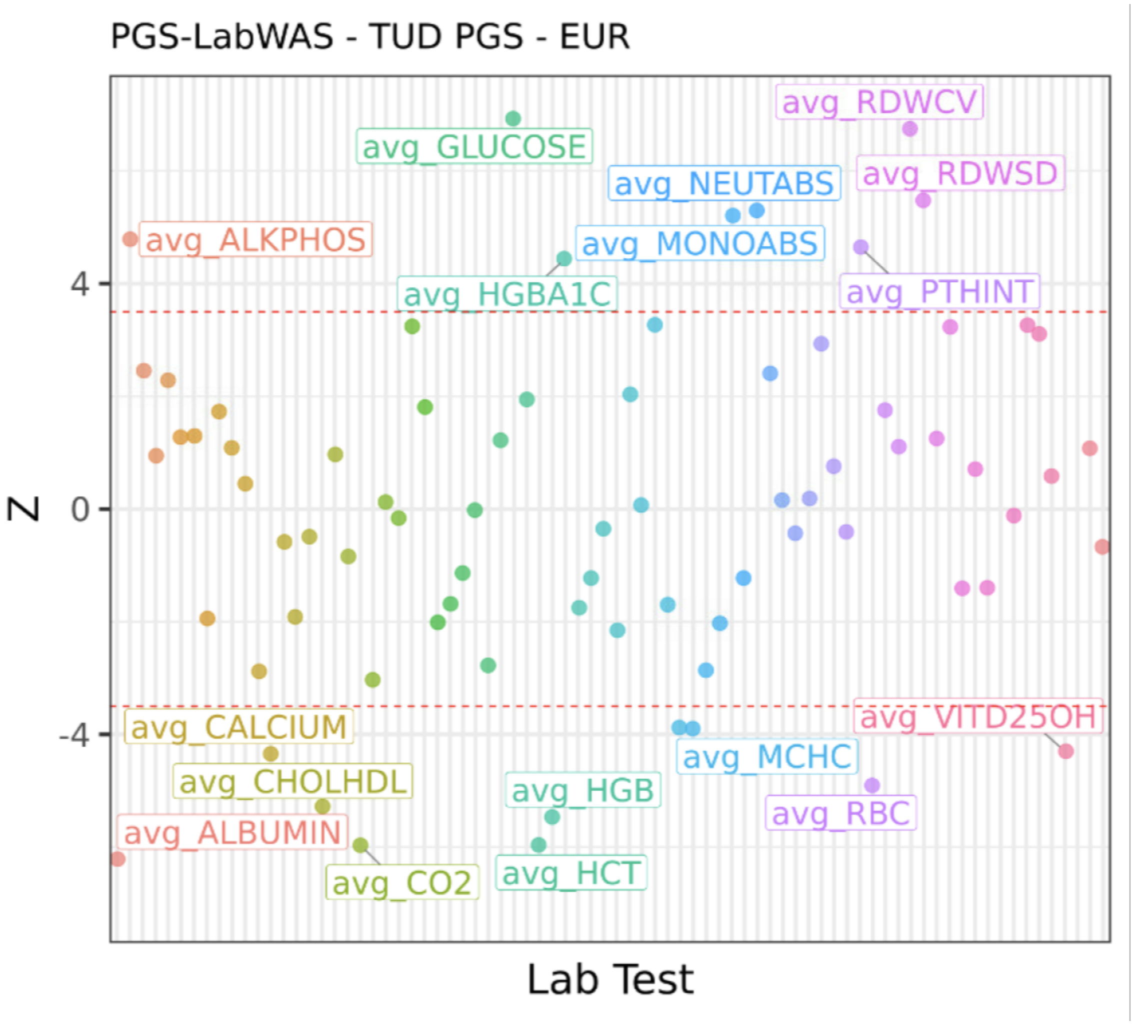


**Figure 2.4: PGS-LabWAS plot for TUD-PGS in the European American GIA group showing significant associations with laboratory tests for glucose, HDL, Vit D, Calcium**



## References

1. Wells QS, Gupta DK, Smith JG, et al. Accelerating Biomarker Discovery Through Electronic Health Records, Automated Biobanking, and Proteomics. *J Am Coll Cardiol*. 2019;73(17):2195-2205. doi:10.1016/j.jacc.2019.01.074

2. Linder JE, Bastarache L, Hughey JJ, Peterson JF. The Role of Electronic Health Records in Advancing Genomic Medicine. Annu Rev Genomics Hum Genet. 2021;22:219-238. doi:10.1146/annurev-genom-121120-125204

3. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205-1210. doi:10.1093/bioinformatics/btq126

4. Zhang X, Li X, He Y, et al. Phenome-wide association study (PheWAS) of colorectal cancer risk SNP effects on health outcomes in UK Biobank. Br J Cancer. 2022;126(5):822-830. doi:10.1038/s41416-021-01655-9

5. Denny JC, Crawford DC, Ritchie MD, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am J Hum Genet. 2011;89(4):529-542. doi:10.1016/j.ajhg.2011.09.008

6. Dennis JK, Sealock JM, Straub P, et al. Clinical laboratory test-wide association scan of polygenic scores identifies biomarkers of complex disease. Genome Med. 2021;13(1):6. Published 2021 Jan 13. doi:10.1186/s13073-020-00820-8

7. Goldstein JA, Weinstock JS, Bastarache LA, et al. LabWAS: Novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. PLoS Genet. 2020;16(11):e1009077. Published 2020 Nov 11. doi:10.1371/journal.pgen.1009077

8. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med*. 2020;12(1):44. Published 2020 May 18. doi:10.1186/s13073-020-00742-5

9. Chang TS, Ding Y, Freund MK, et al. Pre-existing conditions in Hispanics/Latinxs that are COVID-19 risk factors. iScience. 2021;24(3):102188. doi:10.1016/j.isci.2021.102188

10. Johnson R, Ding Y, Bhattacharya A, et al. The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. Cell Genom. 2023;3(1):100243. Published 2023 Jan 11. doi:10.1016/j.xgen.2022.100243

11. Johnson R, Ding Y, Venkateswaran V, et al. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative [published correction appears in Genome Med. 2022 Nov 16;14(1):128]. Genome Med. 2022;14(1):104. Published 2022 Sep 9. doi:10.1186/s13073-022-01106-x

12. G. Abraham, Y. Qiu, and M. Inouye, ``FlashPCA2: principal component analysis of biobank-scale genotype datasets'', (2017) Bioinformatics 33(17): 2776-2778. doi:10.1093/bioinformatics/btx299 (bioRxiv preprint https://doi.org/10.1101/094714)

13. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015). https://doi.org/10.1038/nature15393

14. Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annu Rev Biomed Data Sci. 2021;4:1-19. doi:10.1146/annurev-biodatasci-122320-112352

15. Hutchison KE, Allen DL, Filbey FM, et al. CHRNA4 and tobacco dependence: from gene regulation to treatment outcome. Arch Gen Psychiatry. 2007;64(9):1078-1086. doi:10.1001/archpsyc.64.9.1078

16. Privé F, Aschard H, Carmi S, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort [published correction appears in Am J Hum Genet. 2022 Feb 3;109(2):373]. Am J Hum Genet. 2022;109(1):12-23. doi:10.1016/j.ajhg.2021.11.008

17. The Python Language Reference. Python documentation. Accessed January 31, 2023. https://docs.python.org/3/reference/index.html

18. The Comprehensive R Archive Network. Accessed January 31, 2023. https://cran.r-project.org/

19. Prevention (US) C for DC and, Promotion (US) NC for CDP and H, Health (US) O on S and. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease. Centers for Disease Control and

Prevention (US); 2010. Accessed January 31, 2023. https://www.ncbi.nlm.nih.gov/books/NBK53017/

20. Roy A, Rawal I, Jabbour S, et al. Tobacco and Cardiovascular Disease: A Summary of Evidence. In: Prabhakaran D, Anand S, Gaziano TA, et al., editors. Cardiovascular, Respiratory, and Related Disorders. 3rd edition. Washington (DC): The International Bank for Reconstruction and Development / The World Bank; 2017 Nov 17. Chapter 4. Available from: https://www.ncbi.nlm.nih.gov/books/NBK525170/ doi: 10.1596/978-1-4648-0518-9_ch4

21. De Angelis F, Wendt FR, Pathak GA, et al. Drinking and smoking polygenic risk is associated with childhood and early-adulthood psychiatric and behavioral traits independently of substance use and psychiatric genetic risk. Transl Psychiatry. 2021;11(1):1-12. doi:10.1038/s41398-021-01713-z

22. Thorgeirsson TE, Gudbjartsson DF, Sulem P, et al. A common biological basis of obesity and nicotine addiction. Transl Psychiatry. 2013;3(10):e308. doi:10.1038/tp.2013.81

23. Maddatu J, Anderson-Baucum E, Evans-Molina C. Smoking and the risk of type 2 diabetes. *Transl Res*. 2017;184:101-107. doi:10.1016/j.trsl.2017.02.004

24. von Mühlen DG, Greendale GA, Garland CF, Wan L, Barrett-Connor E. Vitamin D, parathyroid hormone levels and bone mineral density in community-dwelling older women: the Rancho Bernardo Study. Osteoporos Int. 2005;16(12):1721-1726. doi:10.1007/s00198-005-1910-8

25. Brot C, Jorgensen NR, Sorensen OH. The influence of smoking on vitamin D status and calcium metabolism. Eur J Clin Nutr. 1999;53(12):920-926. doi:10.1038/sj.ejcn.1600870

26. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities [published correction appears in Nat Genet. 2021 May;53(5):763]. Nat Genet. 2019;51(4):584-591. doi:10.1038/s41588-019-0379-x

# Chapter 3

## Polygenic scores for tobacco use provide insights into systemic health risks in a diverse EHR-linked biobank in Los Angeles

### 3.1 Abstract

Tobacco use is a major risk factor for many diseases and is heavily influenced by environmental factors with significant underlying genetic contributions. Here, we evaluated the predictive performance, and potential systemic health effects of tobacco use disorder (TUD) predisposing germline variants using a polygenic score (PGS) in 24,202 participants from the hospital-based UCLA ATLAS biobank. Among genetically inferred ancestry groups (GIAs), TUD-PGS was significantly associated with TUD in European American (EA) (OR: 1.20, CI: [1.16, 1.24]), Hispanic/Latin American (HL) (OR:1.19, CI: [1.11, 1.28]), and East Asian American (EAA) (OR: 1.18, CI: [1.06, 1.31]) GIAs but not in African American (AA) GIA (OR: 1.04, CI: [0.93, 1.17]). In a cross-ancestry phenome-wide association meta-analysis, TUD-PGS was associated with cardiometabolic, respiratory, and psychiatric phecodes (17 phecodes at P < 2.7E-05). When restricted to never-smokers, the top TUD-PGS associations were obesity and alcohol-related disorders (P = 3.54E-07, 1.61E-06). Mendelian Randomization (MR) analysis provides evidence of a causal association between adiposity measures and tobacco use. Inconsistent predictive performance of the TUD-PGS across GIAs emphasizes the need to include participants of diverse ancestries at all levels of genetic research for equitable clinical translation of TUD-PGS. Our results suggest that TUD-

predisposed individuals may require comprehensive tobacco use prevention and management approaches to address underlying addictive tendencies.

## 3.2 Introduction

Tobacco use causes significant global mortality and morbidity, contributing to several systemic conditions, including cardiometabolic diseases and cancers[1,2]. Tobacco use could be viewed as a complex psychiatric trait with environmental risk factors[3] and genetic contributions[4,5]. Multi-ancestry genetic studies report an estimated SNP-based heritability of tobacco use behaviors ranging between 5-18%[4,5].  Twin and family studies report heritability estimates of 40%-56% for cigarette smoking and 72% for nicotine dependence. These heritability estimates vary widely between different tobacco use traits and between males and females[6]. Prevention and management strategies for tobacco use can benefit from precision medicine approaches, with the inclusion of baseline genetic risk to develop individualized preventive and therapeutic strategies. These efforts require a thorough understanding of the effects of a genetic predisposition to tobacco use and the impact of tobacco predisposition on the overall systemic health of an individual.

Researchers use genome-wide association studies (GWAS) to identify single nucleotide polymorphisms (SNPs) associated with tobacco use disorder. GWAS have identified over 2000 loci associated with tobacco use traits, such as smoking behaviors and nicotine dependence[4,5]. However, single variants rarely capture a large proportion of phenotypic variation for a complex behavioral trait like tobacco use.  polygenic scores (PGS) sum the weighted effects for multiple variants of interest, capturing a larger proportion of phenotypic variation than single variants. Polygenic scores have been used in research for disease prediction and to evaluate disease correlations, with the potential for clinical translation to identify high-risk individuals[7]. In particular, tobacco use behaviors have shown genetic correlations with diseases such as schizophrenia and substance use disorders[8–12].

Using the variants identified by GWAS, phenome-wide association studies (PheWAS) test the association of a single genetic variant across multiple phenotypes[13]. PheWAS identifies other traits or disorders upon which the single genetic variant could exert an effect. Generally, PheWAS use phenotypes that are identified using phecodes, which are ICD codes that are aggregated into clinically-meaningful groupings.

In our analysis, we combined a PGS for tobacco use disorder (TUD) with a PheWAS approach to create a PGS-PheWAS to examine the potential pleiotropic effects of multiple genetic variants that predispose to tobacco use disorder and identify systemic disease risks for individuals with a genetic predisposition to tobacco use[14]. We used a publicly available PGS for tobacco use disorder, developed in European-ancestry individuals in UK Biobank[15] and imputed these scores into the UCLA ATLAS biobank which comprises consented and genotyped UCLA patients representing diverse ancestry groups and phenotypes drawn from their electronic health records[16–20]. We found that the TUD-PGS demonstrated inconsistent predictive performance and risk stratification in non-European ancestry groups within the UCLA ATLAS biobank. In a PGS-PheWAS, we identified several phecodes associated with a genetic predisposition to tobacco use, mainly in cardiometabolic, respiratory, and neuropsychiatric phenotype categories. Next, to separate out the effects of tobacco use behavior from a genetic predisposition to tobacco use, we restricted a PGS-PheWAS to patients with no smoking history and identified persistent associations with obesity and alcohol-related disorders, suggesting shared genetic etiologies for these complex traits. Finally, we used publicly available GWAS summary statistics to perform Mendelian randomization[21] to evaluate the nature of the persistent tobacco use-obesity associations. We found evidence of causality between adiposity measures and tobacco use. Our work underscores the need to expand the diversity of study populations to generalize findings and to equitably translate genetic research to patient care. Further, the potential pleiotropic effects of tobacco-predisposing genetic variants suggest a more comprehensive approach to addressing tobacco use addiction that includes due consideration to other associated behavioral traits.

## 3.3 Methods

### 3.3.1 Study population

All analyses were performed with UCLA ATLAS Biobank data, a biobank embedded within the UCLA Health medical system[16–20]. UCLA Health is a comprehensive healthcare system serving the population in and around the greater Los Angeles area. The UCLA Institute for Precision Health is home to the UCLA ATLAS biobank with >40k participants genotyped, of which 24,202 participants were included in this study. This large-scale collection of genotyped biospecimens is integrated with the UCLA Data Discovery Repository (DDR), containing de-identified patient electronic health records (EHR) which include clinical, procedural, laboratory, prescription, and demographic information.

Final analyses included 24,202 ATLAS participants (7,902 cases and 16,283 controls) with complete information on the outcome and covariates including smoking status and insurance information. For ancestry-specific analysis, we included European American (N = 15,780), Hispanic/Latin American (N = 4,412), East Asian American (N = 2,377), and African American (N = 1,633) ancestry groups with sufficient sample sizes for analysis.

### 3.3.2 Data processing and population stratification

Detailed information on data processing can be found in previous publications[16–20]. Briefly, blood samples were collected from consented participants and genotyped on a custom array[22]. Initial array-level quality control measures included removing strand ambiguous SNPs and variants with >5% missingness and filtering out SNPs that do not pass the Hardy-Weinberg equilibrium test with a p-value set at ("–hwe 0.001"). After restricting to unrelated individuals, the QC-ed genotypes were imputed to the TOPMed Freeze5 reference using the Michigan Imputation Server[23,24]. The final QC steps were to filter the variants at the threshold of $R^2 > 0.90$ and minor allele frequency > 1%. All quality control steps were conducted using PLINK 1.9[25].

We computed the top 10 principal components for the study population using FlashPCA2 software[26]. We then grouped the study population into genetically inferred ancestry groups (GIAs) - European American (EA), Hispanic/Latin American (HL), East Asian American (EAA), African American (AA) - by k-nearest neighbor (KNN) stratification of the principal components, using the continental ancestry populations from 1000 Genomes Project[27,28] as a reference. To account for differences in population stratification between GIA groups, for the PGS-PheWAS analysis, we conducted individual PGS-PheWAS within each GIA group and then meta-analyzed across GIA groups to obtain cross-ancestry results.

### 3.3.3 Polygenic score imputation within ATLAS

We used a publicly available polygenic score trained on 391,124 European individuals (21954 cases and 357624 controls) from the UK biobank for the trait 'tobacco use disorder' from the PGS catalog (PGS002037)[15,29]. This trait, 'tobacco use disorder' was identified using phecode 318.0 which corresponds to ICD-codes F17.0, F17.1, F17.2, F17.3, F17.4, F17.9, Z72.0, 305.1, 305.10, 305.11, 305.12, 305.13, 649.0, 649.00, 649.01, 649.02, 649.03, 649.04 and V15.82. This PGS was selected for two reasons: (1) the PGS was trained on the same phecode for TUD that is available in ATLAS and (2) there is a high degree of overlap with ATLAS genotyped variants (800,381 of 847,691 total variants in TUD-PGS overlapping with ATLAS data - 94.4% overlap). The original PGS training analyses were performed using LDpred2[30] and adjusted for the following covariates: sex, age, birth date, Townsend's deprivation index, and the first 16 principal components of the genotype matrix. We computed the PGS for each ATLAS participant by multiplying the individual risk allele dosages by their corresponding weights that are provided by the PGS catalog[29]. The PGS was mean-centered and standardized by the standard deviation within each GIA group to generate a PGS Z-score.

We also tested the predictive performance of 16 multi-ancestry PGS from Saunders et al, Nature 2022 [5], trained on European, Admixed, East Asian and African ancestry populations for traits ‘Smoking initiation’, ‘Age of smoking initiation’, ‘Cigarettes smoked per day’ and, ‘Smoking cessation’. We downloaded these PGS (PGS003357-PGS003372) from the PGS Catalog[29] and tested their predictive performance on 4 genetically inferred ancestry groups within ATLAS for phecode 318.0 for tobacco use disorder, since we do not have information on the traits that the PGS were originally trained in.

### 3.3.4 Phecodes

ICD9 and ICD10 billing codes were aggregated into clinically meaningful groupings called phecodes using mappings derived from the PheWAS catalog, v1.2[31]. Cases were defined by the presence of an ICD code tagged by the respective phecode and controls by the absence of the ICD codes. Tobacco use disorder diagnosis was derived from the presence of “tobacco use disorder” phecode (318.00) within an individual’s health record which groups ICD codes (F17.200, F17.201, F17.210, F17.211, F17.220, F17.221, F17.290, F17.291, O99.33, O99.330, O99.331, O99.332, O99.333, O99.334, O99.335, Z87.891) for tobacco use disorder (TUD). For the PheWAS analysis, we used 1847 phecodes, extracted from each individual’s health record as described above, to capture phenotypes across the phenotypic spectrum[31].

### 3.3.5 Statistical Analysis

All analysis was conducted in either Python 2.6.8[32] or R 4.2.1[33].

a) *Predictive Performance and Risk Stratification*

We analyzed the predictive performance of the standardized TUD-PGS across ancestry groups and risk quantiles using GIA-stratified logistic regression models, with the phecode for TUD as the outcome and with predictors including terms for age, sex, the first five principal components of the genotype matrix, and insurance class.

We include insurance class information as the closest proxy to bias introduced by participation and access to healthcare within the de-identified electronic health records[34]. This insurance class variable consists of the type of insurance used by the patient in their clinical encounters. The classes include “Public”, “Private” or “Self-pay”. Public class includes ‘Medicare’, ‘Medicare Advantage’, ‘Medicare Assigned’, ‘Medi-Cal’, ‘Medicaid’, and ‘Medi-Cal Assigned’. Private class includes ‘International Payor’, ‘Group Health Plan’, ‘Worker's Comp’, ‘Tricare’, ‘UCLA Managed Care’, ‘Blue Shield’, ’Commercial’, ’Blue Cross’, ‘Package Billing’ and ‘Other’.

Odds ratios were calculated within each GIA, with P-values from Wald-type test statistics and a Bonferroni-corrected significance level of 0.0125 = (0.05/4). For risk stratification analysis, we grouped individuals of each GIA group into 5 groups of equal size based on their PGS and compared the quintile with the highest score with the quintile with the lowest scores. This model can be represented as

*Tobacco use disorder phecode (318.0) ~ PGS_Z (or) PGS quintile + Age + Sex + PCs1-5 + Insurance Class*

### b) <u>Phenome-wide association meta-analysis</u>

For the phenome-wide association analysis, we tested the association between the standardized TUD-PGS and 1847 electronic health record-derived phecodes across the phenome. Each GIA-specific PheWAS analysis consisted of logistic regressions across 1847 EHR-derived phecodes, controlling for age, sex, first 5 PCs, and insurance class. For the cross-ancestry meta-analysis, we use the PGS-PheWAS results computed within each GIA group and meta-analyze across these ancestry groups using a random effect, inverse variance weighted model using the metafor (version 3.4) package in R[35]. We use a phenome-wide Bonferroni-corrected p-value significance threshold of 2.7e-05 to adjust for the multiple testing burden (P = 0.05/1847 tests for each trait identified by phecodes). The never-smoker analysis followed a similar analysis plan, restricted to individuals of European American GIA with no history of smoking recorded by their provider within their medical records (n=9,921).

### *c) Mendelian Randomization*

We evaluated causality using Mendelian Randomization (MR) methods to test for and evaluate the causality between tobacco use and obesity[21]. We used summary statistics from GSCAN Consortium GWAS for “Cigarettes Smoked Per Day” (249,752 participants of European Ancestry and 12,003,613 SNPs)[36] and summary statistics from MRC Integrative Epidemiology Unit - the University of Bristol and UKBB GWAS for “Waist Circumference” (462,166 participants of European Ancestry and 9,851,867 SNPs)[37] as the instrumental variables to test the causal association between tobacco use and obesity measures. We performed a second MR analysis to validate the previous analysis using summary statistics for ‘Body Mass Index - BMI’ using summary statistics from UK Biobank[37] (461,460 individuals and 9,851,867 SNPs), using the same ‘Cigarettes smoked per day’ summary statistics from GSCAN as the outcome.

Lastly, GSCAN consortium and UK Biobank have approximately 35% sample overlap and hence we also tested this association using summary statistics for BMI from

GIANT consortium (322,154 individuals and 2,554,668 SNPs)[38]. We used the 'TwoSampleMR' R package to extract instruments, harmonize and obtain effect sizes from multiple MR methods (MR Egger, Weighted median, Inverse variance weighted, Simple mode, and Weighted mode)[39].

## 3.4 Results

### 3.4.1 Baseline characteristics of included ATLAS Biobank participants

The final analysis included n = 24,202 individuals with complete information on all covariates. Within the "TUD" phecode, the study population consisted of 7,902 cases and 16,283 controls. The average age of individuals with a TUD phecode was 64.3 years. Participant sex was significantly associated with TUD phecode with 55.1% of the phecode represented by the male sex. Four genetically inferred ancestry groups had sufficient sample size to perform the analyses: European American (EA), Hispanic/Latin American (HL), East Asian American (EAA), and African American ancestry (AA) (n=15,780, 4,412, 2,377, and 1,633, respectively). ***Table 3.1*** summarizes the demographics of the study sample.

### 3.4.2 Prediction and risk stratification of TUD using TUD-PGS across genetically inferred ancestry groups

We first evaluated how well the TUD-PGS predicts TUD across the multi-ancestry study sample within the ATLAS biobank. The TUD-PGS associated significantly with the phecode for TUD within the ATLAS biobank for individuals of European American (EA) GIA (OR:1.20, CI: [1.16, 1.24]), showing an increase in odds of TUD by 20% for each standard deviation increase in the TUD-PGS. Similarly, we observed significant associations between TUD-PGS and TUD among Hispanic/Latin American (HL) GIA (OR:1.19, CI: [1.11, 1.28]), and East Asian American (EAA) GIA groups (OR: 1.18, CI: [1.06, 1.31]). However, the TUD-PGS was not associated with TUD in individuals of African American (AA) GIA group (OR: 1.04, CI: [0.93, 1.17]). ***Supp Table 3.1*** summarizes these associations.

In addition, we used multi-ancestry PGS (PGS003357- PGS003372) and tested their predictive performance in the ancestry group corresponding to their training group. These PGS showed inconsistent albeit significant associations in EA GIA and insignificant associations in non-European GIAs with TUD in ATLAS (***Supp Table 3.2***)

Next, we assessed if the TUD-PGS could stratify individuals by risk for tobacco use disorder. Based on TUD-PGS, we divided the study sample into quintiles and estimated the odds ratio of TUD for each quintile compared to the bottom quintile. When compared to the quintile with the lowest TUD-PGS, the quintile with the highest TUD-PGS demonstrated an OR = 1.69 (CI: [1.51, 1.88]) in EA and 1.71 (CI: [1.36, 2.14]) in HL ancestry groups. The TUD-PGS offered strong risk stratification for individuals of EA GIA and for the top two risk quintiles in HL. Risk stratification was weaker and inconsistent in the EAA, (OR = 1.60, CI = [1.15, 2.24]) and AA ancestry groups (OR = 1.02, CI = [0.71, 1.47]) (***Fig 3.1, Supp Table 3.3***). This TUD-PGS risk stratifies individuals in EA and HL ancestry groups, potentially identifying individuals at a higher risk of tobacco use disorder within these ancestry groups. However, this risk stratification was inconsistent or absent in EAA and AA ancestry groups.

### 3.4.3 Systemic comorbidities in TUD-predisposed individuals identified by TUD-PGS-PheWAS

Next, we systematically evaluated associations between a genetic predisposition to TUD with 1847 traits or diseases across the phenome. The TUD-PGS captures the genetic predisposition to TUD and the 1847 traits are captured using phecodes extracted from each individual's electronic health record. In a PheWAS of the TUD-PGS across 1847 phecodes (***Supp Fig 3.1a***), meta-analyzed across 4 GIAs, we found 17 significant associations at Bonferroni-adjusted P < 0.05 after adjusting for age, sex, first 5 principal components of the genotype matrix, and health insurance information. The top phecodes associated with the TUD-PGS were 'morbid obesity', 'obstructive chronic bronchitis', 'substance addiction and disorders', and 'ischemic heart disease' (P = 1.38E-09, 2.73E-09, 4.45E-08, 1.61E-07) (***Fig 3.2a***). Phecode categories with multiple associations were circulatory (n=5), respiratory (n=3), neurological (n=2), and metabolic (n=2) phenotypes **(*Supp Table 3.4)*.** The results of this analysis systematically identify the health risks associated with a genetic predisposition to tobacco use captured by the PGS.

However, it must be noted that these associations may reflect the traits and diseases associated with tobacco use behavior, which lie on the TUD-PGS to trait/disease pathway (***Supp Fig 3.1b***). To study the potential pleiotropic effects of germline variants that predispose to TUD, we leveraged the fact that individuals with genetic predisposition to TUD may choose not to engage in tobacco use behaviors. We can thus account for the effect of tobacco use behavior to identify systemic risks of TUD genetic predisposition by stratifying to individuals with no smoking history recorded in their electronic health records. Accordingly, we repeated the PGS-PheWAS association analysis, restricting to “never-smokers” in individuals of EA ancestry, i.e. individuals who reported that they have never smoked tobacco (***Supp Fig 3.1b***). In this analysis, the TUD-PGS demonstrated associations with obesity, alcohol-related disorders, cancer of the esophagus, and hypertension (P = 3.54E-07, 1.61E-06, 3.05E-06, 2.62E-05) (***Figure 3.2b, Supp Table 3.5***).

In an evaluation of the trends of obesity and alcohol-related disorders across quintiles of the TUD-PGS, we observed higher ORs among never-smokers compared to ever-smokers for obesity and alcohol-related disorders. TUD-PGS offered inconsistent risk stratification for obesity and alcohol-related disorders in ever-smokers, or individuals with a history of smoking (***Figure 3.3***). In contrast, a reverse trend is noted in lung cancer, an established trait associated with smoking behavior, which can thus serve as a negative control, where we observed higher ORs in ever-smokers compared to never-smokers. (***Supp figure 3.2, Supp Table 3.6***) We can conclude from this analysis that, individuals predisposed to TUD show associations with obesity and alcohol-related disorder even in the absence of tobacco use behavior.

### 3.4.4 Mendelian randomization analysis finds evidence of causality in the association between obesity and tobacco use

To evaluate if the association between obesity and tobacco use can be given a directional and causal interpretation, we performed Mendelian randomization (MR) analysis between quantitative measures of obesity and tobacco use using publicly available GWAS of “waist circumference”[36] and “cigarettes smoked per day”[35]. From the results of multiple MR methods, we observed that the exposure “waist circumference” demonstrated significant positive causal associations with the outcome “cigarettes smoked per day” across all methods used to test this association (MR Egger, Weighted median, Inverse variance weighted, Simple mode, Weighted mode with P = 2.39E-03, 1.50E-32, 1.49E-46, 8.22E-05, 2.05E-08, respectively). A second MR analysis of “body mass index” as the exposure and “cigarettes smoked per day” as the outcome showed similar positive causal associations (MR Egger, Weighted median, Inverse variance weighted, Simple mode, Weighted mode P = 2.65E-03, 8.34E-33, 1.17E-45, 8.23E-06, 5.78E-07). An MR analysis of the reverse direction, with “cigarettes smoked per day” as the exposure and “waist circumference” and “body mass index” as outcomes did not show significant causal effects. ***Supp Figure 3.3a and b*** presents the causal effect estimates and confidence intervals. In a subsequent MR analysis in both directions using summary statistics for BMI from GIANT consortium, we find similar results, shown in ***Supp Table 3.7.***

## 3.5 Discussion

In this study, we examined the predictive performance and risk stratification of a publicly available TUD-PGS in the UCLA ATLAS biobank. We demonstrate that this TUD-PGS predicts TUD and risk-stratifies EA and HL GIA groups. However, inconsistent prediction and risk stratification was noted in the EAA and AA GIA groups.

There are two drawbacks to using this TUD-PGS clinically to identify individuals at high risk for tobacco use. First, the inconsistent predictive performance across GIA groups will result in inequitable clinical translation. Second, individual-level clinical decisions must be validated with clinical history in addition to genetic risk. At present, being classified as “high risk” by TUD-PGS is unreliable due to large uncertainty in imputed polygenic scores at an individual level[40]

Next, we evaluated the potential pleiotropic effects of TUD predisposing variants using the PGS to conduct a phenome-wide association analysis. Additionally, we repeated this analysis in a subgroup of participants without a reported history of smoking behavior, to evaluate the systemic associations of a genetic predisposition to tobacco use in the absence of tobacco use behavior. The PGS-PheWAS cross ancestry meta-analysis demonstrated significant associations with respiratory and cardiovascular phenotypes, both of which have robust clinical and biological evidence [41,42]. Other significant associations were in the category of psychiatric disorders, including associations with anxiety disorders and substance addiction disorders. These psychiatric disorder associations have been consistently reported in past genetic studies of smoking and tobacco use[43].

In the PGS-PheWAS analysis of never-smokers, phenotypes associated with tobacco use behaviors, namely, respiratory and cardiovascular phecodes, did not demonstrate statistical significance. Instead, we observed associations with psychiatric phecodes including alcohol-related disorders, and metabolic phecodes with potential behavioral contributions such as obesity. The MR analysis results suggest a causal association between adiposity and tobacco use, in line with other published literature with similar directionality and effect sizes[44]. Together, the associations between tobacco use, obesity, and alcoholism are suggestive of shared genetic architecture between these traits, likely originating from the biological regulation of impulsivity and addictive behaviors[45].

While this TUD-PGS cannot yet be translated clinically, these findings nevertheless have implications for patients with tobacco use disorder. We demonstrate the systemic comorbidities associated with a genetic propensity to TUD and that genetically predisposed individuals may be at risk for obesity and alcohol use disorder even when tobacco use behavior is absent. For patients in the TUD high-risk genetic propensity group, these findings would necessitate broadening the focus of the preventive and therapeutic strategy to include a more comprehensive regulation of biological pathways that underlie addiction and impulsivity.

A major strength of this study is that we evaluated TUD-PGS in an information-rich biobank across multiple genetically inferred ancestry groups. The rich phenotypic information available in the biobank allowed us to test associations across the phenome in a hypothesis-free manner, allowing for the discovery of disease associations. Another strength of the paper is that we accounted for possible confounding bias introduced by participation/access to healthcare bias, which can arise from using data from a hospital-based biobank, by using an insurance class variable as a proxy marker for participation and access.

Previous work has shown that PGS accuracy decreases linearly when there is a large difference in genetic ancestry between the training sample and the target sample. These differences in performance lead to bias and imprecision in risk stratification when PGS are applied clinically for complex traits such as TUD. Our results add to these results and motivate more sophisticated computational methods to improve the portability of PGS, especially for complex traits, like TUD, that are influenced greatly by both genetics and the environment and are risk factors for other diseases.

We conclude with limitations and future considerations of our work. Our study included a multi-ancestry sample of patients, but non-European ancestries are represented at smaller sample sizes for most analyses using the UCLA ATLAS biobank. With continued enrollment, we hope to increase the non-European sample sizes and evaluate differential genetic effects in these ancestries. Next, phecodes are derived from ICD codes which are billing codes and, accordingly, may not always capture the full extent of an individual’s disease history. The interpretations of our analyses are within the limitations of these phenotype definitions. We emphasize that the risk of having a phecode in the electronic health record does not accurately reflect the risk of having the disease. Phecode assignments come with biases, including access to healthcare. We have attempted to address this bias introduced by healthcare access by including an insurance class information variable. Nevertheless, this difference must be considered when applying these results to the general population. Lastly, the MR analysis has a partial sample overlap which might offer biased results. However, subsequent analysis with summary statistics from GWAS without sample overlap demonstrates similar results as the original MR analysis, supporting a conclusion of a potential causal association between measures of adiposity and tobacco use.

The results of our study have implications for public health and clinical approaches to the treatment of tobacco use disorder. Future research should strive to improve the prediction and risk stratification of TUD-PGS in all ancestry groups. With consistent performance across ancestry groups and improved individual-level prediction, TUD-PGS can be useful to identify individuals who can benefit from comprehensive preventive and therapeutic strategies to manage their underlying addictive tendencies. Given the growing evidence on health risks associated with obesity and tobacco use, our results suggest possible shared genetic etiology between these two risk factors, strengthening the argument that public health approaches must consider this shared risk while formulating interventions.

## 3.6 Tables

**Table 3.1: Baseline characteristics of ATLAS participants included in this study**

| | | **Overall** |
|---|---|---|
| **n** | | 24202 |
| **Age, median [Q1, Q3]** | | 61.0 [46.0,72.0] |
| **Sex, n (%)** | Female | 13277 (54.9) |
| | Male | 10914 (45.1) |
| **Insurance Class, n (%)** | Private | 14996 (62.0) |
| | Public | 8431 (34.8) |
| | Self-Pay | 775 (3.2) |
| **Tobacco Use Disorder, n (%)** | Controls | 16283 (67.3) |
| | Cases | 7902 (32.7) |
| **Genetically Inferred Ancestry, n (%)** | African American (AA) | 1633 (6.7) |
| | Hispanic/Latin American (HL) | 4412 (18.2) |
| | East Asian American (EAA) | 2377 (9.8) |
| | European American (EA) | 15780 (65.2) |

## 3.7 Figures

**Figure 3.1: TUD-PGS correlates with TUD phecode in EA, HL, and EAA ancestries across risk quintiles**

The X-axis represents the top 4 quintiles grouped according to TUD-PGS. Y axis represents effect sizes represented by odds ratios. The red line indicates OR = 1. Effect sizes between TUD-PGS and TUD phecode vary across PGS-quintiles in 4 genetically inferred ancestry groups with strong risk stratification noted in EA and HL and inconsistent risk stratification in AA and EAA groups.

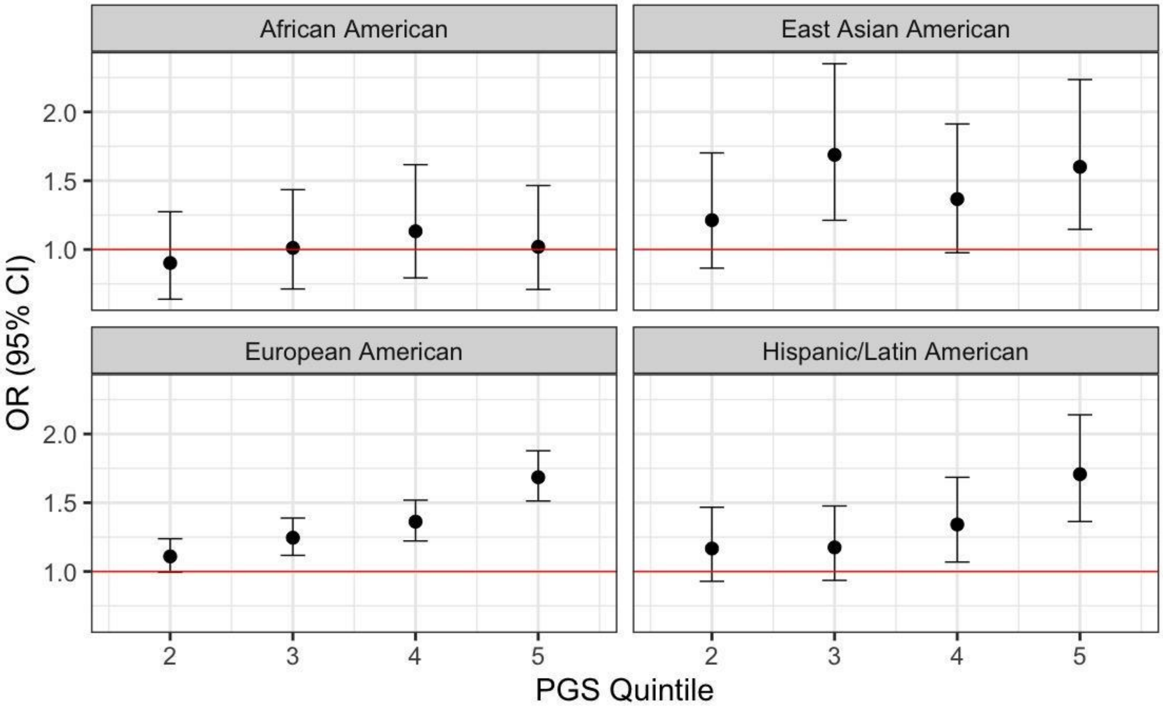


**Figure 3.2a: TUD-PGS-PheWAS plot across 1847 phecodes (cross-ancestry meta-analysis)**

Associations between TUD-PGS and 1847 phecodes across the phenome, meta-analyzed across 4 GIA groups with significant associations labeled. The X-axis represents the Z value (beta/SE). Each color represents a phecode category and each dot represents a phecode. Phenome-wide significance is represented by the red dashed line at a Z value = 4.2 which corresponds to a P value of 2.57e-5 (1847 tests/0.05). Top associations were noted in circulatory, metabolic, mental and respiratory phenotype categories.

**Figure 3.2b: TUD-PGS-PheWAS plot across 1847 phecodes in never-smokers of EA ancestry group**

Associations between TUD-PGS and 1847 phecodes across the phenome in never-smokers of EA ancestry with significant associations labeled. The X-axis represents the Z value (beta/SE). Each color represents a phecode category and each dot represents a phecode. Phenome-wide significance is represented by the red dashed line at a Z value = 4.2 which corresponds to a P value of 2.57e-5 (1847 tests/0.05). In TUD-PGS-PheWAS restricted to ‘never smokers’, top associations were obesity and alcohol-related disorders.

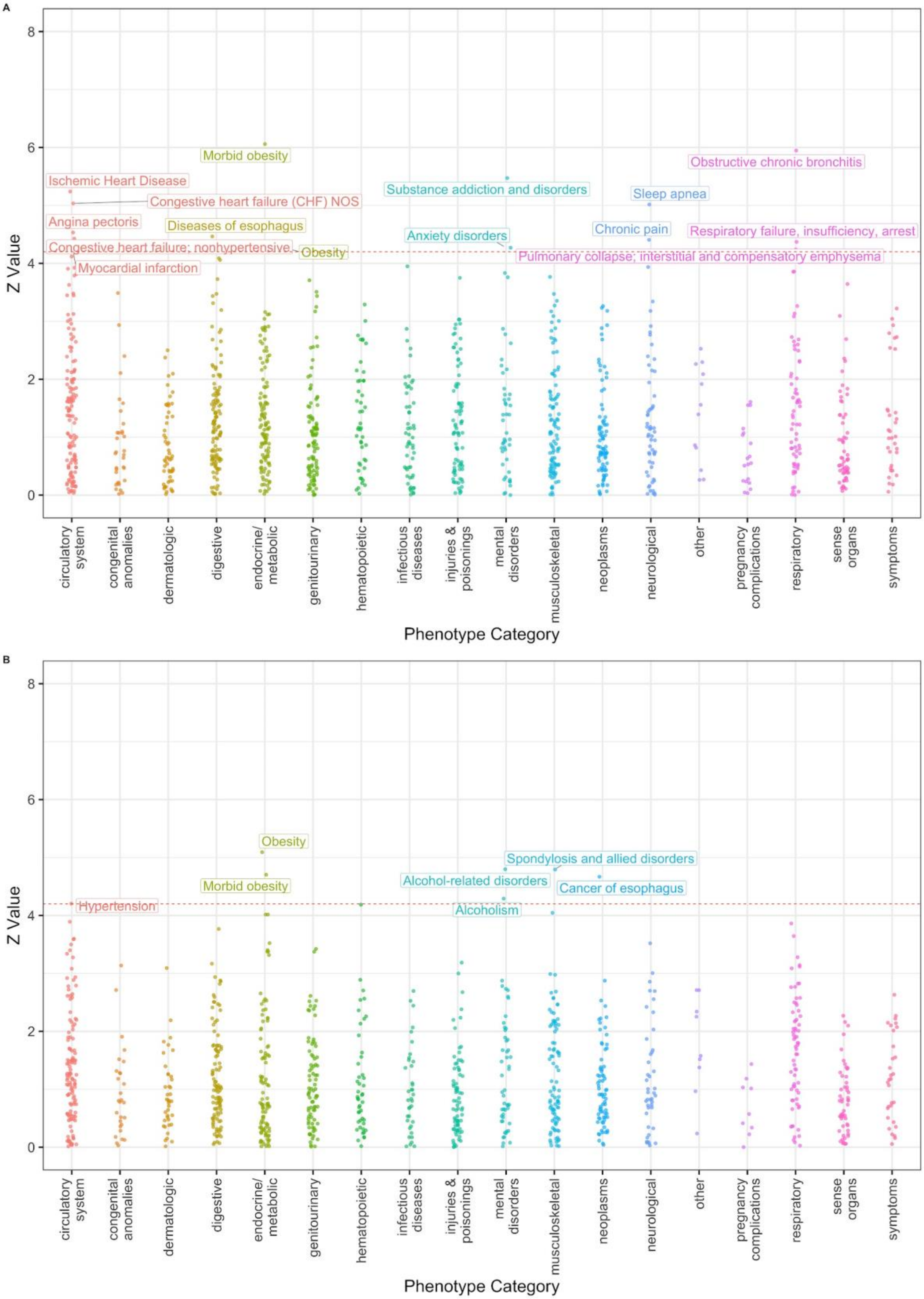
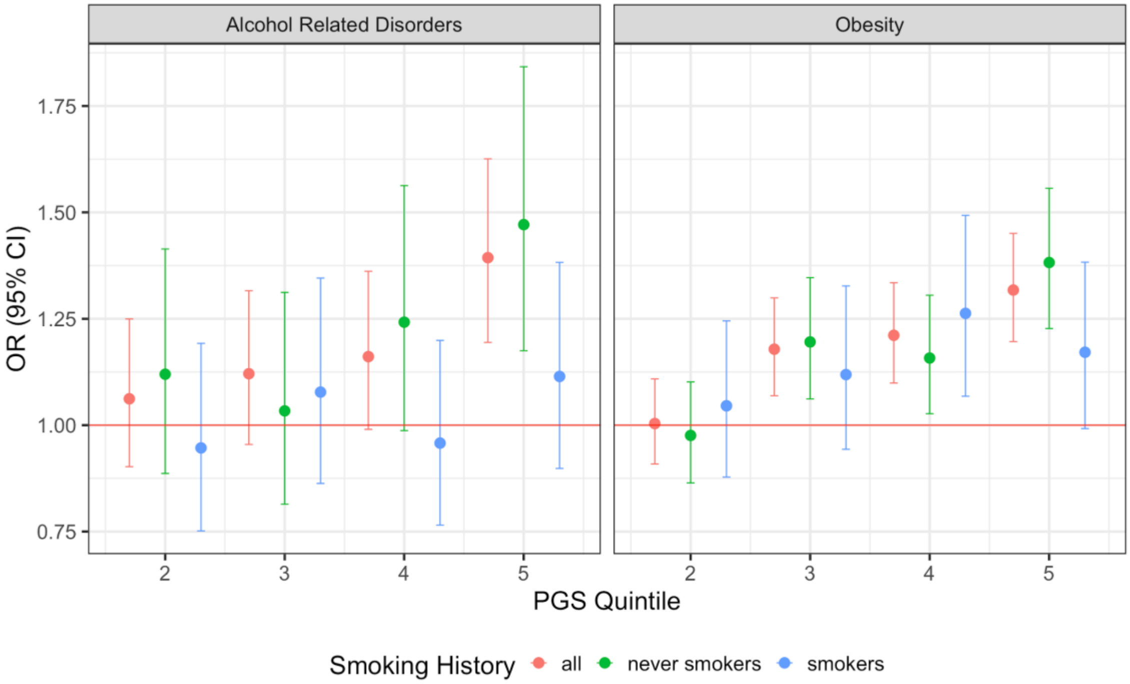

A
Z Value
8
6
4
2
0
Morbid obesity
Ischemic Heart Disease
Congestive heart failure (CHF) NOS
Angina pectoris
Diseases of esophagus
Congestive heart failure; nonhypertensive
Obesity
Myocardial infarction
Substance addiction and disorders
Anxiety disorders
Sleep apnea
Chronic pain
Obstructive chronic bronchitis
Respiratory failure, insufficiency, arrest
Pulmonary collapse; interstitial and compensatory emphysema
circulatory system
congenital anomalies
dermatologic
digestive
endocrine/ metabolic
genitourinary
hematopoietic
infectious diseases
injuries & poisonings
mental disorders
musculoskeletal
neoplasms
neurological
other
pregnancy complications
respiratory
sense organs
symptoms
Phenotype Category
B
Z Value
8
6
4
2
0
Obesity
Morbid obesity
Hypertension
Spondylosis and allied disorders
Alcohol-related disorders
Cancer of esophagus
Alcoholism
circulatory system
congenital anomalies
dermatologic
digestive
endocrine/ metabolic
genitourinary
hematopoietic
infectious diseases
injuries & poisonings
mental disorders
musculoskeletal
neoplasms
neurological
other
pregnancy complications
respiratory
sense organs
symptoms
Phenotype Category

**Figure 3.3: TUD-PGS associations with Alcohol-related disorders and Obesity among all vs ever vs never-smokers across TUD-PGS quintiles**

Associations between TUD-PGS quintiles and Alcohol-related disorders (phecode = 317.0) and Obesity (phecode = 278.1). The X-axis represents the top 4 quintiles grouped according to TUD-PGS. Y axis represents effect sizes represented by odds ratios. The red line indicates OR =1. TUD-PGS risk-stratifies for the phecodes for alcohol-related disorders and obesity in 'never smokers' but not in 'ever-smokers'.



## 3.8 Supplementary Tables

### Supplementary Table 3.1: TUD-PGS association with TUD across GIAs

| GIA | β | SE | P-Value | [0.025 | 0.975] | OR | OR_lower_CI | OR_upper_CI |
|---|---|---|---|---|---|---|---|---|
| European American | 0.2 | 0.02 | 1.66E-25 | 0.15 | 0.22 | 1.20 | 1.16 | 1.24 |
| Hispanic/Latin American | 0.2 | 0.04 | 2.24E-06 | 0.10 | 0.24 | 1.19 | 1.11 | 1.28 |
| East Asian American | 0.2 | 0.05 | 1.93E-03 | 0.06 | 0.27 | 1.18 | 1.06 | 1.31 |
| African American | 0.04 | 0.06 | 5.07E-01 | -0.08 | 0.16 | 1.04 | 0.93 | 1.17 |

### Supplementary Table 3.2: Evaluation of predictive performance of 16 multi ancestry PGS for phecode for 'tobacco use disorder' (TUD) in ATLAS: Associations of PGS002037 (trained for TUD) for TUD in ATLAS

| Trait | Training Pop | Testing Pop in ATLAS | PGS | coef | SE | P>\|z\| | 0.025 | 0.975 |
|---|---|---|---|---|---|---|---|---|
| TUD | European | European American | PGS002037 | 0.20 | 0.02 | <0.0001 | 0.16 | 0.24 |
| TUD | European | Admixed American | PGS002037 | 0.21 | 0.04 | <0.0001 | 0.13 | 0.29 |
| TUD | European | East Asian American | PGS002037 | 0.20 | 0.07 | 0.003 | 0.07 | 0.33 |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| TUD | European | African American | PGS002037 | 0.04 | 0.08 | 0.591 | -0.12 | 0.20 |

### Associations of the multi-ancestry PGS for TUD - Trained in European ancestry, tested in European ancestry in ATLAS

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| Smoking initiation | European | European American | PGS003360 | -0.3 | 0.02 | <0.0001 | -0.37 | -0.30 |
| Age of smoking initiation | European | European American | PGS003364 | 0.1 | 0.02 | <0.0001 | 0.10 | 0.17 |
| Cigarettes smoked per day | European | European American | PGS003368 | -0.1 | 0.02 | <0.0001 | -0.15 | -0.08 |
| Smoking cessation | European | European American | PGS003372 | -0.1 | 0.02 | <0.0001 | -0.14 | -0.07 |

### Associations of the multi-ancestry PGS for TUD- Trained in Admixed ancestry, tested in Admixed American ancestry in ATLAS

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| Smoking initiation | Admixed | Admixed American | PGS003358 | -0.09 | 0.05 | 0.06 | -0.19 | 0.01 |
| Age of smoking initiation | Admixed | Admixed American | PGS003362 | -0.06 | 0.04 | 0.1 | -0.14 | 0.01 |
| Cigarettes smoked per day | Admixed | Admixed American | PGS003366 | 0.01 | 0.04 | 0.8 | -0.066 | 0.08 |
| Smoking cessation | Admixed | Admixed American | PGS003370 | 0.024 | 0.04 | 0.6 | -0.06 | 0.11 |

**Associations of the multi-ancestry PGS for TUD - Trained in East Asian ancestry, tested in East Asian American ancestry in ATLAS**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Smoking initiation | East Asian | East Asian American | PGS003359 | -0.2 | 0.06 | <0.0001 | -0.33 | -0.11 |
| Age of smoking initiation | East Asian | East Asian American | PGS003363 | -0.03 | 0.06 | 0.6 | -0.14 | 0.08 |
| Cigarettes smoked per day | East Asian | East Asian American | PGS003367 | -0.05 | 0.06 | 0.4 | -0.16 | 0.06 |
| Smoking cessation | East Asian | East Asian American | PGS003371 | -0.08 | 0.05 | 0.1 | -0.18 | 0.02 |

**Associations of the multi-ancestry PGS for TUD - Trained in African ancestry, tested in African American ancestry in ATLAS**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Smoking initiation | African | African American | PGS003357 | -0.08 | 0.10 | 0.46 | -0.28 | 0.13 |
| Age of smoking initiation | African | African American | PGS003361 | -0.01 | 0.07 | 0.94 | -0.15 | 0.14 |
| Cigarettes smoked per day | African | African American | PGS003365 | 0.09 | 0.06 | 0.1 | -0.02 | 0.20 |
| Smoking cessation | African | African American | PGS003369 | -0.18 | 0.06 | 0.004 | -0.29 | -0.06 |

**Supplementary Table 3.3: TUD-PGS association with TUD across quintiles and GIAs**

| PGS_ Quantile | Coef | SE | P>\|z\| | 0.025 | 0.975 | OR | OR_ lower_CI | OR_ upper_CI | GIA |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.10 | 0.06 | 6.27E-02 | -0.005 | 0.21 | 1.11 | 0.99 | 1.24 | EA |
| 3 | 0.22 | 0.06 | 7.70E-05 | 0.11 | 0.33 | 1.25 | 1.12 | 1.39 | EA |
| 4 | 0.31 | 0.06 | 2.47E-08 | 0.20 | 0.42 | 1.36 | 1.22 | 1.52 | EA |
| 5 | 0.52 | 0.06 | 3.25E-21 | 0.41 | 0.63 | 1.69 | 1.51 | 1.88 | EA |
| 2 | 0.15 | 0.12 | 1.86E-01 | -0.07 | 0.38 | 1.17 | 0.93 | 1.47 | HL |
| 3 | 0.16 | 0.12 | 1.66E-01 | -0.07 | 0.39 | 1.18 | 0.94 | 1.48 | HL |
| 4 | 0.29 | 0.12 | 1.13E-02 | 0.07 | 0.52 | 1.34 | 1.07 | 1.68 | HL |
| 5 | 0.54 | 0.12 | 3.27E-06 | 0.31 | 0.76 | 1.71 | 1.36 | 2.14 | HL |
| 2 | 0.19 | 0.17 | 2.64E-01 | -0.15 | 0.53 | 1.21 | 0.86 | 1.70 | EAA |
| 3 | 0.52 | 0.17 | 1.94E-03 | 0.19 | 0.85 | 1.69 | 1.21 | 2.35 | EAA |
| 4 | 0.31 | 0.17 | 6.87E-02 | -0.02 | 0.65 | 1.37 | 0.98 | 1.91 | EAA |
| 5 | 0.47 | 0.17 | 5.77E-03 | 0.14 | 0.80 | 1.60 | 1.15 | 2.23 | EAA |
| 2 | -0.10 | 0.18 | 5.58E-01 | -0.45 | 0.24 | 0.90 | 0.63 | 1.27 | AA |
| 3 | 0.01 | 0.18 | 9.49E-01 | -0.34 | 0.36 | 1.01 | 0.71 | 1.43 | AA |
| 4 | 0.12 | 0.18 | 4.93E-01 | -0.23 | 0.48 | 1.13 | 0.79 | 1.62 | AA |
| 5 | 0.02 | 0.18 | 9.18E-01 | -0.34 | 0.38 | 1.02 | 0.71 | 1.47 | AA |

**Supplementary Table 3.4: Significant associations between TUD-PGS and 1847 traits in the PGS-PheWAS cross-ancestry meta-analysis**

| Phecode | beta | SE | P Value | CI.LB | CI.UB | Phenotype | Category |
|---|---|---|---|---|---|---|---|
| 278.11 | 0.12 | 0.02 | 1.38E-09 | 0.08 | 0.17 | Morbid obesity | endocrine/metabolic |
| 496.21 | 0.25 | 0.04 | 2.73E-09 | 0.17 | 0.33 | Obstructive chronic bronchitis | respiratory |
| 316 | 0.12 | 0.02 | 4.45E-08 | 0.08 | 0.16 | Substance addiction and disorders | mental disorders |
| 411 | 0.09 | 0.02 | 1.61E-07 | 0.05 | 0.12 | Ischemic Heart Disease | circulatory system |
| 228.1 | -0.10 | 0.02 | 3.49E-07 | -0.14 | -0.06 | Hemangioma of skin and subcutaneous tissue | neoplasms |
| 428.1 | 0.12 | 0.02 | 4.80E-07 | 0.07 | 0.16 | Congestive heart failure (CHF) NOS | circulatory system |
| 327.3 | 0.08 | 0.02 | 5.29E-07 | 0.05 | 0.11 | Sleep apnea | neurological |
| 228 | -0.09 | 0.02 | 1.74E-06 | -0.13 | -0.05 | Hemangioma and lymphangioma, any site | neoplasms |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 411.3 | 0.12 | 0.03 | 5.83E-06 | 0.07 | 0.17 | Angina pectoris | circulatory system |
| 530 | 0.06 | 0.01 | 8.05E-06 | 0.03 | 0.09 | Diseases of esophagus | digestive |
| 428 | 0.09 | 0.02 | 9.57E-06 | 0.05 | 0.14 | Congestive heart failure; nonhypertensive | circulatory system |
| 338.2 | 0.06 | 0.01 | 1.05E-05 | 0.03 | 0.09 | Chronic pain | neurological |
| 509 | 0.09 | 0.02 | 1.24E-05 | 0.04817 | 0.13 | Respiratory failure, insufficiency, arrest | respiratory |
| 278.1 | 0.11 | 0.03 | 1.42E-05 | 0.064552 | 0.17 | Obesity | endocrine/metabolic |
| 411.2 | 0.11 | 0.03 | 1.45E-05 | 0.060558 | 0.16 | Myocardial infarction | circulatory system |
| 300 | 0.06 | 0.01 | 1.95E-05 | 0.031596 | 0.09 | Anxiety disorders | mental disorders |
| 508 | 0.07 | 0.02 | 2.23E-05 | 0.035846 | 0.097 | Pulmonary collapse; interstitial and compensatory emphysema | respiratory |

**Supplementary Table 3.5: Significant associations between TUD-PGS and 1847 traits in the 'never-smoker' PGS-PheWAS in EA ancestry group**

| Phecode | Coef | SE | P Value | 0.025 | 0.975 | Phenotype | Category |
|---|---|---|---|---|---|---|---|
| 278.1 | 0.13 | 0.03 | 3.54E-07 | 0.08 | 0.18 | Obesity | endocrine/metabolic |
| 317 | 0.23 | 0.05 | 1.61E-06 | 0.14 | 0.32 | Alcohol-related disorders | mental disorders |
| 721 | 0.12 | 0.03 | 1.64E-06 | 0.074 | 0.17 | Spondylosis and allied disorders | musculoskeletal |
| 278.11 | 0.16 | 0.03 | 2.56E-06 | 0.09 | 0.23 | Morbid obesity | endocrine/metabolic |
| 150 | 0.71 | 0.15 | 3.05E-06 | 0.41 | 1.01 | Cancer of esophagus | neoplasms |
| 228 | -0.12 | 0.03 | 4.67E-06 | -0.17 | -0.07 | Hemangioma and lymphangioma, any site | neoplasms |
| 317.1 | 0.23 | 0.05 | 1.78E-05 | 0.12 | 0.33 | Alcoholism | mental disorders |
| 401 | 0.09 | 0.02 | 2.62E-05 | 0.05 | 0.14 | Hypertension | circulatory system |

**Supplementary Table 3.6: Associations between Alcohol-Related Disorders, Obesity, and Lung cancer and PGS quantiles traits in the 'ever-smoker' and 'never-smoker' groups**

| PGS_Quantile | SE | P>\|z\| | Phecode | Smoking History | OR | OR_Lower_CI | OR_Upper_CI | Phenotype |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.05 | 9.40E-01 | 278.1 | all | 1.00 | 0.91 | 1.11 | Obesity |
| 3 | 0.05 | 9.38E-04 | 278.1 | all | 1.18 | 1.07 | 1.30 | Obesity |
| 4 | 0.05 | 1.11E-04 | 278.1 | all | 1.21 | 1.1 | 1.33 | Obesity |
| 5 | 0.05 | 2.02E-08 | 278.1 | all | 1.32 | 1.2 | 1.45 | Obesity |
| 2 | 0.09 | 6.17E-01 | 278.1 | smokers | 1.05 | 0.88 | 1.25 | Obesity |
| 3 | 0.09 | 1.97E-01 | 278.1 | smokers | 1.12 | 0.94 | 1.33 | Obesity |
| 4 | 0.09 | 6.34E-03 | 278.1 | smokers | 1.26 | 1.07 | 1.49 | Obesity |
| 5 | 0.08 | 6.22E-02 | 278.1 | smokers | 1.17 | 0.99 | 1.38 | Obesity |
| 2 | 0.06 | 6.93E-01 | 278.1 | never smokers | 0.98 | 0.86 | 1.10 | Obesity |
| 3 | 0.06 | 3.22E-03 | 278.1 | never smokers | 1.20 | 1.06 | 1.35 | Obesity |
| 4 | 0.06 | 1.67E-02 | 278.1 | never smokers | 1.16 | 1.03 | 1.31 | Obesity |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 5 | 0.06 | 9.73E-08 | 278.1 | never smokers | 1.38 | 1.23 | 1.56 | Obesity |
| 2 | 0.08 | 4.69E-01 | 317 | all | 1.06 | 0.90 | 1.25 | Alcohol Related Disorders |
| 3 | 0.08 | 1.63E-01 | 317 | all | 1.12 | 0.95 | 1.32 | Alcohol Related Disorders |
| 4 | 0.08 | 6.60E-02 | 317 | all | 1.16 | 0.99 | 1.36 | Alcohol Related Disorders |
| 5 | 0.08 | 2.44E-05 | 317 | all | 1.39 | 1.19 | 1.63 | Alcohol Related Disorders |
| 2 | 0.12 | 6.41E-01 | 317 | smokers | 0.95 | 0.75 | 1.19 | Alcohol Related Disorders |
| 3 | 0.11 | 5.09E-01 | 317 | smokers | 1.08 | 0.86 | 1.35 | Alcohol Related Disorders |
| 4 | 0.11 | 7.08E-01 | 317 | smokers | 0.96 | 0.77 | 1.20 | Alcohol Related Disorders |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 5 | 0.11 | 3.24E-01 | 317 | smokers | 1.11 | 0.90 | 1.38 | Alcohol Related Disorders |
| 2 | 0.12 | 3.42E-01 | 317 | never smokers | 1.12 | 0.88 | 1.41 | Alcohol Related Disorders |
| 3 | 0.12 | 7.85E-01 | 317 | never smokers | 1.03 | 0.81 | 1.31 | Alcohol Related Disorders |
| 4 | 0.12 | 6.42E-02 | 317 | never smokers | 1.24 | 0.99 | 1.56 | Alcohol Related Disorders |
| 5 | 0.11 | 7.64E-04 | 317 | never smokers | 1.47 | 1.17 | 1.84 | Alcohol Related Disorders |
| 2 | 0.11 | 4.65E-01 | 165.1 | all | 1.09 | 0.87 | 1.36 | Lung Cancer |
| 3 | 0.11 | 2.39E-01 | 165.1 | all | 1.14 | 0.91 | 1.42 | Lung Cancer |
| 4 | 0.11 | 9.08E-01 | 165.1 | all | 1.01 | 0.81 | 1.27 | Lung Cancer |
| 5 | 0.11 | 5.50E-01 | 165.1 | all | 1.07 | 0.85 | 1.34 | Lung Cancer |
| 2 | 0.16 | 7.82E-01 | 165.1 | smokers | 1.05 | 0.76 | 1.44 | Lung Cancer |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| 3 | 0.15 | 1.15E-01 | 165.1 | smokers | 1.28 | 0.94 | 1.73 | Lung Cancer |
| 4 | 0.16 | 8.13E-01 | 165.1 | smokers | 1.04 | 0.76 | 1.42 | Lung Cancer |
| 5 | 0.16 | 4.52E-01 | 165.1 | smokers | 1.12 | 0.83 | 1.53 | Lung Cancer |
| 2 | 0.16 | 6.18E-01 | 165.1 | never smokers | 1.08 | 0.79 | 1.49 | Lung Cancer |
| 3 | 0.17 | 6.40E-01 | 165.1 | never smokers | 0.92 | 0.66 | 1.29 | Lung Cancer |
| 4 | 0.17 | 5.88E-01 | 165.1 | never smokers | 0.91 | 0.65 | 1.27 | Lung Cancer |
| 5 | 0.18 | 3.81E-01 | 165.1 | never smokers | 0.86 | 0.60 | 1.21 | Lung Cancer |

**Supplementary Table 3.7: Mendelian Randomization results between Cigarettes smoked per day (GSCAN) and Body Mass Index (GIANT Consortium)**

| outcome | exposure | method | nsnp | b | se | pval |
|---|---|---|---|---|---|---|
| Body mass index \|\| id:ieu-a-835 | Cigarettes smoked per day \|\| id:ieu-b-142 | MR Egger | 16 | -0.082 | 0.044 | 0.083 |
| Body mass index \|\| id:ieu-a-835 | Cigarettes smoked per day \|\| id:ieu-b-142 | Weighted median | 16 | -0.046 | 0.021 | 0.032 |
| Body mass index \|\| | Cigarettes smoked | Inverse | 16 | - | 0.029 | 0.622 |

| | | | | | | |
|---|---|---|---|---|---|---|
| id:ieu-a-835 | per day \|\| id:ieu-b-142 | variance weighted | | 0.014 | | |
| Body mass index \|\| id:ieu-a-835 | Cigarettes smoked per day \|\| id:ieu-b-142 | Simple mode | 16 | 0.060 | 0.065 | 0.372 |
| Body mass index \|\| id:ieu-a-835 | Cigarettes smoked per day \|\| id:ieu-b-142 | Weighted mode | 16 | -0.049 | 0.021 | 0.0360 |
| Cigarettes smoked per day \|\| id:ieu-b-142 | Body mass index \|\| id:ieu-a-835 | MR Egger | 65 | 0.23 | 0.123 | 0.070 |
| Cigarettes smoked per day \|\| id:ieu-b-142 | Body mass index \|\| id:ieu-a-835 | Weighted median | 65 | 0.33 | 0.047 | 1.29E-12 |
| Cigarettes smoked per day \|\| id:ieu-b-142 | Body mass index \|\| id:ieu-a-835 | Inverse variance weighted | 65 | 0.26 | 0.042 | 2.93E-10 |
| Cigarettes smoked per day \|\| id:ieu-b-142 | Body mass index \|\| id:ieu-a-835 | Simple mode | 65 | 0.47 | 0.127 | 4.44e-04 |
| Cigarettes smoked per day \|\| id:ieu-b-142 | Body mass index \|\| id:ieu-a-835 | Weighted mode | 65 | 0.45 | 0.086 | 1.87E-06 |

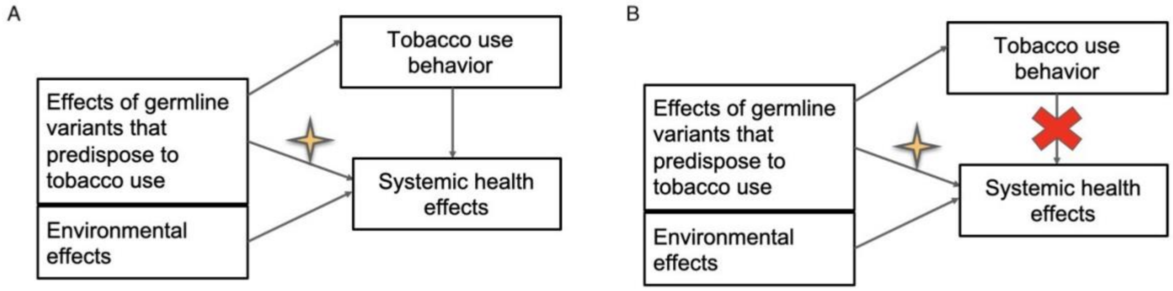
## 3.9 Supplementary Figures

**Supplementary Figures 3.1A and B**

a) **DAG showing the relationship evaluated in the PGS-PheWAS meta-analysis**

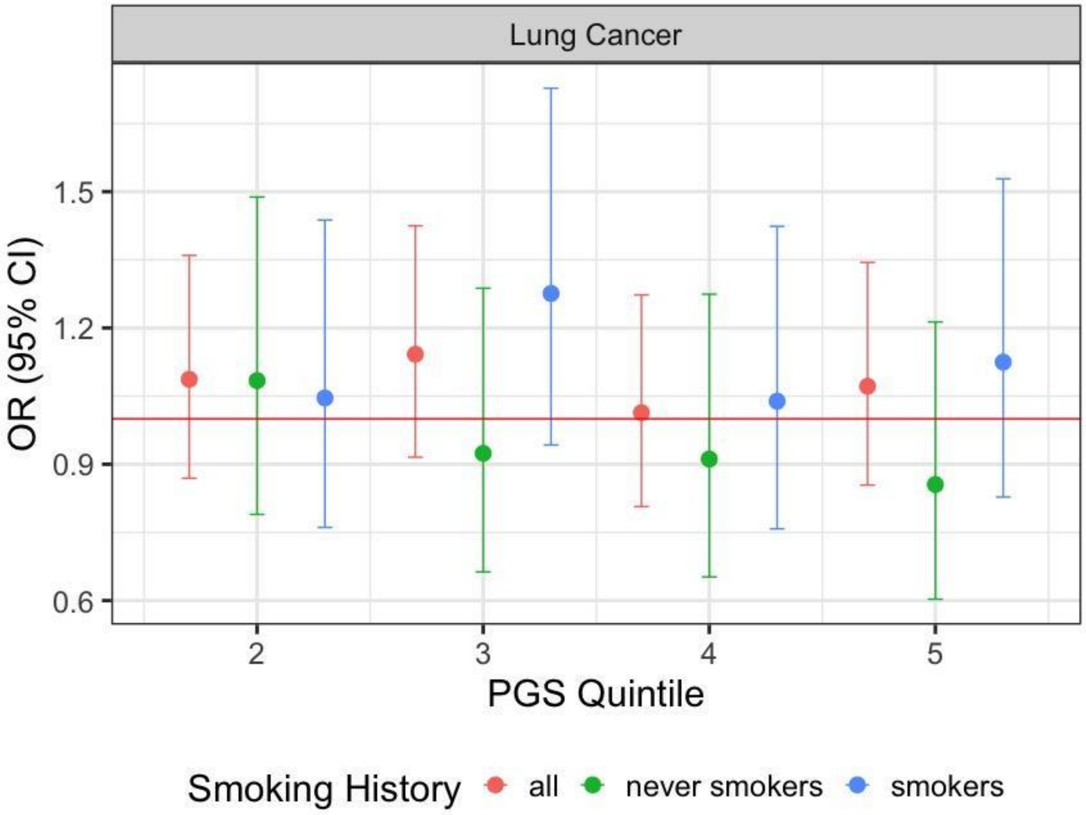
b) **DAG showing the relationship evaluated in the PGS-PheWAS never-smoker analysis**

Effects of germline variants that predispose to tobacco use are captured by the TUD-PGS. Systemic health effects are captured by 1847 phecodes. The star indicates the relationship evaluated in the PGS-PheWAS analysis. The red X in 1B denotes that the effect of tobacco use behavior on those systemic health effects is accounted for in the PGS-PheWAS analysis in never-smokers.



**Supplementary Figure 3.2: TUD-PGS association with Lung Cancer across PGS quintiles among ever-smokers and never smokers**

X-axis represents the top 4 quintiles grouped by TUD-PGS. Y axis represents effect sizes represented by odds ratios. Red line indicates OR =1.



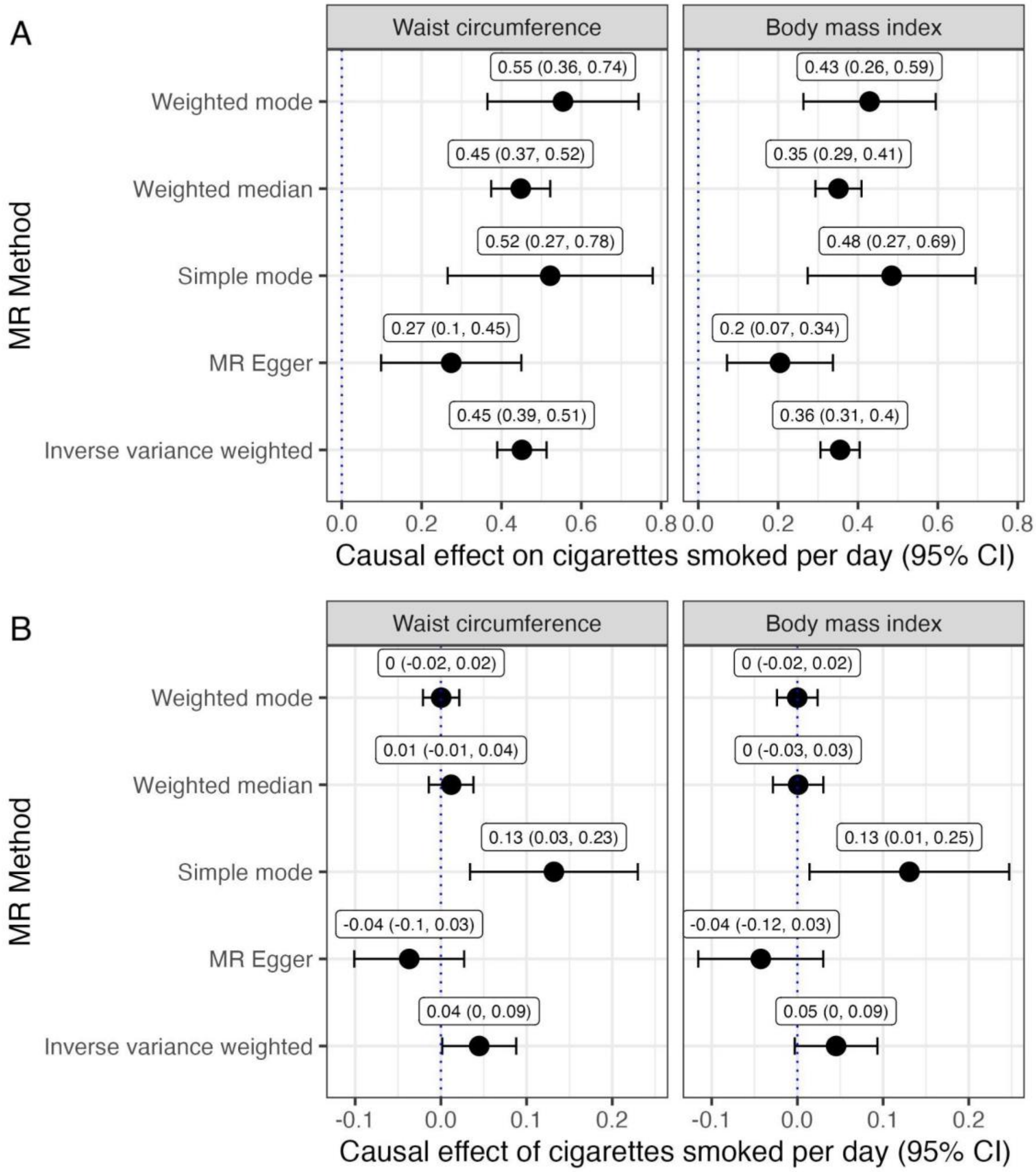
**Supplementary Figures 3.3A and 3.3B: Mendelian Randomization Between Waist Circumference, BMI, and Cigarettes Smoked Per Day**

a) **MR results between Outcome - cigarettes smoked per day and Exposure - waist circumference and body mass index**

Mendelian Randomization results across multiple MR methods using summary statistics for cigarettes smoked per day from the GSCAN Consortium and for waist circumference from MRC-UBristol and body mass index from UKBB. X axis represents the effect sizes and Y axis represents the MR method used.

b) **MR results between Exposure - cigarettes smoked per day and Outcome - waist circumference and body mass index**

Mendelian Randomization results across multiple MR methods using summary statistics for cigarettes smoked per day from the GSCAN Consortium and for waist circumference from MRC-UBristol and body mass index from UKBB. X axis represents the effect sizes and Y axis represents the MR method used.



## References

1. World Health Organization. WHO Report on the Global Tobacco Epidemic, 2017. Geneva: World Health Organization, 2017.

2. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Centers for Disease Control and Prevention (US); 2014. Accessed July 14, 2022. http://www.ncbi.nlm.nih.gov/books/NBK179276/

3. Caraballo RS, Rice KL, Neff LJ, Garrett BE. Social and Physical Environmental Characteristics Associated With Adult Current Cigarette Smoking. *Prev Chronic Dis*. 2019;16:180373. doi:10.5888/pcd16.180373

4. Evans LM, Jang S, Hancock DB, et al. Genetic architecture of four smoking behaviors using partitioned SNP heritability. *Addict Abingdon Engl*. 2021;116(9):2498-2508. doi:10.1111/add.15450

5. Saunders GRB, Wang X, Chen F, et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*. 2022;612(7941):720-724. doi:10.1038/s41586-022-05477-4

6. Kaprio J. Genetic epidemiology of smoking behavior and nicotine dependence. *COPD*. 2009;6(4):304-306. doi:10.1080/15412550903049165

7. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med*. 2020;12(1):44. doi:10.1186/s13073-020-00742-5

8. Ohi K, Nishizawa D, Muto Y, et al. Polygenic risk scores for late smoking initiation associated with the risk of schizophrenia. *Npj Schizophr*. 2020;6(1):1-7. doi:10.1038/s41537-020-00126-z

9. Al-Soufi L, Martorell L, Moltó MD, et al. A polygenic approach to the association between smoking and schizophrenia. Addict Biol. 2022;27(1):e13104. doi:10.1111/adb.13104

10. Deak JD, Clark DA, Liu M, et al. Alcohol and nicotine polygenic scores are associated with the development of alcohol and nicotine use problems from adolescence to young adulthood. Addiction. 2022;117(4):1117-1127. doi:10.1111/add.15697

11. Cooke ME, Clifford JS, Do EK, et al. Polygenic score for cigarette smoking is associated with ever electronic-cigarette use in a college-aged sample. Addiction. 2022;117(4):1071-1078. doi:10.1111/add.15716

12. Bray M, Chang Y, Baker TB, et al. The Promise of Polygenic Risk Prediction in Smoking Cessation: Evidence From Two Treatment Trials. *Nicotine Tob Res Off J Soc Res Nicotine Tob*. 2022;24(10):1573-1580. doi:10.1093/ntr/ntac043

13. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205-1210. doi:10.1093/bioinformatics/btq126

14. Pendergrass SA, Brown-Gentry K, Dudek S, et al. Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLOS Genet*. 2013;9(1):e1003087. doi:10.1371/journal.pgen.1003087

15. Privé F, Aschard H, Carmi S, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort

[published correction appears in Am J Hum Genet. 2022 Feb 3;109(2):373]. Am J Hum Genet. 2022;109(1):12-23. doi:10.1016/j.ajhg.2021.11.008

16. Chang TS, Ding Y, Freund MK, et al. Pre-existing conditions in Hispanics/Latinxs that are COVID-19 risk factors. *iScience*. 2021;24(3). doi:10.1016/j.isci.2021.102188
17. Lajonchere C, Naeim A, Dry S, Wenger N, Elashoff D, Vangala S, Petruse A, Ariannejad M, Magyar C, Johansen L, Werre G, Kroloff M, Geschwind D, An Integrated, Scalable, Electronic Video Consent Process to Power Precision Health Research: Large, Population-Based, Cohort Implementation and Scalability Study. J Med Internet Res 2021;23(12):e31121; doi: 10.2196/31121 : https://www.jmir.org/2021/12/e31121
18. Johnson R, Ding Y, Venkateswaran V, et al. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. *Genome Med*. 2022;14(1):1-23. doi:10.1186/s13073-022-01106-x
19. Ruth Johnson, Yi Ding, Arjun Bhattacharya, Sergey Knyazev, Alec Chiu, Clara Lajonchere, Daniel H. Geschwind, Bogdan Pasaniuc, The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank, Cell Genomics, Volume 3, Issue 1, 2023, 100243, ISSN 2666-979X, https://doi.org/10.1016/j.xgen.2022.100243.
20. Naeim A, Dry S, Elashoff D, et al. Electronic Video Consent to Power Precision Health Research: A Pilot Cohort Study [published correction appears in JMIR Form Res. 2021 Oct 21;5(10):e33891]. JMIR Form Res. 2021;5(9):e29123. Published 2021 Sep 8. doi:10.2196/29123

21. Sanderson E, Glymour MM, Holmes MV, et al. Mendelian randomization. *Nat Rev Methods Primer*. 2022;2(1):1-21. doi:10.1038/s43586-021-00092-5

22. Infinium Global Screening Array-24 Kit | Population-scale genetics. Accessed January 31, 2023. https://www.illumina.com/products/by-type/microarray-kits/infinium-global-screening.html

23. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021;590(7845):290-299. doi:10.1038/s41586-021-03205-y

24. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7

25. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4(1). doi:10.1186/s13742-015-0047-8

26. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. 2017;33(17):2776–8.

27. Data | 1000 Genomes. Accessed January 31, 2023. https://www.internationalgenome.org/data

28. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

29. Samuel A. Lambert, Laurent Gil, Simon Jupp, Scott C. Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, John Danesh, Jacqueline A. L. MacArthur and Michael Inouye. The Polygenic Score

Catalog as an open database for reproducibility and systematic evaluation Nature Genetics doi: 10.1038/s41588-021-00783-5 (2021).

30. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger [published online ahead of print, 2020 Dec 16]. Bioinformatics. 2020;36(22-23):5424-5431. doi:10.1093/bioinformatics/btaa1029

31. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1111. doi:10.1038/nbt.2749

32. The Python Language Reference. Python documentation. Accessed January 31, 2023. https://docs.python.org/3/reference/index.html

33. The Comprehensive R Archive Network. Accessed January 31, 2023. https://cran.r-project.org/

34. Services I of M (US) C on MA to PHC, Millman M. *A Model for Monitoring Access*. National Academies Press (US); 1993. Accessed January 31, 2023. https://www.ncbi.nlm.nih.gov/books/NBK235891/

35. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *J Stat Softw*. 2010;36:1-48. doi:10.18637/jss.v036.i03

36. Liu M, Jiang Y, Wedow R, 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. 2019;51(2):237-244. doi:10.1038/s41588-018-0307-5

37. Elsworth B, Lyon M, Alexander T, et al. The MRC IEU OpenGWAS data infrastructure. *bioRxiv*. Published online August 10, 2020. doi:10.1101/2020.08.10.244293
38. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015;518(7538):197-206. doi:10.1038/nature14177
39. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife. doi:10.7554/eLife.34408
40. Ding Y, Hou K, Burch KS, et al. Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. Nat Genet. 2022;54(1):30-39. doi:10.1038/s41588-021-00961-5
41. Prevention (US) C for DC and, Promotion (US) NC for CDP and H, Health (US) O on S and. *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease*. Centers for Disease Control and Prevention (US); 2010. Accessed January 31, 2023. https://www.ncbi.nlm.nih.gov/books/NBK53017/
42. Roy A, Rawal I, Jabbour S, et al. Tobacco and Cardiovascular Disease: A Summary of Evidence. In: Prabhakaran D, Anand S, Gaziano TA, et al., editors. Cardiovascular, Respiratory, and Related Disorders. 3rd edition. Washington (DC): The International Bank for Reconstruction and Development / The World Bank; 2017 Nov 17. Chapter 4. Available from: https://www.ncbi.nlm.nih.gov/books/NBK525170/ doi: 10.1596/978-1-4648-0518-9_ch4
43. De Angelis F, Wendt FR, Pathak GA, et al. Drinking and smoking polygenic risk is associated with childhood and early-adulthood psychiatric and behavioral traits

independently of substance use and psychiatric genetic risk. *Transl Psychiatry*. 2021;11(1):1-12. doi:10.1038/s41398-021-01713-z

44. Carreras-Torres R, Johansson M, Haycock PC, et al. Role of obesity in smoking behaviour: Mendelian randomisation study in UK Biobank. *BMJ*. 2018;361. doi:10.1136/bmj.k1767

45. Thorgeirsson TE, Gudbjartsson DF, Sulem P, et al. A common biological basis of obesity and nicotine addiction. *Transl Psychiatry*. 2013;3(10):e308. doi:10.1038/tp.2013.81

# Chapter 4

# EHR-Data And Polygenic Scores Reveal The Interplay Of Serum Bilirubin, Smoking, And Cancer

## 4.1 Abstract

Bilirubin is a potent antioxidant with a protective role in many diseases. We examined the interplay of serum bilirubin (SB) levels, tobacco smoking (a known cause of low SB), head and neck cancer (HNC), and lung cancer (LC).

Using data from 393,252 participants from UCLA Health we used regression models, propensity score matching, and polygenic scores to examine bidirectional causal effects between smoking, HNC and LC risk, and SB levels.

Current tobacco smokers showed lower SB (-0.038mg/dL, 95% CI: [-0.043, -0.032]), compared to never-smokers. HNC and LC cases demonstrated lower SB levels (-0.11mg/dL, [-0.13, -0.09] and -0.085mg/dL, CI [-0.10, -0.07] respectively) compared to cancer-free controls. This effect persisted after adjusting for smoking. SB levels are associated with HNC and LC risk (ORs: 0.27, CI [0.20,0.37] and 0.57, CI [0.43,0.75], respectively). Lastly, a polygenic score for SB is associated with LC but not with HNC (OR: 0.78, CI [0.65,0.94] and 1.01, [0.86,1.19], respectively).

Low SB levels are associated with a risk of HNC and LC independent of the effect of tobacco smoking. Observed low SB is associated with LC and HNC, while genetically predicted low SB (from polygenic scores) is associated with LC only. These findings suggest that, with further validation, SB could serve as a potential early biomarker for LC and HNC.

## 4.2 Introduction

Bilirubin is a compound found in the blood, derived from the breakdown of heme, a component of hemoglobin found in red blood cells. At the end of their life cycle, red blood cells are broken down in the spleen and bone marrow, forming bilirubin, which is then transported to the liver and excreted in bile. Thus, the level of serum bilirubin (SB) is an indicator of liver health, and high SB levels are clinically used markers of liver disease[1]. However, low levels of SB have not been categorized broadly in clinical settings and do not factor into diagnosis or treatment planning in most diseases. Recent studies, however, have redefined the role of SB in the human body, recognizing it as a metabolic hormone with potent antioxidant effects[2]. Low SB levels are associated with increased risk of diseases including cancers, and metabolic and cardiovascular diseases[2,3,4]. Low SB levels have also been associated with poorer survival and worse prognosis in lung and oral cancers[5,6].

A commonly reported cause of low SB is tobacco smoking, likely secondary to the oxidative stress placed on the body under tobacco smoking conditions, disrupting normal bilirubin metabolism[7]. Tobacco smoking is also a well-established risk factor for cancers, especially lung and head and neck cancers (HNC). Thus, it can be theorized that low SB levels linked to cancers could be secondary to interaction with tobacco smoking. Studies have reported interactions between low SB levels and tobacco use in the risk of LC[8,9]. However, this relationship is not well-studied in HNC.

Here, we examined the interplay of SB with tobacco smoking, HNC, and LC using lab values and electronic health records (EHR) information from a hospital-based biobank in Los Angeles. We examined differences in SB levels across sex, self-reported ethnicity/race (SIRE), and cancer case and control status. Next, we examined the effects of smoking on SB and of HNC and LC on SB. We assessed the validity of using SB to predict HNC and LC by using observed SB levels from the patient's EHR and evaluating the combined effect of tobacco smoking and SB on both cancers. Lastly, we used a polygenic score for SB to evaluate the predictive ability of genetically predicted SB to predict HNC and LC.

## 4.3 Methods

### 4.3.1 Study Population: The UCLA ATLAS biobank

The study population is derived from UCLA Health which is a de-identified patient electronic health records (EHR) linked biobank with over 2 million patients. The biobank includes information on demographics, diagnostic codes, laboratory values, prescriptions data, and procedure codes. The UCLA ATLAS Community Health Initiative also includes > 60,000 genotyped participants with their de-identified EHR linked through the DDR. A more extensive description of this resource including information on recruitment, consent, sample processing, and quality control pipelines can be found in previous publications[10-14].

In this study, we included 393,252 participants from the DDR aged 18 and above, of which 15,023 participants were of European American genetically inferred ancestry group with SB measurements and genotypes available for polygenic score construction and analyses. All participants consented to participate in the research.

### 4.3.2 Bilirubin Biomarker Measurements

The primary biomarker analyzed in this study is serum total bilirubin (SB), measured in mg/dL. The SB measurements were extracted from the patient's electronic health records. Since most patients have multiple blood tests over their encounter history with multiple results for serum total bilirubin laboratory tests, we extracted the maximum and minimum total bilirubin values in each patient's record and computed their mean to get a mid-range total bilirubin value for each individual. For quality control and to filter out errors in data entry and significant outliers, we excluded SB values that were less than the 0.1th percentile or greater than the 95th percentile. This step also served to exclude patients with consistently high bilirubin values secondary to liver disease or medication side effects.

### 4.3.3 HNC, LC, and Tobacco Smoking Ascertainment

Smoking history was extracted from the patient's self-reported demographic information. We binarized smoking history to ever and never smokers.

- Ever-smokers included participants who selected any of the following options in their health history: 'Former', 'Heavy Smoker', 'Light Smoker', 'Smoker, Current Status Unknown', and 'Some Days'.
- Never smokers included participants who selected the following options: 'Never', 'Passive Smoke Exposure - Never Smoker'.

We excluded participants who selected 'Never Assessed' or whose smoking history was 'Unknown'. Further, we created a 'Current Smoker' label for participants who answered yes to 'Every Day', 'Some Days', 'Light Smoker', or 'Heavy Smoker' in their smoking history question. This variable was used for additional analysis of tobacco smoking and SB.

Head and neck cancer (HNC) and LC status were ascertained by extracting ICD codes from the participant's EHR. A detailed list of the ICD codes used to ascertain HNC and LC status is included in the supplementary section. Participants were designated as a ‘Case’ if the ICD codes for head and neck cancer or LC existed in their EHR. Participants were assigned to ‘Controls’ when their EHR did not contain any ICD codes for malignancy or cancers.

### 4.3.4 Polygenic score selection and imputation

A publicly available polygenic score for ‘Total Bilirubin’ was used for polygenic score computation ( PGS002160 from the PGS Catalog[15,16]) trained using 391,124 European individuals from the UK biobank and includes 120,068 SNPs. PGS training analyses were adjusted for sex, age, birth date, Townsend’s deprivation index, and the first 16 principal components of the genotype matrix to account for population stratification. We computed the PGS for each genotyped UCLA ATLAS participant by multiplying the individual risk allele dosages by their corresponding weights provided by the PGS catalog using the ‘pgsc_calc’ workflow[16]. The PGS was mean-centered and standardized by the standard deviation within the European American genetic ancestry group to generate a PGS Z-score which was used in further analysis (n = 15,023). We restricted the PGS analysis to European American ancestry group since the original PGS was trained in Europeans, and studies have shown that the predictive performance of PGS is unreliable when used in ancestries that are genetically dissimilar to the trained population[17].

### 4.3.5 Statistical Analyses

All analysis was conducted in either Python 2.6.8[18] or R 4.2.1[19]. We used linear regression models to evaluate the associations between the dependent variable of SB levels and independent variables of tobacco smoking, HNC, and LC. We adjusted for participant age, sex, and self-reported race/ethnicity (SIRE) in all models. In a subsequent model to evaluate cancer effect sizes on SB, we additionally account for tobacco smoking by including it as an independent variable. Linear coefficients and confidence intervals were calculated, with P-values from Wald-type test statistics.

Propensity scores were estimated by logistic regression analysis, with cancer status as the dependent variable and age, sex, SIRE, and tobacco smoking as independent variables. We used 1:1 propensity matching to create a more balanced analysis group, creating two groups - 2,037 HNC cases and 2,037 HNC controls, and 2,373 LC cases and 2,373 LC controls with no significant differences in age, sex, SIRE or tobacco smoking.

To evaluate the effect of observed SB and genetically predicted SB on HNC and LC, we used logistic regression models with HNC/LC as the dependent variable and SB as the independent variable. The SB variable was the extracted SB levels from the EHR for the observed SB analysis and the polygenic score for SB for the genetically predicted SB analysis. Odds ratios and confidence intervals were calculated, with P-values from Wald-type test statistics for these analyses.

### 4.3.6 Ethical Approval

Patient Recruitment and Sample Collection for Precision Health Activities at UCLA is an approved study by the UCLA Institutional Review Board (UCLA IRB #17-001013). All participants provided informed consent to participate in the research.

### 4.3.7 Data Sharing

All shareable data produced in the present work are contained in the manuscript.

## **4.4 Results**

### 4.4.1 Participant characteristics and effect of demographic factors on SB

The baseline characteristics of 393,252 participants with 2,039 HNC cases, 2,378 LC cases, and 388,835 cancer controls are shown in ***Supp Table 4.1***. We used a multivariable linear regression model to systematically evaluate the differences in SB level distributions among demographic factors. Males on average had 0.15 mg/dL higher SB compared to females (CI [0.146, 0.151]), and older individuals had a 0.0009 mg/dL increase per year of age, CI [0.0009,0.001]). When compared to self-reported White or Caucasian group, Asians had on average 0.019 mg/dL higher SB values, (CI [0.015, 0.023]) and African Americans had 0.01 mg/dL lower SB values (CI [-0.016, -0.005]). Lastly, individuals who self-identified as 'Hispanic/Latin American' ethnicity had 0.024 mg/dL higher SB compared to individuals who self-identified as 'Not Hispanic/Latin American' CI [0.020, 0.028] (***Table 4.1***). These results suggest that SB levels vary significantly by participant sex, age and SIRE in this study population.

### 4.4.2 Lower SB levels are associated with smokers compared to never smokers

Next, we evaluated whether tobacco smoking is associated with SB levels. For this, we compared the SB levels of ever-smokers to never-smokers and further, we compared the SB levels of current smokers to  never-smokers.

In a linear regression model adjusting for age, sex, and SIRE, ever-smokers demonstrated a negative association with SB (*-0.017 mg/dL, CI [-0.019, -0.014]*). This effect size persists and slightly increases when comparing 'current smokers' (*-0.038 mg/dL, CI [-0.043, -0.032]*) to 'never smokers'. Both of these results suggest that tobacco smoking is associated with low SB levels with current smokers demonstrating much lower SB levels when compared to ever-smokers ***(Table 4.2).***

### 4.4.3 Lower SB levels are associated with LC and HNC independent of smoking status

Next, we evaluated the associations between LC and HNC and SB levels. In a linear regression model, HNC and LC demonstrate negative associations with SB after adjusting for age, sex, SIRE (*-0.11 mg/dL, CI [-0.13, -0.1] and -0.09 mg/dL, CI [-0.1, -0.07], respectively*). The negative associations between HNC, LC, and SB persist after adjusting additionally for smoking history (*-0.11 mg/dL, CI [-0.13, -0.09] and -0.08 mg/dL, CI [-0.1, -0.07], respectively*). An interaction term between LC and smoking demonstrated a significant effect on SB (*-0.06 mg/dL, CI [-0.09, -0.02]* ) while an interaction term between HNC and smoking did not (*-0.01 mg/dL, CI [-0.04, 0.03]*) *(**Table 4.3**).*

We additionally explored this association using propensity matched groups. The baseline characteristics of the propensity-matched HNC and LC groups are shown in Supplementary Tables 2a and 2b. In two linear regression models, HNC and LC individually demonstrated negative associations with SB *( -0.11mg/dL, CI [-0.13, -0.09]) and -0.083mg/dL, CI [-0.10, -0.07] respectively)*. These results suggest that SB has an inverse relationship with LC and HNC which is not mediated solely by tobacco smoking. Additionally, the significant interaction between LC and smoking on SB suggests that the negative association between LC and SB is exacerbated with smoking, though a similar trend is not observed for HNC.

### 4.4.4 Observed SB Levels are associated with LC and HNC

Next, to obtain the effect sizes of SB on the risk of LC and HNC, we used logistic regression models with HNC and LC as the dependent variables. In this model, lower levels of SB predict LC and HNC after adjusting for age, sex, SIRE, and smoking. *(OR: 0.57, CI [0.43,0.75] and 0.27, CI [0.20, 0.37]) respectively)*. Additionally, interaction terms of SB and smoking demonstrated significant effects on LC but not on HNC *(OR: 0.5, CI [0.35, 0.72]) and 0.95, CI [0.62,1.46] respectively). (**Table 4.4)***

### 4.4.5 A polygenic score for SB (SB-PGS) is associated with LC

In a sub-group of genotyped European ancestry patients from the UCLA ATLAS biobank, we first validated a PGS for total serum bilirubin (SB-PGS). This SB-PGS demonstrated a strong association with observed SB levels extracted from the patient's EHR lab values (*0.094 mg/dL per standard deviation of SB-PGS, CI [0.091, 0.097*], demonstrating reliability in predicting SB.

We then evaluated the effect sizes of genetically predicted SB on LC and HNC. In this European ancestry sub-group, after adjusting for age, sex, and the first 5 principal components, SB-PGS predicts LC (*n=124 cases, OR: 0.78, CI [0.65,0.94]*) but not HNC, *(n = 152 cases; OR = 1.01, CI [0.86,1.19])*. When restricted to only ever-smokers in this subgroup, SB-PGS demonstrates a slightly higher effect size (*OR: 0.70, CI [0.56,0.88*]) on LC and no effect with HNC *(OR:0.95, CI [0.74,1.20]*).

## 4.5 Discussion

In this study, we analyzed the associations of SB with the risks of HNC and LC using the UCLA ATLAS EHR-linked biobank. SB was inversely associated with HNC and LC, even after adjusting for tobacco smoking. Tobacco smoking significantly interacted with SB on LC risk but not HNC. Lastly, observed SB predicts both HNC and LC while genetically predicted SB predicted LC and not HNC.

This study is the first large-scale study to report on the negative associations between SB and HNC. These results parallel the negative association between SB and LC noted in our study and other studies of SB and LC [8,9,20]. One potential mechanism for this observed association is likely through SB's action on reactive oxygen species and inflammatory factors, which are known to promote the etiopathogenesis of cancers[21,22]. The protective effects of SB on these two cancers are likely secondary to its antioxidant function in the body and endogenous anti-inflammatory activity[23].

The associations between SB and cancers could be mediated by intrinsic or extrinsic factors that have an effect on one or both of these factors. Among extrinsic factors, tobacco smoking is a well-known contributor of free radicals and reactive oxygen species, playing an important role in carcinogenesis[24]. Studies including this one find evidence of strong negative associations between tobacco smoking and

SB[7,8,9], necessitating further inquiry into the role of tobacco smoking in the SB-cancer pathway. The persistent associations of the SB-HNC/LC associations after adjusting for tobacco smoking suggest that these associations are not solely driven by tobacco smoking. In fact, tobacco smoking interacts with SB in LC, a finding we did not observe in HNC. The difference in the SB-tobacco smoking interaction results between HNC and LC could likely be explained by differences in etiology (alcohol consumption and human papillomavirus infection for HNC), differential tissue response, local environment, and microbiome differences between the two sites. From a clinical perspective, these results suggest that SB could be a low-cost laboratory biomarker for HNC and LC. However, further studies are needed to understand these complex pathway mechanisms and to validate these findings.

Next, an intrinsic factor that might play a role in the SB-cancer association is the genetic control of SB. A validated polygenic score for SB predicted LC but not HNC in this study. Additionally, a larger effect size is noted in ever-smokers for LC, i.e. ever-smokers with genetically predicted high SB levels had a lower risk of LC when compared to ever-smokers with lower SB levels. A recent study used Mendelian Randomization methods to evaluate the association between two genetic variants which account for ~40% of population-level SB variability and LC and found evidence of a causal association[25]. These findings suggest the possible influence of genetic control of SB on the susceptibility to LC. The inability of the SB-PGS to predict HNC could be attributed to the multifactorial nature of HNC development. It is also likely that the genetic effects of the variants in the SB-PGS have very small effects on HNC that we could not capture. Further validation studies are required with adequate sample sizes across the subtypes of HNC and including information on a larger spectrum of environmental risk factors, including alcohol consumption and human papillomavirus status. Multiple stages of research must be completed before clinical translation is possible and equitable — such as ensuring the robust predictive performance of SB-PGS across diverse ancestral groups.

The main strengths of our study are that we use data from an EHR-linked biobank, thus including patients from a real-world hospital system of diverse racial and ethnic groups. The large sample size adequately powers the analysis of SB, HNC, and LC. Given the significant role that tobacco smoking plays in the development of HNC and LC, we thoroughly evaluated the effects of tobacco smoking on SB levels and potential interactions with SB levels in the context of the cancer risk. Lastly, we used a polygenic score to explore the role that genetics may play in SB levels and cancer risk, to identify potential genetic biomarkers that could be used to predict cancer.

We end this discussion with some limitations of our study that must be kept in mind when interpreting and applying the results of this study. The observational nature of our study prevents us from making causal interpretations and assigning directions to the associations observed between SB levels, tobacco smoking, HNC, and LC. Next, the de-identified nature of an EHR-linked biobank prevents us from including socioeconomic variables that could have effects on tobacco smoking and cancer risk. Lastly, we lacked information on cancer subtype, staging, and grading and could not investigate these factors that could throw further light on the mechanisms behind SB and cancer associations.

In conclusion, our study finds associations between low SB levels and HNC, LC risk, and tobacco smoking. Observed SB predicts both HNC and LC while genetically predicted SB predicts LC. Further research is needed to validate SB as a potential laboratory and genetic biomarker for HNC and LC.

## 4.6 Tables

**Table 4.1: Effects of demographic factors on SB**

**Linear Regressions adjusted for age, sex, SIRE. Significance threshold is 0.05/8**

| Reference | Variables | Effect Size (mg/dL) | 95% CI | P-Value |
|---|---|---|---|---|
| Female | Male | 0.15 | 0.15,0.15 | < 0.0001 |
|  | Patient Age (per year) | 0.001 | 0.0009,0.001 | < 0.0001 |
| White or Caucasian | American Indian or Alaska Native | -0.02 | -0.03,-0.002 | 0.03 |
| White or Caucasian | Asian | 0.02 | 0.015,0.02 | < 0.0001 |
| White or Caucasian | Black or African American | -0.01 | -0.016,-0.004 | 0.0004 |
| White or Caucasian | Native Hawaiian or Other Pacific Islander | 0.03 | -0.004,0.06 | 0.09 |
| White or Caucasian | Other Race | 0.004 | -0.0001,0.009 | 0.06 |
| Not Hispanic or Latin ethnicity | Hispanic or Latin ethnicity | 0.02 | 0.019,0.03 | < 0.0001 |

**Table 4.2: Effect of tobacco Smoking on SB, comparing ever-smokers and current smokers to never-smokers**

| Reference | Variables | Effect on SB mg/dL | 95% CI | P-Value |
|---|---|---|---|---|
| Never smokers | Ever-smokers | -0.017 | -0.019, -0.014 | < 0.0001 |
| Never smokers | Current smokers | -0.038 | -0.043, -0.032 | < 0.0001 |

**Table 4.3: Effects of HNC and LC on SB, after adjusting for tobacco smoking and HNC/LC: smoking interaction**

| Variables | Effect on SB mg/dL | 95% CI | P-Value |
|---|---|---|---|
| HNC | -0.11 | -0.13, -0.10 | < 0.0001 |
| HNC (additionally adjusted for smoking) | -0.11 | -0.13, -0.09 | < 0.0001 |
| HNC: Smoking interaction | -0.01 | -0.04, 0.03 | 0.8 |
| LC | -0.09 | -0.1, -0.07 | < 0.0001 |
| LC (additionally adjusted for smoking) | -0.08 | -0.1, -0.07 | < 0.0001 |
| LC: Smoking interaction | -0.06 | -0.09, -0.02 | 0.0007 |

Table 4.4: Effects of SB levels on HNC and LC

| | OR | 95% CI | P-Value |
|---|---|---|---|
| **Outcome - HNC** | | | |
| Serum Bilirubin | 0.27 | 0.20, 0.37 | < 0.0001 |
| Bilirubin and smoking interaction variable | 0.95 | 0.62, 1.46 | 0.8 |
| **Outcome - LC** | | | |
| Serum Bilirubin | 0.57 | 0.43, 0.75 | 0.0001 |
| Bilirubin and smoking interaction variable | 0.50 | 0.35, 0.72 | 0.0002 |

## 4.7 Supplementary Tables

### ICD Codes for HNC and LC

HNC ICD Codes: '140', '140.1', '140.3', '140.4', '140.5', '140.6', '140.8', '140.9', '141', '141.1', '141.2', '141.3', '141.4', '141.5', '141.6', '141.8', '141.9', '142', '142.1', '142.2', '142.8', '142.9','143', '143.1', '143.8', '143.9', '144', '144.1', '144.8', '144.9', '145', '145.1', '145.2', '145.3', '145.4', '145.5', '145.6', '145.8', '145.9', '146', '146.1', '146.2', '146.3', '146.4', '146.5', '146.6', '146.7', '146.8', '146.9', '147', '147.1', '147.2', '147.3', '147.8', '147.9’, '148', '148.1', '148.2', '148.3', '148.8', '148.9', '149', '149.1', '149.8', '149.9', '160', '160.1', '160.2', '160.3', '160.4', '160.5', '160.8', '160.9’, '161', '161.1', '161.2', '161.3', '161.8', '161.9', 'C00.0', 'C00.1', 'C00.2', 'C00.3', 'C00.4', 'C00.5', 'C00.6', 'C00.8', 'C00.9', 'C02.0', 'C02.1', 'C02.2', 'C02.3', 'C02.4', 'C02.8', 'C02.9', 'C06.0', 'C06.1', 'C06.2', 'C06.8', 'C06.80', 'C06.89’, 'C06.9', 'C08.0', 'C08.1', 'C08.9', 'C09.0', 'C09.1', 'C09.8', 'C09.9', 'C10.0', 'C10.1', 'C10.2', 'C10.3', 'C10.4', 'C10.8', 'C10.9', 'C11.0', 'C11.1',’'C11.2', 'C11.3', 'C11.8', 'C11.9', 'C13.0', 'C13.1', 'C13.2', 'C13.8', 'C13.9', 'C14.0','C14.2', 'C14.8', 'C30.0'.

LC ICD Codes: 'C34.0', 'C34.1', 'C34.2', 'C34.3', 'C34.8', 'C34.9','162.0', '162.2', '162.3', '162.4', '162.5', '162.8', '162.9'

**Supplementary Table 4.1: Baseline characteristics of participants, stratified by cancer status**

| | | Cancer Controls | LC Cases | HNC Cases |
|---|---|---|---|---|

|  |  |  |  |  |
|---|---|---|---|---|
| n |  | 388835 | 2378 | 2039 |
| Patient Age, mean (SD) |  | 57.0 (18.7) | 75.8 (10.9) | 69.7 (13.5) |
| SB mg/dL, mean (SD) |  | 0.6 (0.3) | 0.5 (0.2) | 0.5 (0.2) |
| Sex, n (%) | Female | 213694 (55.0) | 1259 (52.9) | 589 (28.9) |
|  | Male | 175141 (45.0) | 1119 (47.1) | 1450 (71.1) |
| Smoking History, n (%) | Never Smokers | 270645 (69.6) | 831 (34.9) | 1010 (49.5) |
|  | Ever Smokers | 118190 (30.4) | 1547 (65.1) | 1029 (50.5) |
| Self-reported Race, n (%) | American Indian or Alaska Native | 2048 (0.7) | 10 (0.5) | 6 (0.4) |
|  | Asian | 36563 (13.4) | 395 (19.6) | 208 (12.8) |
|  | Black or African American | 15834 (5.8) | 111 (5.5) | 59 (3.6) |
|  | Native Hawaiian or Other Pacific Islander | 527 (0.2) | 9 (0.4) | 2 (0.1) |
|  | Other Race | 32190 (11.8) | 203 (10.1) | 171 (10.5) |
|  | White or Caucasian | 161651 (59.1) | 1111 (55.2) | 1078 (66.2) |

|  |  |  |  |  |
|---|---|---|---|---|
| Self-Reported Ethnicity, n (%) | Hispanic or Latin | 50630 (14.3) | 156 (7.0) | 170 (9.0) |
|  | Not Hispanic or Latin | 290389 (82.2) | 1961 (88.3) | 1655 (87.9) |

**Supplementary Table 4.2a: Propensity matched group baseline characteristics stratified by HNC**

|  |  | Overall | HNC Controls | HNC Cases |
|---|---|---|---|---|
| n |  | 4074 | 2037 | 2037 |
| Patient Age, mean (SD) |  | 69.7 (13.5) | 69.7 (13.5) | 69.7 (13.5) |
| Serum Total Bilirubin, [Q1,Q3] |  | 0.5 [0.4,0.7] | 0.6 [0.4,0.8] | 0.5 [0.4,0.6] |
| Sex, n (%) | Female | 1174 (28.8) | 587 (28.8) | 587 (28.8) |
|  | Male | 2900 (71.2) | 1450 (71.2) | 1450 (71.2) |
| Smoking History, n (%) | 0 | 2018 (49.5) | 1009 (49.5) | 1009 (49.5) |
|  | 1 | 2056 (50.5) | 1028 (50.5) | 1028 (50.5) |

| | | | | |
|---|---|---|---|---|
| Self-reported Race, n (%) | American Indian or Alaska Native | 11 (0.3) | 6 (0.4) | 5 (0.3) |
| | Asian | 414 (12.7) | 207 (12.7) | 207 (12.7) |
| | Black or African American | 118 (3.6) | 59 (3.6) | 59 (3.6) |
| | Native Hawaiian or Other Pacific Islander | 4 (0.1) | 2 (0.1) | 2 (0.1) |
| | Other Race | 342 (10.5) | 171 (10.5) | 171 (10.5) |
| | Unknown Race | 210 (6.5) | 105 (6.4) | 105 (6.5) |
| | White or Caucasian | 2156 (66.2) | 1078 (66.2) | 1078 (66.3) |
| Self-Reported Ethnicity, n (%) | Hispanic or Latin | 343 (9.1) | 173 (9.2) | 170 (9.0) |
| | Not Hispanic or Latin | 3310 (87.9) | 1655 (87.8) | 1655 (88.0) |
| | Unknown Ethnicity | 112 (3.0) | 56 (3.0) | 56 (3.0) |

**Supplementary Table 4.2b: Propensity matched group baseline characteristics stratified by LC**

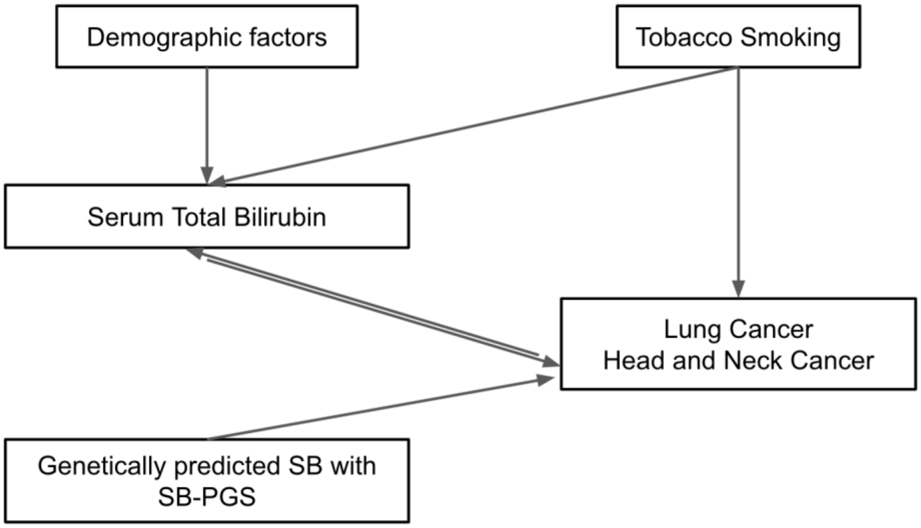
| | | Overall | LC Controls | LC Cases |
|---|---|---|---|---|
| n | | 4746 | 2373 | 2373 |

| | | | | |
|---|---|---|---|---|
| Patient Age, mean (SD) | | 75.8 (10.9) | 75.8 (10.9) | 75.8 (10.9) |
| Serum Total Bilirubin, [Q1,Q3] | | 0.5 [0.4,0.7] | 0.5 [0.4,0.7] | 0.5 [0.4,0.7] |
| Sex, n (%) | Female | 2512 (52.9) | 1256 (52.9) | 1256 (52.9) |
| | Male | 2234 (47.1) | 1117 (47.1) | 1117 (47.1) |
| Smoking History, n (%) | Never Smokers | 1656 (34.9) | 828 (34.9) | 828 (34.9) |
| | Ever Smokers | 3090 (65.1) | 1545 (65.1) | 1545 (65.1) |
| Self-Reported Race, n (%) | American Indian or Alaska Native | 15 (0.4) | 5 (0.2) | 10 (0.5) |
| | Asian | 786 (19.6) | 393 (19.6) | 393 (19.6) |
| | Black or African American | 222 (5.5) | 111 (5.5) | 111 (5.5) |
| | Native Hawaiian or Other Pacific Islander | 12 (0.3) | 6 (0.3) | 6 (0.3) |
| | Other Race | 406 (10.1) | 203 (10.1) | 203 (10.1) |
| | Unknown Race | 348 (8.7) | 174 (8.7) | 174 (8.7) |

| | | | | |
|---|---|---|---|---|
| | White or Caucasian | 2222 (55.4) | 1111 (55.5) | 1111 (55.3) |
| Self-Reported Ethnicity, n (%) | Hispanic or Latin | 326 (7.3) | 171 (7.7) | 155 (7.0) |
| | Not Hispanic or Latin | 3918 (88.1) | 1959 (87.8) | 1959 (88.4) |
| | Unknown Ethnicity | 202 (4.5) | 101 (4.5) | 101 (4.6) |

## 4.8 Supplementary Figures

**Supplementary Figure 4.1: Directed acyclic graph representing the associations of interest. Relevant result sections are indicated in the DAG**



## References

1. Kwo PY, Cohen SM, Lim JK. ACG Clinical Guideline: Evaluation of Abnormal Liver Chemistries. Am J Gastroenterol. 2017;112(1):18-35. doi:10.1038/ajg.2016.517
2. Creeden JF, Gordon DM, Stec DE, Hinds TD Jr. Bilirubin as a metabolic hormone: the physiological relevance of low levels. Am J Physiol Endocrinol Metab. 2021;320(2):E191-E207. doi:10.1152/ajpendo.00405.2020
3. Monroy-Iglesias MJ, Moss C, Beckmann K, et al. Serum Total Bilirubin and Risk of Cancer: A Swedish Cohort Study and Meta-Analysis. Cancers (Basel). 2021;13(21):5540. Published 2021 Nov 4. doi:10.3390/cancers13215540
4. Wang X, Wu D, Zhong P. Serum bilirubin and ischaemic stroke: a review of literature. Stroke Vasc Neurol. 2020;5(2):198-204. doi:10.1136/svn-2019-000289
5. Kunutsor SK, Bakker SJ, Gansevoort RT, Chowdhury R, Dullaart RP. Circulating total bilirubin and risk of incident cardiovascular disease in the general population. Arterioscler Thromb Vasc Biol. 2015;35(3):716-724. doi:10.1161/ATVBAHA.114.304929
6. Song YJ, Gao XH, Hong YQ, Wang LX. Direct bilirubin levels are prognostic in non-small cell lung cancer. Oncotarget. 2017;9(1):892-900. Published 2017 Dec 12. doi:10.18632/oncotarget.23184
7. Zhang H, Li G, Zhu Z, et al. Serum bilirubin level predicts postoperative overall survival in oral squamous cell carcinoma. J Oral Pathol Med. 2018;47(4):382-387. doi:10.1111/jop.12693

8. Jo J, Kimm H, Yun JE, Lee KJ, Jee SH. Cigarette smoking and serum bilirubin subtypes in healthy Korean men: the Korea Medical Institute study. J Prev Med Public Health. 2012;45(2):105-112. doi:10.3961/jpmph.2012.45.2.105
9. Lim JE, Kimm H, Jee SH. Combined effects of smoking and bilirubin levels on the risk of lung cancer in Korea: the severance cohort study. PLoS One. 2014;9(8):e103972. Published 2014 Aug 6. doi:10.1371/journal.pone.0103972
10. Wen CP, Zhang F, Liang D, et al. The ability of bilirubin in identifying smokers with higher risk of lung cancer: a large cohort study in conjunction with global metabolomic profiling. Clin Cancer Res. 2015;21(1):193-200. doi:10.1158/1078-0432.CCR-14-0748
11. Chang TS, Ding Y, Freund MK, et al. Pre-existing conditions in Hispanics/Latinxs that are COVID-19 risk factors. iScience. 2021;24(3). doi:10.1016/j.isci.2021.102188
12. Lajonchere C, Naeim A, Dry S, Wenger N, Elashoff D, Vangala S, Petruse A, Ariannejad M, Magyar C, Johansen L, Werre G, Kroloff M, Geschwind D, An Integrated, Scalable, Electronic Video Consent Process to Power Precision Health Research: Large, Population-Based, Cohort Implementation and Scalability Study. J Med Internet Res 2021;23(12):e31121; doi: 10.2196/31121 : https://www.jmir.org/2021/12/e31121
13. Johnson R, Ding Y, Venkateswaran V, et al. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. Genome Med. 2022;14(1):1-23. doi:10.1186/s13073-022-01106-x

14. Ruth Johnson, Yi Ding, Arjun Bhattacharya, Sergey Knyazev, Alec Chiu, Clara Lajonchere, Daniel H. Geschwind, Bogdan Pasaniuc, The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank, Cell Genomics, Volume 3, Issue 1, 2023, 100243, ISSN 2666-979X, https://doi.org/10.1016/j.xgen.2022.100243.
15. Naeim A, Dry S, Elashoff D, et al. Electronic Video Consent to Power Precision Health Research: A Pilot Cohort Study [published correction appears in JMIR Form Res. 2021 Oct 21;5(10):e33891]. JMIR Form Res. 2021;5(9):e29123. Published 2021 Sep 8. doi:10.2196/29123
16. Samuel A. Lambert, Laurent Gil, Simon Jupp, Scott C. Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, John Danesh, Jacqueline A. L. MacArthur and Michael Inouye. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nature Genetics. doi: 10.1038/s41588-021-00783-5 (2021).
17. Privé F, Aschard H, Carmi S, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort [published correction appears in Am J Hum Genet. 2022 Feb 3;109(2):373]. Am J Hum Genet. 2022;109(1):12-23. doi:10.1016/j.ajhg.2021.11.008
18. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities [published correction appears in Nat Genet. 2021 May;53(5):763]. Nat Genet. 2019;51(4):584-591. doi:10.1038/s41588-019-0379-x

19. The Python Language Reference. Python documentation. Accessed January 31, 2023. https://docs.python.org/3/reference/index.html

20. The Comprehensive R Archive Network. Accessed January 31, 2023. https://cran.r-project.org/

21. Horsfall LJ, Rait G, Walters K, et al. Serum bilirubin and risk of respiratory disease and death. JAMA. 2011;305(7):691-697. doi:10.1001/jama.2011.124

22. Nakamura H, Takada K. Reactive oxygen species in cancer: Current findings and future directions. Cancer Sci. 2021;112(10):3945-3952. doi:10.1111/cas.15068

23. Greten FR, Grivennikov SI. Inflammation and Cancer: Triggers, Mechanisms, and Consequences. Immunity. 2019;51(1):27-41. doi:10.1016/j.immuni.2019.06.025

24. Vogel ME, Zucker SD. Bilirubin acts as an endogenous regulator of inflammation by disrupting adhesion molecule-mediated leukocyte migration. Inflamm Cell Signal. 2016;3(1):e1178. doi:10.14800/ics.1178

25. Valavanidis A, Vlachogianni T, Fiotakis K. Tobacco smoke: involvement of reactive oxygen species and stable free radicals in mechanisms of oxidative damage, carcinogenesis and synergistic effects with other respirable particles. Int J Environ Res Public Health. 2009;6(2):445-462. doi:10.3390/ijerph6020445

26. Horsfall LJ, Burgess S, Hall I, Nazareth I. Genetically raised serum bilirubin levels and lung cancer: a cohort study and Mendelian randomisation using UK Biobank. Thorax. 2020;75(11):955-964. doi:10.1136/thoraxjnl-2020-214756

# Chapter 5

## Conclusions and Future Directions

Bioinformatics and computational methods are vital additions to our research arsenal in the era of precision medicine and dentistry. In this thesis, we utilized novel methods including phenome-wide association tests, lab-wide association tests, and polygenic scores, to create computational pipelines streamlined for the UCLA biobank data. We then used these pipelines to study two phenotypes - tobacco use disorder and head and neck cancer. We evaluated the potential of a polygenic score (PGS) to predict tobacco use disorder in 4 genetic ancestry groups. We also used this PGS to evaluate the phenome-wide associations of tobacco use predisposing genetic variants. We found that PGS trained in European ancestry populations did not reliably predict or risk-stratify participants of non-European ancestry groups for tobacco use disorder. Next, we found that individuals with a genetic predisposition to tobacco use demonstrate associations with circulatory, psychiatric, metabolic, and respiratory phenotypes. Lastly, we found that when individuals with a genetic predisposition to tobacco use disorder did not engage in tobacco smoking, they were at risk for other disorders including obesity and substance addiction disorders.  With further validation in other biobanks and across diverse populations, these findings could have implications for the management of tobacco use disorders. Individuals with a genetic predisposition to tobacco use disorder might require more comprehensive management to manage underlying addictive tendencies.

Next, in the study examining serum bilirubin (SB) levels, tobacco use, head and neck (HNC), and lung cancer (LC), we found that SB demonstrates a bidirectional negative association with HNC and LC. This association persists after adjusting for tobacco smoking, suggesting that the pathway between SB and HNC/LC is mediated by factors other than tobacco smoking. Lastly, we also found that a polygenic score for SB is significantly associated with LC. While these results are promising, clinical translation requires more precise measures of SB levels with clear risk thresholds for different sex, race and ethnicity groups. With further extensive validation of our results, SB levels could potentially serve as a low-cost biomarker for HNC and LC.

In future studies, I aim to employ the computational methods described here, including genome-wide-, phenome-wide analysis, and polygenic scores to study head and neck cancers, specifically oropharyngeal cancers (OPC). Germline studies of OPC and head and neck cancer (HNC) patients indicate that, like other more widely studied cancers, several germline variants are associated with OPC[1-5]. Additionally, germline variants can interact with somatic mutations to influence the course of HNC[6]. However, there are some critical knowledge gaps:

1) Current OPC genome-wide association studies (GWAS) are not ancestrally diverse. GWAS are powerful genetic tools to identify germline variants associated with traits of interest such as OPC and HNC[1-5]. However, these studies are done primarily in European ancestry populations and the results are not generalizable to individuals of other ancestries[7,8]. In order to translate these research findings to the clinic, it is critical to increase the diversity of populations included in genetic studies.

2) The pleiotropic effects of OPC-associated germline variants are largely unknown. Risk variants identified by GWAS for cancers often show associations with other cancer and developmental phenotypes. For example, a CDH1 germline variant is

associated with breast cancer, gastric carcinoma, and cleft/lip and palate[9]. These pleiotropic associations provide key information on the biological mechanisms underlying the risk conferred by these variants.

To address these gaps, I plan to conduct ancestry-specific OPC GWAS and cross-ancestry meta-analyses within ATLAS using an analytic pipeline developed for the ATLAS biobank. I will then meta-analyze OPC ATLAS GWAS results with publicly available OPC GWAS summary statistics from patient-recruited cohorts and other biobank data, to create the largest and most diverse GWAS of OPC. Next, I plan to develop an OPC-PGS using the top variants from the results of the meta-analysis, excluding data from the ATLAS biobank to prevent sample overlap using methods optimized for multi-ancestry PGS development. I will thoroughly evaluate the PGS performance using performance metrics within all available genetic ancestry groups in UCLA ATLAS. I will utilize the phenome-wide association (PheWAS) analysis pipeline, integrated with the OPC-PGS to perform a PGS-PheWAS analysis to evaluate the pleiotropic effects of common variants across 1847 phenotypes. Lastly, with further IRB approval, I plan to evaluate the associations between the OPC-PGS and clinical characteristics such as human papillomavirus infection, tumor stage and grade, pathological findings and imaging presentations. Establishing clinical correlation with genetic factors will provide important therapeutic insights and will aid in the development of reliable genetic testing recommendations to advance precision medicine for OPC. Inclusion of diverse populations will help advance the goal of equitable clinical translation of research findings.

## References

1. Lesseur C, Diergaarde B, Olshan AF, et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. Nat Genet. 2016;48(12):1544-1550. doi:10.1038/ng.3685
2. Ferreiro-Iglesias A, McKay JD, Brenner N, et al. Germline determinants of humoral immune response to HPV-16 protect against oropharyngeal cancer. Nat Commun. 2021;12(1):5945. Published 2021 Oct 12. doi:10.1038/s41467-021-26151-9
3. Huang KL, Mashl RJ, Wu Y, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell. 2018;173(2):355-370.e14. doi:10.1016/j.cell.2018.03.039
4. Fostira F, Koutsodontis G, Vagia E, et al. Predisposing Germline Mutations in Young Patients With Squamous Cell Cancer of the Oral Cavity. JCO Precision Oncology. 2018 Nov;2:1-8. DOI: 10.1200/po.18.00022. PMID: 35135152.
5. Shete S, Liu H, Wang J, et al. A Genome-Wide Association Study Identifies Two Novel Susceptible Regions for Squamous Cell Carcinoma of the Head and Neck. Cancer Res. 2020;80(12):2451-2460. doi:10.1158/0008-5472.CAN-19-2360
6. Feng G, Feng H, Qi Y, et al. Interaction analysis between germline genetic variants and somatic mutations in head and neck cancer. Oral Oncol. 2022;128:105859. doi:10.1016/j.oraloncology.2022.105859
7. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019;570(7762):514-518. doi:10.1038/s41586-019-1310-4

8. Carlson CS, Matise TC, North KE, et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol. 2013;11(9):e1001661. doi:10.1371/journal.pbio.1001661

9. Vogelaar IP, Figueiredo J, van Rooij IA, et al. Identification of germline mutations in the cancer predisposing gene CDH1 in patients with orofacial clefts. Hum Mol Genet. 2013;22(5):919-926. doi:10.1093/hmg/dds497