Evaluating the effects of depression genetics and treatment on clinical laboratory values

By

Julia Mae Sealock

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

May 13th, 2022

Nashville, Tennessee

Approved:

Lea K Davis, PhD, Advisor

Jennifer Below, PhD, Committee Chair

Lisa Bastarache, MS

Nancy J Cox, PhD

Douglas Ruderfer, PhD

Colin Walsh, MD

To my grandmother, Dorothy Mae Marco Farmer

from whom I inherited a middle name and a strong will,

whose life motivated me to

ask questions, seek answers, and help others.

To my amazing family,

my mom, dad, and sister:

who supported my dreams,

who encouraged me always,

who believed in me when I did not believe in myself.

ACKNOWLEDGEMENTS

This work was funded by the Vanderbilt Training Program on Genetic Variation and Human Phenotypes (T32-GM080178and the Ruth L. Kirschstein National Research Service Award Fellowship (1F31-MH124306-01A1)

This work was made possible through incredible mentorship from Dr. Lea K. Davis, wonderful lab mates and collaborators, and a supportive training program in the Vanderbilt Genetics Institute. Thank you, Lea, for creating an environment that puts people first, for encouraging and shaping my scientific interests, and helping me become a better scientist and person. I look forward to many more years of your mentorship. I am particularly thankful for the VGI for creating a welcoming, friendly environment with rigorous science and a focus on trainee development and well-being. Finally, thank you to the members of my Dissertation Committee who always asked insightful questions and provided sound scientific and professional guidance.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

The connection between the mind and body in depression

Depression is a common psychiatric disorder estimated to affect 264 million individuals worldwide[1]. Diagnostic criteria for depression include clinical evaluation of self-reported psychiatric symptoms, such as depressed mood, irritability, anhedonia, or suicidal thoughts. Depression can also manifest with physical symptoms such as weight change, appetite change, sleep disturbances, and psychomotor changes. Additionally, depression diagnosis associates with increased risk for cardiovascular disease[2–4], autoimmune disease[5], and diabetes[6–9]. The physical symptoms and increased risk of peripheral diseases suggest a connection between the mind and the body in depression.

Hippocrates was the first to describe "melancholia", now termed depression, as "fear or sadness that lasts a long time"[10]. In addition to psychiatric symptoms, Hippocrates mentioned several physical manifestations still used in diagnostics today, such as "aversion to food, sleeplessness, restlessness". Ancient Greeks believed in a connection between the mind and the body through black bile, one of the four humors thought to control health. Melancholia was believed to be caused by an excess of black bile. Treatment aimed at balancing humors through air, exercise, or restorative sleep, and was often intertwined with religious or folklore methods over time. The humor theory of depression lasted in medicine until the 1850s[11].

In contrast to the Hippocratic connection between the mind and the body, modern mental healthcare tends to be separated from somatic healthcare. The separation of mental and physical healthcare finds its origins in Descartes's theory of dualism. Dualism states two worlds exist: the outer, physical world and the inner, spiritual world. Dualism still permeates modern psychiatric medicine today through the physical separation and lack of integrated mental and physical healthcare, dismissal of somatic concerns among individuals with severe mental illness, and the disproven theory of immune privilege of the brain[12].

In contrast to dualism, many associations exist between mental and physical health. For example, individuals with severe mental illness, including depression, have a life expectancy 10 years shorter than the general population[13,14]. The decreased life expectancy is only partially explained by suicide. Somatic disorders such as cardiovascular disease, diabetes, and cancer account for the majority of premature death, suggesting a biologic connection between psychiatric and somatic conditions[14]. Additionally, depression has a number of common somatic co-morbidities, including cardiovascular disease, diabetes, and autoimmune disorders. Interestingly, all of these disorders are linked to an activated immune system, leading to the hypothesis that the brain-body connection in these instances are, at least in part, due to inflammation, as opposed to previous ideas of common environment[15].

The direction of association between depression and somatic conditions remains unclear. Investigators in Denmark used a population-wide health registry to examine whether hospitalization for depression or autoimmune disease came first. While one group found depression precedes autoimmune disease[16], another group found the opposite[5], suggesting a bidirectional relationship and shared biology. Other methods of dissecting the relationship

between depression and somatic conditions used genetics. Large-scale genome-wide studies (GWAS) of depression significantly correlate with somatic conditions, including inflammatory conditions, such as Crohn's disease and irritable bowel syndrome, as well as cardiovascular disease, and lung cancer[17,18]. The epidemiologic and genetic link between depression and somatic conditions, many of which involve immune or inflammatory biology, suggests a possible inflammatory link between the two traits.

The role of inflammation in depression

Biomarkers are accurate and reproducible measurements that indicate the medical state of an individual[19]. Increased levels of circulating pro-inflammatory markers can indicate a disease state, such as an autoimmune disorder. Slightly elevated but not abnormal levels of pro-inflammatory markers can be indicative of chronic inflammation[20]. In cross-sectional studies, depression cases consistently have slightly increased levels of pro-inflammatory markers than controls, including white blood cell count (WBC)[21,22], C-reactive protein (CRP)[23], interleukin-6 (IL-6), and tumor necrosis factor-alpha (TNF-α). Multiple meta-analyses confirmed the association between depression and WBC[24], CRP[25,26], IL-6[25,27,28], and TNF-α[27] across multiple studies. Cross-sectional studies helped establish an association between inflammatory markers and depression, however, they cannot indicate whether increased inflammation preceded or followed depressive symptoms.

Longitudinal studies involve following individuals over a period of time and collecting repeated observations, such as biomarker measurements. Longitudinal studies of depression and inflammatory markers have shown that inflammation precedes and, in some cases, can

predict the onset of depression. Several studies indicated increased CRP associated with later depression events, such as increased risk of hospitalization due to depression, increased reporting of depressive symptoms, and increased risk of depression diagnosis[23,26,29]. Additionally, a study reported increased levels of IL-6 at 9 years of age associated with an increased risk of depression at 18 years of age[30]. A study of WBC count in urban adults showed higher baseline levels associated with a faster increase of depressive symptoms over time[22]. At the same time, longitudinal studies have shown clinical depression also associates with subsequent increase in pro-inflammatory markers, notably CRP and IL-6[31,32]. Overall, longitudinal studies have established increases in pro-inflammatory markers can precede depression symptoms and diagnosis, with emerging evidence suggesting the relationship is bidirectional.

Alternative ways of discerning the direction between depression and inflammation include inducing or reducing inflammation and monitoring depression symptoms. Both typhoid vaccination and interferon-α treatment activate the immune system and associate with increases in depressive symptoms. Typhoid vaccination associates with a transient decrease in mood that subsides 24 hours after vaccination[33,34]. Interferon-α is used for the treatment of Hepatitis C, often for lengths of 6 months to 1 year. Several studies have shown interferon-α associates with increases in depressive symptoms that return to pre-treatment levels after treatment is stopped[35–37].

Other studies report that decreasing inflammation using anti-inflammatory medications shows modest antidepressant effects. Three meta-analyses demonstrated anti-cytokine therapies associated with decreases in depressive symptoms compared to placebo[38–40]. Most of

these studies were conducted using secondary endpoints for medications developed for treatment of inflammatory conditions, raising the hypothesis that depressive symptoms can be improved through modulation of the immune system. One meta-analysis found the reduction in depressive symptoms remained even after controlling for improvement in physical health[40]. To date, only one double-blind randomized controlled trial has been conducted to investigate if an anti-inflammatory medication reduces depressive symptoms in a sample diagnosed with major depression. The study did not show improvement in depressive symptoms after treatment with an anti-TNF-α therapy, however, the sample size was small (N=60)[41].

Development of antidepressants

Antidepressants are first-line treatment for depression and are among the most commonly prescribed medications in the United States. However, antidepressants were not initially developed for depression. In the 1950s, iproniazid was developed as an antibiotic for treatment of tuberculosis. In addition to successfully treating tuberculosis, patients exhibited boosts in mood and energy, with reporters noting there was "dancing in the halls tho' there were holes in their lungs"[42]. Dr. Nathan Kline of Rockland State Hospital was eager to use iproniazid in treatment in depression, even though others believed the mood elevation seen in tuberculosis patients was a result of improved physical condition rather than anti-depressive effects of the medication. Improvements in mood remained when Kline tested iproniazid in non-tuberculosis patients, leading to its use as a treatment for depression starting in 1957. After approval of iproniazid, drugs chemically similar to iproniazid were developed, tested, and

approved for depression treatment, creating the first class of antidepressants, monoamine oxidase inhibitors (MAOIs).

The next class of drugs approved for treatment of depression was also an accidental discovery. In 1950s, imipramine was tested in schizophrenic patients with agitation. Patients diagnosed with depressive psychosis and treated with imipramine showed improvement in their general state and depressive symptoms, spurring imipramine to be tested and subsequently approved for treatment of depression, creating the tricyclic class of antidepressants (TCAs)[42].

Despite two classes of antidepressants approved for use, the mechanisms of action were unknown. A deeper look at the mechanism of MAOIs and TCAs revealed the drugs inhibited the uptake of monoamines and enhanced monoaminergic neurotransmission. This discovery led to the catecholamine hypothesis of depression which states depression is the result in a deficiency in monoaminergic function[43]. However, it remains unknown whether the increased availability of monoamines from antidepressant treatment leads to an improvement in mood, or if an alternative mechanism of action produces antidepressant effects.

The first antidepressant intentionally developed for treatment of depression was fluoxetine (brand name “Prozac”) in 1974. Following the catecholamine theory of depression, fluoxetine was designed to specifically increase the availability of serotonin in synapses[43]. Fluoxetine was approved for depression treatment in 1987 and was the first selective serotonin reuptake inhibitor (SSRI) to be used in the United States. With similar efficacy as MAOIs and TCAs but with fewer side effect, SSRIs are the most widely prescribed antidepressants to date. The final class of antidepressants are selective serotonin and norepinephrine reuptake

inhibitors (SNRIs). Similar to SSRIs, SNRIs increase the availability of serotonin and norepinephrine. There have not been any major advances in depression treatment since the 1990s.

Anti-inflammatory action of antidepressants

The pro-inflammatory state in depression and the origins of antidepressants from antibiotics led to investigations of the anti-inflammatory action of antidepressants. Overall, studies show mixed results on whether antidepressants associate with changes in inflammatory markers. A meta-analysis of 22 studies found no association between antidepressants and TNF-alpha or IL-6, and a very weak association with IL-2B[44]. A separate meta-analysis reported antidepressants associated with decreases in IL-6 and CRP[45]. Both meta-analyses reported high heterogeneity between studies, complicating interpretability of the meta-analyses. Only one small study (N=15) examined the effect of antidepressants on WBC and reported no association[46].

Associations between decreases in inflammatory markers and antidepressant response also show mixed results. The most recent meta-analysis demonstrated IL-6 decreases with treatment regardless of treatment response and persistently high levels of TNF-a associated with treatment resistance[47]. No other associations with inflammatory markers emerged.

The role of genetics in depression

While observational studies have been useful in establishing a link between the immune system and depression, another method to investigate the biology of depression is to use

genetics. Prior to genome-wide association studies, the genetics of depression was investigated using family studies. Twin studies evaluate the contribution of genetics to a trait while controlling for a common environment. In twin studies of depression, the estimated heritability ranged from 31-42%[48,49]. Twin studies provide valuable estimates of heritability that include both common and rare genetic variation, however, they are unable to pinpoint exact genes or mechanisms underlying the trait. Linkage analysis uses families with multiple affected individuals to identify regions in the genome inherited from a recent common ancestor and more common in affected individuals. Two separate reviews of linkage results in depression found 9-14 regions that replicated in at least two studies[50,51]. However, a recent re-analysis of linkage results in depression from 1998-2010 showed no significant overlap with hits from genome-wide association scans[48].

Another approach to identify genes involved in complex traits is a candidate gene study which requires a selection of a gene to compare allele frequencies in cases versus controls. In depression, candidate gene studies focused primarily on monoamine and serotonin transporter genes based on the mechanisms of antidepressants[51]. However, a recent re-analysis of candidate genes by Border et al. found no support for any proposed gene and concluded previous results were likely false positives[52].

The introduction of genome-wide association studies (GWAS) in 2006 allowed for large-scale analysis of polymorphisms associated with depression across the entire genome. No genome-wide significant associations were found for depression until 2015 in the CONVERGE study[53]. The latest GWAS conducted by the Psychiatric Genetics Consortium found 102 loci associated with depression[17]. The largest GWAS to date conducted in the Million Veteran

Program reported 178 significant loci[54]. GWAS has not solved the biology of depression, but it has significantly advanced the field by yielding loci for downstream analyses to investigate associated brain regions, overlap with other traits, and potential druggable targets. For example, the Million Veteran Program analysis found significant genetic correlation between depression and 669 traits, including neuroticism, cardiovascular diseases, and rheumatological diseases. Analysis of predicted tissue expression using the GWAS data implicated altered gene expression in several brain regions, such as the hypothalamus and the nucleus acumbens[55]. Changes in gene expression in the brain has been used to investigate new drug targets and propose existing drugs for repurposing. Gaspar et al. reported expression changes in 24 genes that belong to the druggable genome, including an enrichment of targets for monoamine reuptake inhibitors, sex hormones, and antihistamines[56]. In summary, GWAS analyses are foundational to the discovery of biology and new treatments in depression. A limiting factor of these downstream analyses[55,56] are they can only use the summary level information from the GWAS, rather than individual level genetic and phenotypic information.

Polygenic scores estimate the genetic risk for depression by aggregating the small effects of thousands of loci across the genome into one score for each individual[57]. Though they are not currently recommended for clinical use, PGS do capture a significant proportion of the variance in depression diagnosis (1.5-3.2%[17]), indicating PGS represent a biologically relevant contribution to depression. Large-scale GWAS have accelerated the use of polygenic scores in depression. A recent systematic review found depression PGS associated with various psychiatric traits such as schizophrenia, bipolar, and increased number of depression episodes[58]. Other studies of depression PGS found associations with alcohol dependence[59],

suicide attempt[60], and better response to lithium treatment in bipolar patients[61]. In addition to psychiatric disorders, depression PGS associated with increased risk for cardiovascular disease in two recent studies[62,63].

Genetic analysis has also been useful in dissecting the relationship between depression and immune traits. Depression GWAS show significant genetic correlation with autoimmune disorders, such as Crohn's disease and irritable bowel disorder, as well as with a pro-inflammatory biomarker, C-reactive protein[17,18,64]. These results suggest that the association between depression, increased inflammatory markers, and autoimmune disease are at least partially driven by shared genetic factors. An emerging resource for investigating genetic factors and shared biology of complex traits are electronic health records.

Utility of electronic health records and biobanks for investigating the biology of depression

Electronic health records (EHRs) store longitudinal information about the health and clinical care of individual patients, including diagnoses, medications, laboratory test results, and clinical notes. Traditional cohort collection for psychiatric research involves expensive ascertainment of cases and controls through clinical interviews. EHRs provide a low-cost resource to identify cases and controls through structured data such as billing codes or unstructured data such as mining clinical notes with natural language processing[65]. Psychiatric cohorts collected through EHRs have high phenotypic concordance and genetic correlation with interview-based collection[65].

Biobanks that link EHRs to DNA provide an opportunity to analyze clinical information along with genetic risk factors. Biobanks are useful in depression genetics research as well.

Currently, the largest GWAS of depression was conducted using the Million Veteran Program biobank. Additionally, previous studies consistently use depression diagnoses derived from EHRs from the UKBiobank and the iPSYCH collection in Denmark for genetic analyses. Biobanks are powerful resources for polygenic score analyses by allowing for the calculation of genetic liability regardless of diagnostic status and downstream associations with a variety of phenotypes stored in the EHR, such as clinical laboratory ("lab") tests.

In the clinic, lab test results are essential to routine care. These targeted biochemical measurements facilitate disease diagnosis and influence health care delivery. Clinical lab values are also monitored as mediators of disease risk, and are targeted by interventions to reduce disease incidence (e.g., cholesterol-lowering medication to reduce the risk of heart disease). Lab test results in EHRs are a vast and growing resource for novel biomarker discovery, especially as EHRs are increasingly linked to patient DNA samples (e.g., the eMERGE consortium the All of Us Program, and the Million Veteran's Program). Despite their potential, however, EHR-based labs have been used in only a handful of prior genetic studies[66–70], and none have systematically interrogated an extended collection of EHR-based lab values.

Pairing genetics with biomarkers can help elucidate the biology underpinning depression and help provide the basis for future development of diagnostic panels. Furthermore, screening for associations with all available clinical lab tests can help replicate known associations (e.g., immune markers), and also yield novel associations. In this thesis, we 1) introduce a method for screening for associations with lab values derived from EHRs, 2) test for lab-wide associations with depression genetics, 3) validate findings between depression

genetics and white blood cell count across three biobanks, and 4) interrogate the effects of antidepressants on white blood cell count.

CHAPTER II

DEVELOPING METHODS TO ANALYZE LABORATORY VALUES STORED IN ELECTRONIC HEALTH RECORDS*[1]

Introduction

In this chapter, we describe a method developed to conduct large-scale analysis on clinical laboratory values ("labs") extracted from electronic health records (EHRs). Our method Lab-Wide Association Scan (LabWAS) finds associations between any variable of interest (genetic or otherwise) and cleaned EHR labs. EHR labs used in LabWAS are cleaned using our recently developed QualityLab pipeline that is used to perform quality control on EHR-derived labs. Briefly, QualityLab removes non-numeric values, filters for a single unit within a lab, and removes outlier values consistent with biologically implausible values or data entry errors (Appendix A). When applied to VUMC data, QualityLab produced cleaned data on 939 labs for downstream analyses described here.

We hypothesized that EHR-based lab values could be used to identify known and novel relationships between genetic risk factors, biomarkers, and disease. We deployed our framework in the Vanderbilt University Medical Center (VUMC) EHR and linked biobank (BioVU) and replicated it in an independent biobank, Massachusetts General Brigham Biobank (MGB). To validate our method, we focused on genetic analysis of blood values of high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), and triglycerides (TG),

[1] *Adapted with permission from Dennis JK & Sealock JM et al., Genome Medicine, 2021

and on coronary artery disease (CAD) as proof-of-principle examples to test the association between PGS for CAD and known biomarkers of disease (LDL, HDL, and TG).

Methods

Study Sample

Our primary analysis was performed at VUMC which is a tertiary care center providing inpatient and outpatient care in Nashville, TN. The VUMC EHR was established in 1994 and includes data on billing codes from the International Classification of Diseases, 9th and 10th editions (ICD-9 and ICD-10), Current Procedural Terminology (CPT) codes, laboratory values, reports, and clinical documentation. The de-identified mirror of the EHR, known as the Synthetic Derivative, includes patient records on more than 3 million individuals. In 2007, VUMC launched a biobank, BioVU, and the BioVU Consent form is provided to patients in the outpatient clinic environments at VUMC. The form states policies on data sharing and privacy, and upon consent, makes any blood leftover from clinical care eligible for BioVU banking[71]. The VUMC Institutional Review Board oversees BioVU and approved this project (IRB# 160302).

Genotyping and Quality Control

We obtained genotype information on 94,474 VUMC individuals of different ancestral backgrounds genotyped on the Illumina MEGA^{EX} array. Using PLINK v1.9[72] genotypes were filtered for SNP and individual call rates, sex discrepancies, and excessive heterozygosity. We selected individuals of European or African ancestry using principal component analysis implemented in Eigenstrat[73,74] and confirmed the absence of genotyping batch effects through

logistic regression with 'batch' as the phenotype. Imputation was completed using the Michigan Imputation Server[75] and the Haplotype Reference Consortium (HRC) reference panel. SNPs were then filtered for SNP imputation quality (R^2>0.3) and converted to hard calls. We restricted to autosomal SNPs, filtered SNPs with minor allele frequency >0.01, or with allele frequencies that differed by more than 10% from the 1000 Genomes Project phase 3 CEU or ASW set respectively [76], and Hardy-Weinberg Equilibrium ($p>1x10^{-10}$). The resulting dataset contained 6,303,629 SNPs on 72,824 individuals of European genetic ancestry and 12,798,111 SNPs on 15,283 individuals of African genetic ancestry.

Lab Heritability and GWAS Analyses

Prior to calculating SNP-based heritability (h^2_{SNP}), we first calculated pairwise relatedness in the VUMC genotyped sample and removed one related individual from pairs with pi-hat greater than 0.05. This stringent threshold was chosen based on prior experience and previously published best practices in the application of restricted maximum likelihood (REML) approaches to the calculation of h^2_{SNP}[77]. After filtering, 45,010 individuals of European genetic ancestry remained. We then used the Genome-wide Complex Trait Analysis (GCTA) package (version 1.92.4)[78] to create a pairwise genetic relationship matrix for all individuals, and heritabilities were calculated using the restricted maximum likelihood (REML) method, which estimates the variance explained by all the SNPs for a trait. We used the median, rank-based inverse normal transformed (INT) lab values from the QualityLab pipeline, and of the 939 analyzed labs, 335 demonstrated non-zero heritability. For GWAS analyses, we used a less stringent relatedness filter appropriate to GWAS (pi-hat>0.2)[79] resulting in a total available

sample of 66,732 European descent individuals. Next, we subset to the heritable labs with at least 1,000 individuals (n=181), and performed GWAS of the median, INT-transformed lab values using fastGWA[80]. All h^2_{SNP} and GWAS analyses included covariates for sex, cubic splines (knots=4) of median age across the medical record (to control for non-linear effects of age), and the top 10 principal components of estimated from the genetic data.

Heritability and GWAS Analyses of Lipids

We benchmarked our lipid h^2_{SNP} estimates against those from two external datasets, the Global Lipids Genetics Consortium (GLGC)[81] and the Million Veterans Program (MVP). GLGC and MVP estimates of h^2_{SNP} for HDL, LDL, and TG were calculated from GWAS summary statistics using LDSC[82]. We computed h^2_{SNP} in VUMC using Linkage Disequilibrium Score regression (LDSC) applied to our fastGWA summary statistics for HDL, LDL, and TG. However, because LDSC can underestimate h^2_{SNP}[83], we also calculated h^2_{SNP} using GCTA. In addition to these h^2_{SNP} comparisons, we calculated the genetic correlations (r_g) between the VUMC lipid GWASs and the GLGC and MVP lipid GWASs using LDSC and the pre-computed European LD scores from 1000 Genomes Phase 3 European data[84]. We also calculated genetic correlations using a new method, High-Definition Likelihood[85], which fully accounts for linkage disequilibrium across the genome and is more suitable for traits with lower heritability than LDSC. In sensitivity analyses, we repeated genetic correlations of LDL after controlling the VUMC GWASs for coronary atherosclerosis or diabetes diagnoses, defined as phecodes 411, “Ischemic heart disease” and 249, “Secondary diabetes mellitus”.

To validate EHR-based lipid values, we tested the robustness of HDL, LDL, and TG h^2_{SNP} estimates to different lab value and patient filters. First, we excluded lipid measurements that occurred after the first mention of lipid-altering mediation in the EHR, and re-calculated each patient's pre-medication median values of HDL, LDL, and TG. Second, we excluded patients with a diagnosis of CAD, defined by the phecode 411.

LabWAS Pipeline

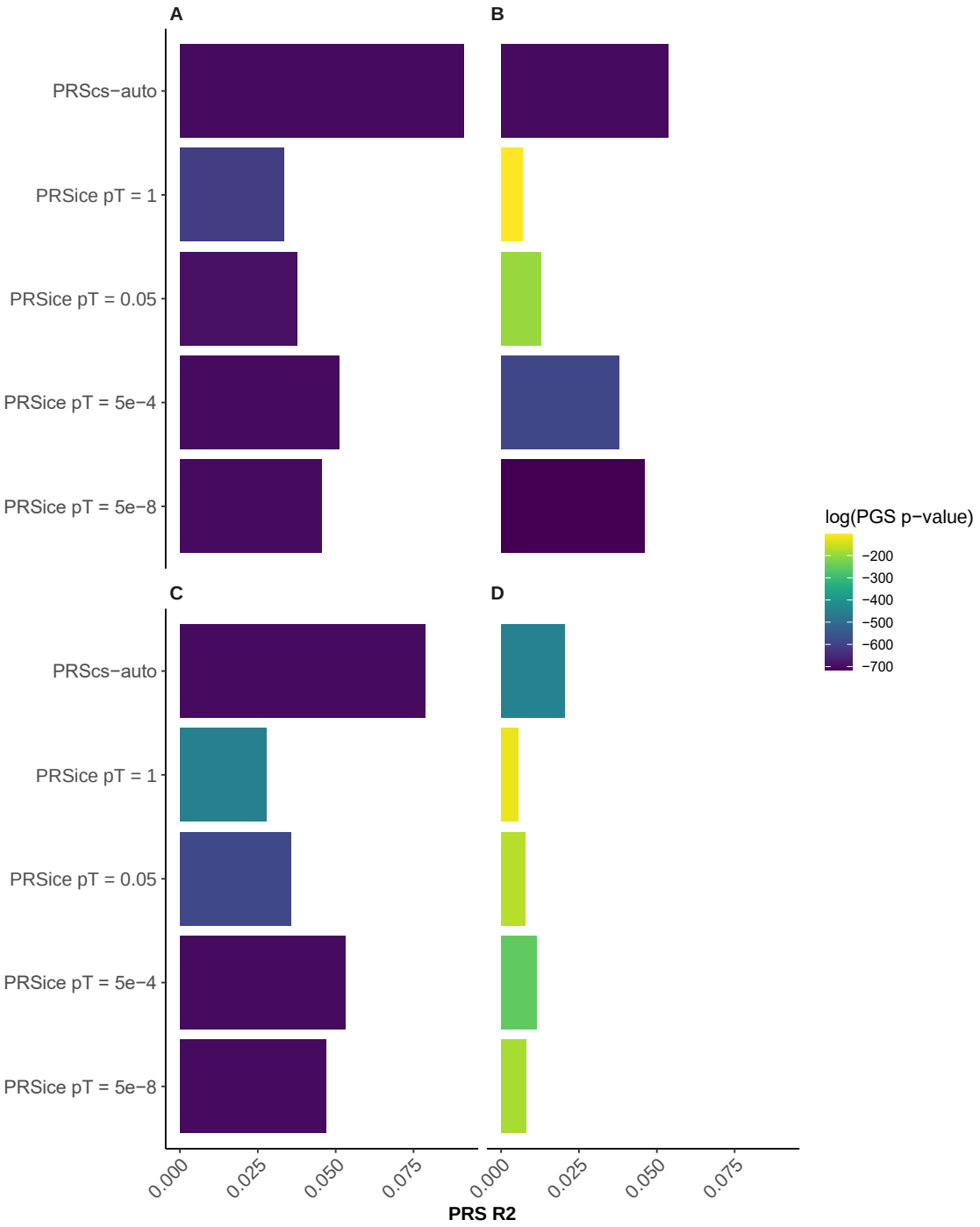
LabWAS uses the median, INT-transformed lab values from the QualityLab pipeline in a linear regression to determine the association with an input variable, adjusting for covariates. In these analyses, a primary goal of the LabWAS was to test common population genetic variation (e.g., PGS) for association with common population variation in lab values. We therefore only included the 335 labs with non-zero h^2_{SNP}. Additionally, we imposed a minimum sample size requirement of 100 for a lab to be included in the LabWAS analysis, bringing the number of labs tested in each scan to 315 in the European ancestry set and 226 in the African ancestry set.

Polygenic Scoring

Prior to polygenic scoring, we randomly removed one related individual from pairs with pi-hat greater than 0.2, leaving 66,732 individuals of European genetic ancestry and 12,383 individuals of African genetic ancestry. We generated lipids PGS for these individuals using PRS-CS[86] with weights derived from the multi-ancestry MVP lipid GWAS summary statistics[69]. PGS for CAD was calculated using SNP weights from CARDIoGRAMplusC4D GWAS summary

statistics[87] using PRS-CS. Because the majority of the MVP multi-ancestry sample was European, linkage disequilibrium was modeled using the pre-calculated European panel. PRS-CS is a recently developed Bayesian polygenic prediction method that imposes continuous shrinkage priors on SNP effect sizes (Polygenic Risk Score – Continuous Shrinkage)[86]. These priors can be represented as global-local scale mixtures of normals which allows the model to flexibly adapt to differing genetic architectures and provides substantial computational advantages. The shrinkage parameter was automatically learnt from the data (i.e., using PRS-CS-auto). SNP effect estimates were obtained from GWAS summary statistics and the score was calculated using a linkage disequilibrium reference panel from 503 European samples in the 1000 Genomes Project phase 3[76]. Although PRS-CS outperformed other polygenic scoring methods across a range of traits in previous experiments, its superiority may not hold across all genetic architectures[86]. We therefore also generated PGS for the European sample using PRSice-2[88] using four p-value thresholds (1, 0.05, $5x10^{-4}$, $5x10^{-8}$), and have automated a pipeline to generate scores across both methods. PGS were scaled to have a mean of zero and SD of one before testing for association with any outcome variables. We validated each score by testing the proportion of trait variability explained by the PGS, controlling for sex, cubic splines of median age (4 knots) across the medical record, and the top 10 principal components to adjust for genetic ancestry.

Figure 1. Predictive abilities of polygenic scores calculated by PRScs and PRSice in VUMC. Predictive ability is measured as the proportion of same-trait variability explained (R^2) by (A) HDL PGS, (B) LDL PGS, (C) TG PGS, and (D) CAD PGS.



LabWAS of Polygenic Scores

PGS for LDL, HDL, and TG, were calculated in VUMC participants using PRS-CS and applying SNP weights from the MVP GWAS summary statistics. We then ran LabWAS of LDL, HDL, and TG polygenic scores to test whether lipid labs were robustly associated with the genetic scores to which they corresponded. Next, a PGS for CAD was calculated using SNP weights from CARDIoGRAMplusC4D GWAS summary statistics[87] and a LabWAS of CAD PGS to test whether the score could identify lab traits associated with genetic risk for CAD, before and after controlling for a CAD diagnosis. Each LabWAS was controlled for sex, cubic splines of median age across the medical record, and the top 10 principal components of ancestry. Results are reported as effect estimates and their 95% confidence intervals per SD increase in the PGS. The Bonferroni-corrected threshold for statistical significance across all tested labs was 3.97×10^{-5} (0.05/(315 x 4)).

Replication in Massachusetts General Brigham Biobank

We next sought to replicate the associations between lipids PGS and referent lipids as well as the significant associations with CAD PGS in an external biobank. The MGB, previously the Partners Biobank, is an ongoing virtual cohort study of patients across the Partners HealthCare hospital system (including Brigham and Women's Hospital, Massachusetts General Hospital, and other affiliated hospitals), which provides a large-scale resource of linked longitudinal EHR data, genomic data, and self-reported survey data[89]. All patients provided informed consent before enrollment, and all study procedures were approved by the Partners HealthCare Institutional Review Board.

Lab values were extracted from EHRs and cleaned using QualityLab, resulting in 759 labs for analysis. The median value for each lab trait for each individual was selected and inverse normalized. Lab heritabilities were calculated using REML in GCTA. Of 759 labs that passed QualityLab, 241 demonstrated measurable heritability and included a sample size of at least 100 individuals.

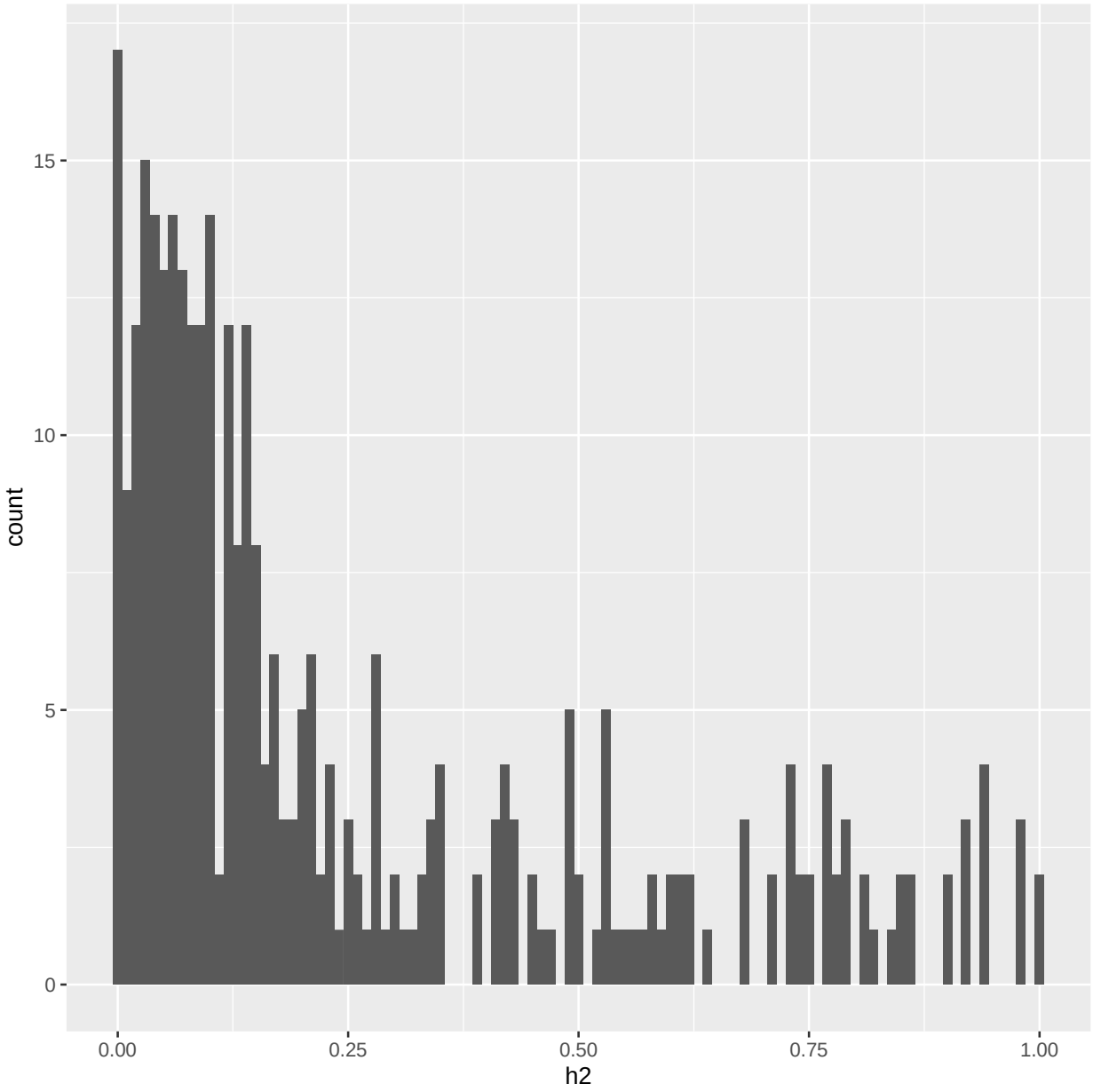
Polygenic scores for HDL, LDL, TG, and CAD were calculated on individuals of European descent in MGB (n=25,698) using the same criteria as VUMC. Lipids and CAD polygenic scores were associated with each of 234 labs using LabWAS. Lastly, the associations between CAD PGS and lab traits were controlled for CAD diagnosis, defined by phecode 411 (N cases = 1,094, N controls = 20,405). All associations were controlled for sex, top 10 principal components, and the first three splines of median age across the medical record.

Results

Heritability and GWAS Analyses of All Labs

In VUMC, out of 939 clean lab traits, 335 demonstrated non-zero h^2_{SNP} and the point estimates ranged from 2×10^{-6} to 0.98. (Figure 2). As a resource for the community, the GWAS summary statistics for the labs with calculable heritability and a minimum sample size of 1,000 individuals (n=181) are available in the GWAS Catalog (Study Number: GCP000091; accession numbers GCST90012603 - GCST90012784).

Figure 2. Histogram of heritability estimates of VUMC labs.



Heritability and GWAS Analyses of Lipids

The h^2_{SNP} estimates in VUMC were robust to removing post-medication observations, and to removing CAD cases. The number of participants included in these analyses, however, was smaller, and so the standard errors of these h^2_{SNP} estimates were larger (Figure 3a; Table 1). Both GCTA and LDSC gave similar estimates of h^2_{SNP} in VUMC (Figure 3b, Table 1), and the LDSC estimates in VUMC were comparable to those in the GLGC and MVP for all lipids.

Genetic correlation between VUMC and GLGC summary statistics was strong for HDL (LDSC: rg=0.96, SE=0.08, p-value=2.69 x 10^{-35}, High-Definition Likelihood: rg=0.92, SE=0.11, p-value=3.25 x 10^{-17}) and TG (LDSC: rg=0.94, SE=0.05, p-value=5.86 x 10^{-97}, High-Definition Likelihood: rg=0.89, SE=0.11, p-value=7.69 x 10^{-17}). When comparing VUMC and MVP, the correlations for HDL (LDSC: rg = 0.99, p-value = 7.51 x 10^{-61}, High-Definition Likelihood: rg=0.89, SE=0.08, p-value=2.24 x 10^{-27}), and TG (LDSC: rg=0.94, p-value=2.28x10^{-99}, High-Definition Likelihood: rg=0.88, SE=0.10, p-value=4.84 x 10^{-18}) were nearly perfect (Figure 4a, Table 2). The LDL and LDL pre-medication genetic correlations between GLGC and VUMC were not calculable using LDSC due to low heritability. Using High-Definition Likelihood, GLGC LDL levels were significantly correlated when median LDL values across the entire EHR (rg = 0.44, SE=0.10, p-value = 1.08 x 10^{-5}) and median pre-medication LDL values (rg = 0.50, SE=0.08, p-value = 6.38 x 10^{-10}). The comparison between VUMC and MVP showed a stronger correlation for LDL (LDSC: rg=0.84, SE=0.17, p-value=1.47 x 10^{-6}; High-Definition Likelihood: rg=0.53, SE=0.09, p-value=1.52 x 10^{-11}). The genetic correlation with MVP increased when we restricted to pre-medication values of LDL in VUMC (LDSC: rg=0.89, SE=0.22, p-value=2.90 x 10^{-5}; High-Definition Likelihood: rg=0.56, SE=0.07, p-value=2.06 x 10^{-15}) (Figure 4a, Table 2), and increased further

when we controlled for coronary atherosclerosis and diabetes diagnoses (GLGC, High Definition Likelihood: rg=0.57, SE=0.09, p-value=8.88 x 10^{-9}, MVP, LDSC: rg = 1.00, SE=0.34, p-value = 0.004), MVP, High Definition Likelihood: rg=0.55, SE=0.09, p-value=1.50 x 10^{-8}) (Figure 4b, Table 2).

Figure 3. Heritability comparison of lipids. (A) Estimates of heritability computed by GCTA in VUMC patients were robust to excluding individuals with a diagnosis of CAD and to removing post-medication observations. (B) Estimates of heritability computed using GWAS summary statistics and LDSC were comparable across VUMC and the Global Lipids Genetic Consortium (GLGC) and Million Veteran's Project (MVP) samples.

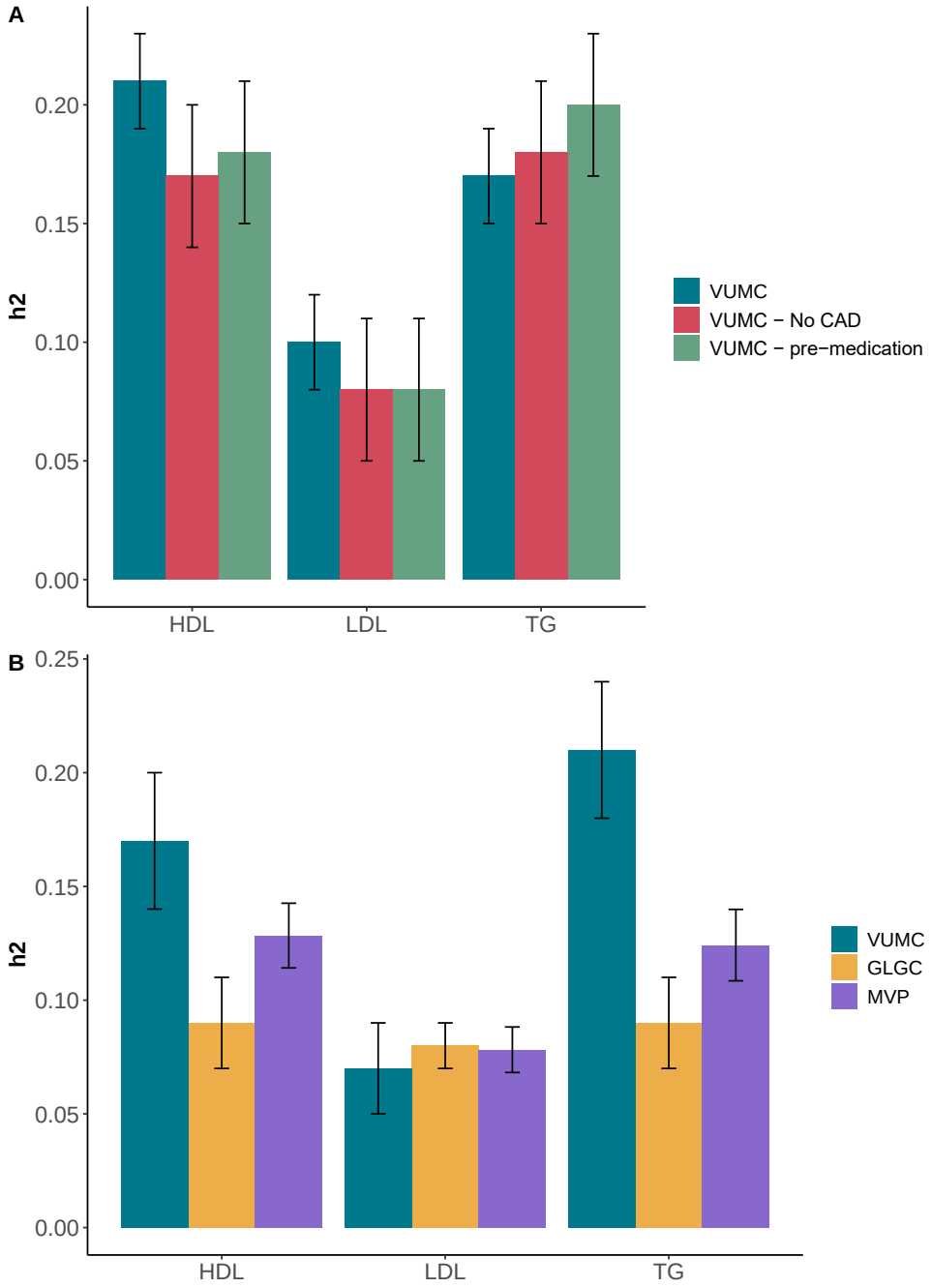


Figure 4. Lipids genetic correlation. (A) Genetic correlations between lipid levels in VUMC and the Global Lipids Genetic Consortium (GLGC) or Million Veteran's Program (MVP) calculated using LDSC or high-definition likelihood (HDL). (B) Genetic correlation between GLGC/MVP LDL and VUMC LDL controlled for CAD and diabetes diagnosis using LDSC (solid) and High-Definition Likelihood (dashed).

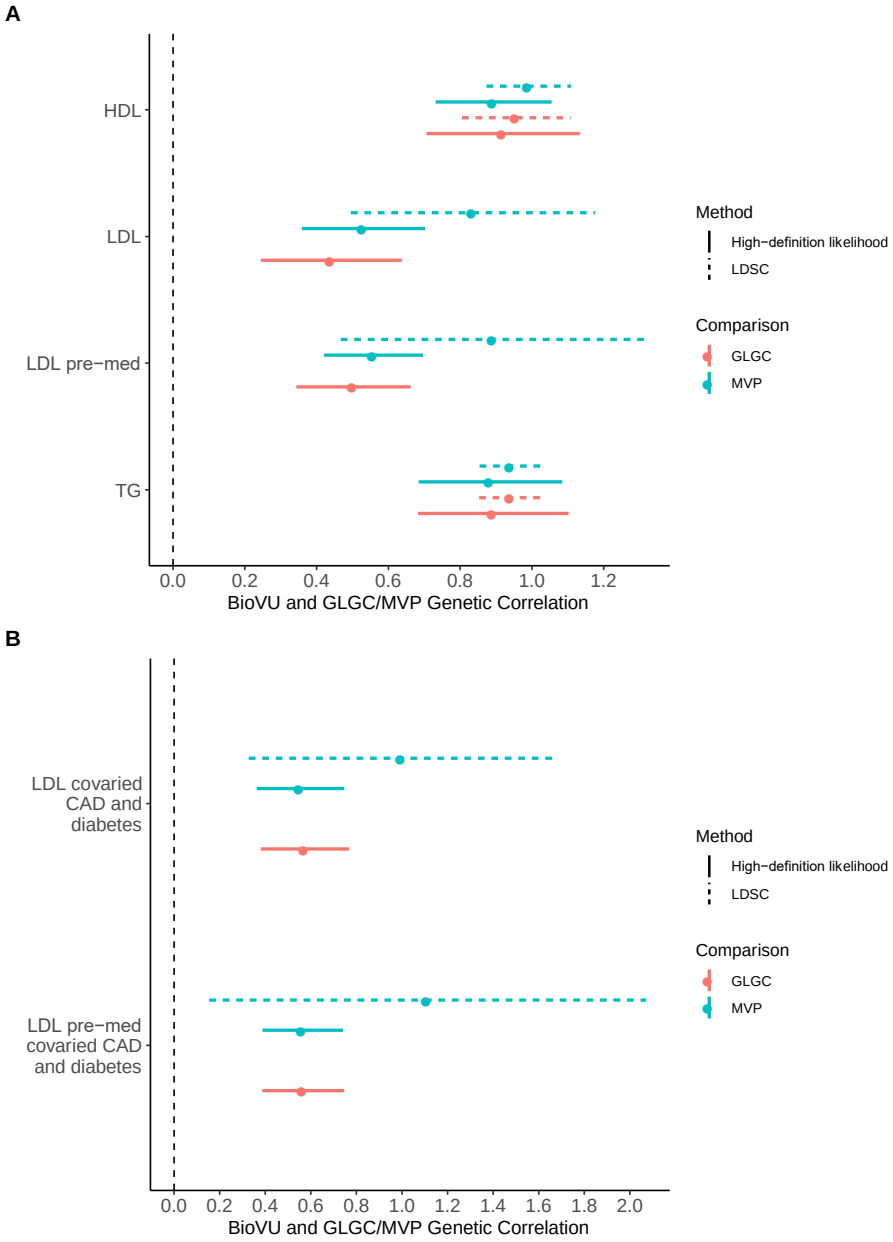


Table 1. Comparison of heritability estimates for lipids between VUMC, GLGC, and MVP. Heritability was calculated using two methods, LDSC and GCTA.

Lipid	h^2_{SNP} Estimation Method	Population	N	h^2_{SNP}	SE
HDL	LDSC	GLGC	188,577	0.09	0.02
		MVP	291,746	0.13	0.01
		VUMC	29,497	0.17	0.03
	GCTA	VUMC	25,179	0.21	0.02
		VUMC pre-medication	22,865	0.18	0.03
		VUMC no CAD	22,123	0.17	0.03
LDL	LDSC	GLGC	188,577	0.08	0.01
		MVP	297,218	0.08	0.01
		VUMC	28,417	0.07	0.02
	GCTA	VUMC	24,320	0.1	0.02
		VUMC pre-medication	21,756	0.08	0.03
		VUMC no CAD	21,440	0.08	0.03
TG	LDSC	GLGC	188,577	0.09	0.02
		MVP	291,933	0.12	0.02
		VUMC	30,597	0.21	0.03
	GCTA	VUMC	26,151	0.17	0.02
		VUMC pre-medication	23,861	0.2	0.03
		VUMC no CAD	23,134	0.18	0.03

Table 2. Genetic correlation of VUMC lipids with GLGC and MVP. Genetic correlation was calculated using two methods, LDSC and High-definition likelihood (HDL).

Lipid	Comparison	Method	rg	SE	p-value
HDL	GLGC	LDSC	0.957	0.077	2.69E-35
		HDL	0.920	0.109	3.25E-17
	MVP	LDSC	0.991	0.060	7.51E-61
		HDL	0.893	0.082	2.24E-27
LDL	GLGC	HDL	0.441	0.100	1.08E-05
	MVP	LDSC	0.836	0.174	1.47E-06
		HDL	0.530	0.088	1.52E-11
LDL pre-medication	GLGC	HDL	0.503	0.081	6.38E-10
	MVP	LDSC	0.893	0.217	3.90E-05
		HDL	0.559	0.070	2.06E-15
LDL covaried CAD and diabetes	GLGC	HDL	0.575	0.099	8.88E-09
	MVP	LDSC	1.001	0.343	0.0036
		HDL	0.555	0.098	1.50E-08
LDL pre-medication covaried CAD and diabetes	GLGC	HDL	0.567	0.092	9.08E-13
	MVP	LDSC	1.113	0.489	0.0228
		HDL	0.565	0.090	4.01E-10
TG	GLGC	LDSC	0.942	0.045	5.86E-97
		HDL	0.893	0.107	7.69E-17
	MVP	LDSC	0.941	0.045	2.28E-99
		HDL	0.884	0.102	4.84E-18

LabWAS of Polygenic Scores for Lipids in European Ancestry Individuals in VUMC

A LabWAS of HDL PGS in the European sample was associated with levels of several metabolic markers (Figure 5a, Table 3), including increased HDL (p-value<2.23 x 10^{-308}, beta = 0.31), decreased TG (p-value = 2.06 x 10^{-171}, beta = -0.16), decreased total cholesterol to HDL ratio (p-value = 2.54 x 10^{-44}, beta= -0.22), increased total blood cholesterol (p-value = 2.51 x 10^{-37}, beta = 0.07), and decreased blood glucose (p-value = 4.62 x 10^{-32}, beta = -0.04), decreased blood urea nitrogen (p-value = 1.48 x 10^{-15}, beta = -0.03), decreased glycated hemoglobin (p-value = 1.52 x 10^{-12}, beta = -0.05), decreased bedside glucose (p-value = 1.03 x 10^{-11}, beta = -0.07), and decreased whole blood glucose (p-value = 2.49 x 10^{-5}, beta = -0.03). HDL_{PGS} was also associated with four immune labs, white blood cell count (p-value = 6.14 x 10^{-13}, beta = -0.03), absolute neutrophil count (p-value = 5.69 x 10^{-7}, beta = -0.03), immature granulocytes (p-value = 7.86 x 10^{-6}, beta = -0.02), and monocyte to leukocyte ratio (p-value = 9.13 x 10^{-6}, beta = 0.02). Five blood biomarkers associated with HDL_{PGS}, mean corpuscular volume (p-value = 3.48 x 10^{-17}, beta = 0.03), blood carbon dioxide (p-value = 6.69 x 10^{-11}, beta = 0.02), mean corpuscular hemoglobin (p-value = 9.53 x 10^{-10}, beta = 0.02), international normalized ratio (p-value = 1.31 x 10^{-6}, beta = -0.03), and red blood cell distribution width (p-value = 2.21 x 10^{-5}, beta = -0.02). Finally, three other labs associated with HDL PGS, urate (p-value = 1.13 x 10^{-11}, beta = -0.07), creatinine (p-value = 1.42 x 10^{-10}, beta = -0.02), and urine pH (p-value = 2.22 x 10^{-8}, beta = 0.02).

The LabWAS of LDL PGS showed associations with four lipid labs (Figure 5b, Table 4). The most significant association was increased calculated LDL (p-value < 2.23 x 10^{-308}, beta = 0.24), followed by increased total blood cholesterol (p-value = 1.30 x 10^{-282}, beta = 0.20), increased directly measured LDL (p-value = 3.79 x 10^{-44}, beta = 0.19), increased non-HDL cholesterol (p-

value = 1.78 x 10^{-31}, beta = 0.19), increased total cholesterol to HDL ratio (p-value = 5.27 x 10^{-17}, beta = 0.13), and increased triglycerides (p-value = 4.47 x 10^{-6}, beta = 0.03). LDL_{PGS} also associated with four blood biomarkers, mean corpuscular hemoglobin (p-value = 5.68 x 10^{-8}, beta = -0.02), total protein in blood (p-value = 2.18 x 10^{-6}, beta = 0.02), total protein in serum (p-value = 3.00 x 10^{-6}, beta = 0.02), and mean corpuscular hemoglobin concentration (p-value = 1.50 x 10^{-5}, beta = -0.02).

The LabWAS of TG PGS was associated with several metabolic measurements (Figure 5c, Table 5), including increased TG (p-value < 2.23 x 10^{-308}, beta = 0.28), followed by decreased HDL (p-value = 4.83 x 10^{-148}, beta = -0.14), increased total cholesterol to HDL ratio (p-value = 2.95 x 10^{-28}, beta = 0.02), increased blood glucose (p-value = 1.20 x 10^{-22}, beta = 0.04), increased lipemic index (p-value = 1.57 x 10^{-18}, beta = 0.01), increased total blood cholesterol (p-value = 1.25 x 10^{-14}, beta = 0.04), increased glycated hemoglobin (p-value = 5.69 x 10^{-9}, beta = 0.04), increased bedside glucose (p-value = 2.99 x 10^{-7}, beta = 0.04), and increased non-HDL cholesterol (p-value = 1.18 x 10^{-6}, beta = 0.08). Additionally, TG PGS showed associations with seven immune labs, white blood cells (p-value = 3.90 x 10^{-30}, beta = 0.04), immature granulocytes (p-value = 1.99 x 10^{-14}, beta = 0.03), absolute lymphocytes (p-value = 2.01 x 10^{-11}, beta = 0.03), monocyte to leukocyte ratio (p-value = 5.21 x 10^{-10}, beta = -0.03), absolute neutrophils (p-value = 1.87 x 10^{-9}, beta = 0.03), complement C4 (p-value = 1.03 x 10^{-8}, beta = 0.09), and monocyte count (p-value = 6.76 x 10^{-8}, beta = -0.03). Several blood associations also emerged with TG_{PGS}, including carbon dioxide (p-value = 2.57 x 10^{-24}, beta = -0.04), total protein in blood (p-value = 4.25 x 10^{-16}, beta = 0.03), mean corpuscular volume (p-value = 9.16 x 10^{-13}, beta = -0.03), mean corpuscular hemoglobin (p-value = 9.75 x 10^{-8}, beta = -0.02), anion gap (p-

value = 2.03 x 10^{-17}, beta = 0.03), total protein in serum (p-value = 2.61 x $10^{-16,}$ beta = 0.04), and calcitriol (p-value = 1.07 x 10^{-10}, beta = -0.05). Lastly, TG PGS associated with albumin to creatinine ratio (p-value = 9.13 x 10^{-8}, beta = 0.10), urate (p-value = 6.58 x 10^{-9}, beta = 0.06), urinary pH (7.66 x 10^{-7}, beta = -0.02), and urinary albumin concentration (p-value = 2.99 x 10^{-5}, beta = 0.06).

Figure 5. Lipid PGS LabWAS in individuals of European ancestry in VUMC. (A) HDL PGS LabWAS, (B) LDL PGS LabWAS, (C) Triglycerides PGS LabWAS. The red line indicates the Bonferroni threshold for statistical significance and the blue line indicates a *p* value of 0.05. Upward triangles indicate that the PGS is associated with increased levels of the lab, while downward triangles indicate an association with reduced levels of the lab.

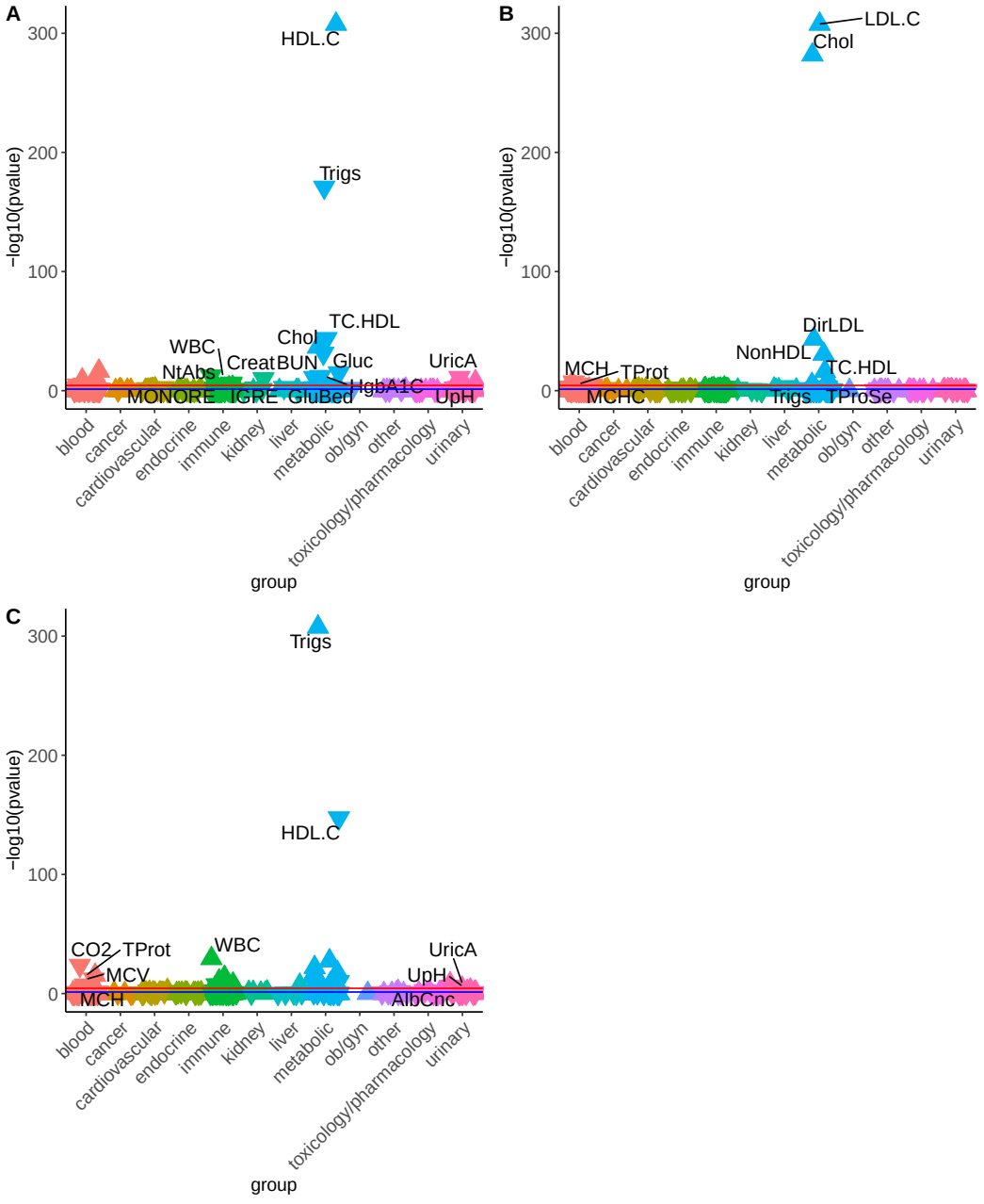


Table 3. Significant associations from the LabWAS of HDL PGS in individuals of European ancestry in VUMC.

Short Name	Full Name	Group	N	p-value	beta	SE
HDL.C	Cholesterol in HDL [Mass/volume] in Serum or Plasma	metabolic	29,497	2.23E-308	0.310	0.005
Trigs	Triglyceride [Mass/volume] in Serum or Plasma	metabolic	30,597	2.06E-171	-0.157	0.006
TC.HDL	Cholesterol.total/Cholesterol in HDL [Molar ratio] in Serum or Plasma	metabolic	3,676	2.54E-44	-0.215	0.015
Chol	Cholesterol [Mass/volume] in Serum or Plasma	metabolic	30,388	2.51E-37	0.071	0.006
Gluc	Glucose lab	metabolic	62,341	4.62E-32	-0.043	0.004
MCV	MCV [Entitic volume] by Automated count	blood	64,844	3.48E-17	0.030	0.004
BUN	Urea nitrogen serum/plasma	metabolic	62,403	1.48E-15	-0.032	0.004
WBC	Leukocytes [#/volume] in Blood by Automated count	immune	64,836	6.14E-13	-0.025	0.003
HgbA1C	Hemoglobin A1c (Glycated)	metabolic	21,558	1.52E-12	-0.047	0.007
GluBed	Glucose [Mass/volume] in Blood by Automated test strip	metabolic	25,193	1.03E-11	-0.041	0.006
UricA	Urate [Mass/volume] in Serum or Plasma	urinary	9,825	1.13E-11	-0.068	0.010
CO2	Carbon dioxide serum/plasma	blood	62,304	6.69E-11	0.023	0.003
Creat	Creatinine [Mass/volume] in Blood	kidney	62,909	1.42E-10	-0.025	0.004
MCH	MCH [Entitic mass] by Automated count	blood	64,869	9.53E-10	0.022	0.004
UpH	pH of Urine by Test strip	urinary	40,674	2.22E-08	0.025	0.004
NtAbs	NtAbs	immune	25,118	5.69E-07	-0.026	0.005
PT.inr	International normalized ratio	blood	34,919	1.31E-06	-0.027	0.006
IGRE	Immature granulocytes/100 leukocytes in Blood	immune	44,827	7.86E-06	-0.020	0.005
MONORE	Monocytes/100 leukocytes in Blood by Automated count	immune	46,569	9.13E-06	0.019	0.004
RDW	Erythrocyte distribution width [Ratio] by Automated count	blood	64,874	2.21E-05	-0.016	0.004
GluWB	Glucose [Mass/volume] in Blood	blood	13,848	2.49E-05	-0.034	0.008

Table 4. Significant associations from the LabWAS of LDL PGS in individuals of European ancestry in VUMC.

Short Name	Full Name	Group	N	p-value	beta	SE
LDL.C	Cholesterol in LDL [Mass/volume] in Serum or Plasma by calculation	metabolic	28,417	2.23E-308	0.236	0.006
Chol	Cholesterol [Mass/volume] in Serum or Plasma	metabolic	30,388	1.30E-282	0.197	0.005
DirLDL	Cholesterol in LDL [Mass/volume] in Serum or Plasma by Direct assay	metabolic	4,859	3.79E-44	0.185	0.013
NonHDL	Cholesterol non HDL [Mass/volume] in Serum or Plasma	metabolic	3,252	1.78E-31	0.192	0.016
TC.HDL	Cholesterol.total/Cholesterol in HDL [Molar ratio] in Serum or Plasma	metabolic	3,676	5.27E-17	0.130	0.015
MCH	MCH [Entitic mass] by Automated count	blood	64,869	5.68E-08	-0.019	0.004
TProt	Tau protein [Presence] in Body fluid	blood	44,329	2.18E-06	0.020	0.004
TProSe	Protein serum/plasma	metabolic	32,127	3.00E-06	0.024	0.005
Trigs	Triglyceride [Mass/volume] in Serum or Plasma	metabolic	30,597	4.47E-06	0.026	0.006
MCHC	MCHC [Mass/volume] by Automated count	blood	54,231	1.50E-05	-0.017	0.004

Table 5. Significant associations from the LabWAS of TG PGS in individuals of European ancestry in VUMC.

Short Name	Full Name	Group	N	p-value	beta	SE
Trigs	Triglyceride [Mass/volume] in Serum or Plasma	metabolic	30,597	2.23E-308	0.287	0.005
HDL.C	Cholesterol in HDL [Mass/volume] in Serum or Plasma	metabolic	29,497	4.83E-148	-0.139	0.005
WBC	Leukocytes [#/volume] in Blood by Automated count	immune	64,836	3.90E-30	0.040	0.003
TC.HDL	Cholesterol.total/Cholesterol in HDL [Molar ratio] in Serum or Plasma	metabolic	3,676	2.95E-28	0.166	0.015
CO2	Carbon dioxide serum/plasma	blood	62,304	2.57E-24	-0.035	0.003
Gluc	Glucose lab	metabolic	62,341	1.20E-22	0.036	0.004
LipIdx	Lipemic index of Serum or Plasma	metabolic	4,500	1.57E-18	0.122	0.014
AN.GAP	Anion gap serum/plasma	metabolic	62,164	2.03E-17	0.030	0.004
TProSe	Protein serum/plasma	metabolic	32,127	2.61E-16	0.041	0.005
TProt	Tau protein [Presence] in Body fluid	blood	44,329	4.25E-16	0.034	0.004
Chol	Cholesterol [Mass/volume] in Serum or Plasma	metabolic	30,388	1.25E-14	0.042	0.006
IGRE	Immature granulocytes/100 leukocytes in Blood	immune	44,827	1.99E-14	0.035	0.005
MCV	MCV [Entitic volume] by Automated count	blood	64,844	9.16E-13	-0.025	0.004
LymAbs	Lymphocytes [#/volume] in Blood by Automated count	immune	25,135	2.01E-11	0.034	0.005
D25OHT	Calcitriol [Mass/volume] in Serum or Plasma	metabolic	18,315	1.07E-10	-0.045	0.007
MONORE	Monocytes/100 leukocytes in Blood by Automated count	immune	46,569	5.21E-10	-0.026	0.004
NtAbs	NtAbs	immune	25,118	1.87E-09	0.032	0.005
HgbA1C	Hemoglobin A1c (Glycated) Hemoglobin A1c (Glycated)	metabolic	21,558	5.69E-09	0.038	0.007
UricA	Urate [Mass/volume] in Serum or Plasma	urinary	9,825	6.58E-09	0.058	0.010
C4Quan	Complement C4 [Mass/volume] in Serum or Plasma	immune	3,667	1.03E-08	0.092	0.016
Monocy	Monocytes [#/volume] in Blood by Manual count	immune	26,166	6.76E-08	-0.029	0.005
AlbCre	Microalbumin/Creatinine [Mass Ratio] in 24 hour Urine	liver	2,935	9.13E-08	0.098	0.018
MCH	MCH [Entitic mass] by Automated count	blood	64,869	9.75E-08	-0.019	0.004
GluBed	Glucose [Mass/volume] in Blood by Automated test strip	metabolic	25,193	2.99E-07	0.031	0.006
UpH	pH of Urine by Test strip	urinary	40,674	7.66E-07	-0.022	0.004
NonHDL	Cholesterol non HDL [Mass/volume] in Serum or Plasma	metabolic	3,252	1.18E-06	0.079	0.016
AlbCnc	Albumin [Presence] in Urine	urinary	5,255	2.99E-05	0.057	0.014

LabWAS of Lipid Polygenic Scores in African Ancestry Individuals in VUMC

In the African ancestry group, HDL PGS significantly associated with increased HDL (p-value = 1.38×10^{-74}, beta = 0.23), decreased triglycerides (p-value = 6.72×10^{-10}, beta = -0.08), and increased total cholesterol (p-value = 4.81×10^{-9}, beta = 0.08) (Figure 6a, Table 8). LDL PGS associated with LDL cholesterol (p-value = 5.71×10^{-63}, beta = 0.24) and increased total cholesterol (p-value = 1.63×10^{-53}, beta=0.21) (Figure 6b, Table 8). TG PGS showed significant associations with increased triglycerides (p-value = 1.66×10^{-53}, beta = 0.19), decreased HDL cholesterol (p-value = 6.08×10^{-11}, beta = -0.08), and increased glucose (p-value = 2.33×10^{-5}, beta = 0.04) (Figure 6c, Table 6).

Figure 6. Lipid PGS LabWAS in individuals of African ancestry in VUMC. (A) HDL PGS LabWAS, (B) LDL PGS LabWAS, (C) Triglycerides PGS LabWAS. The red line indicates the Bonferroni threshold for statistical significance and the blue line indicates a *p* value of 0.05. Upward triangles indicate that the PGS is associated with increased levels of the lab, while downward triangles indicate an association with reduced levels of the lab.

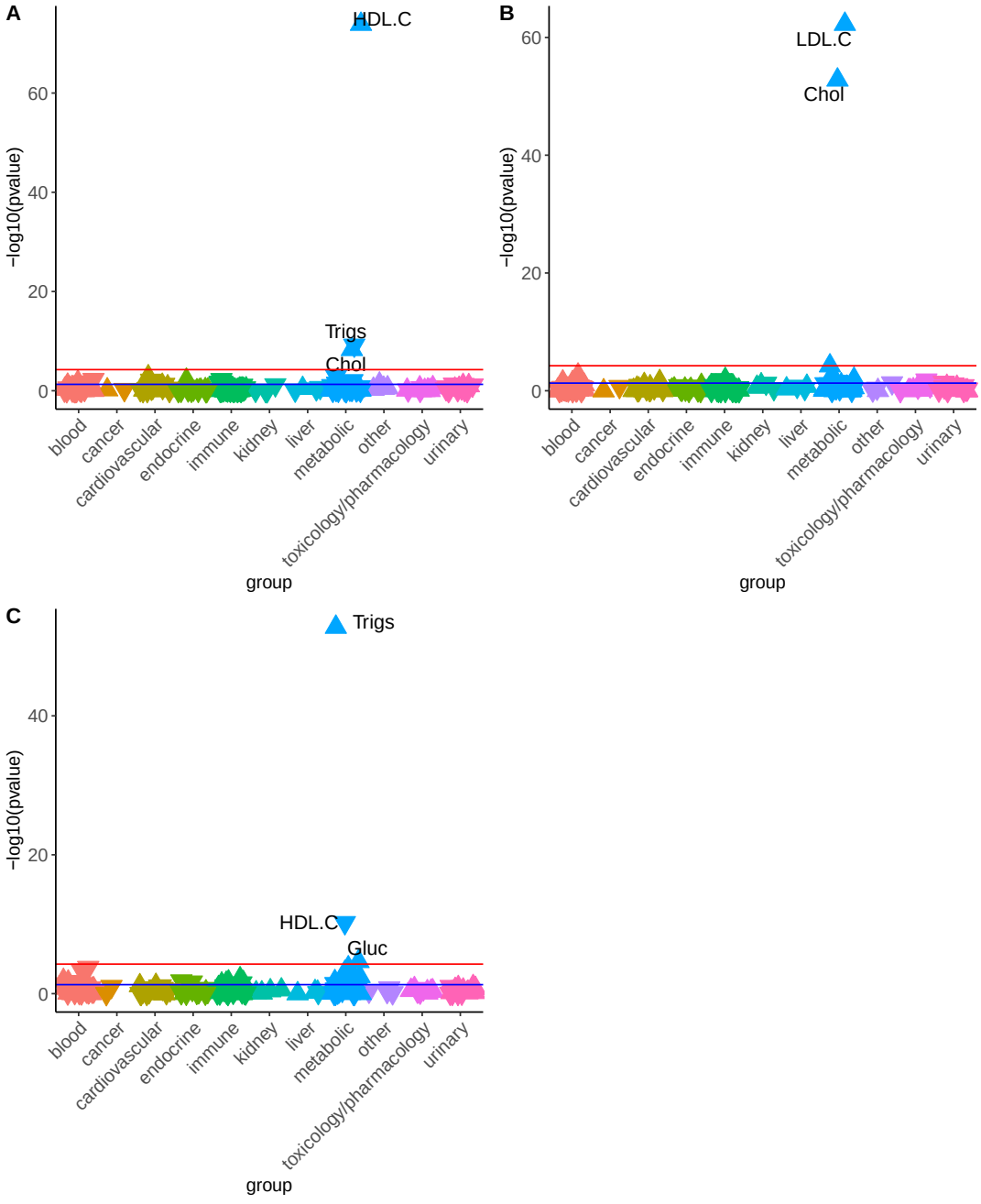


Table 6. Significant associations from the LabWAS of lipid and CAD PGS in individuals of African ancestry in VUMC.

PGS	Short Name	Full Name	Group	N	p-value	beta	SE
HDL	HDL.C	Cholesterol in HDL [Mass/volume] in Serum or Plasma	metabolic	5,607	1.38E-74	0.232	0.012
	Trigs	Triglyceride [Mass/volume] in Serum or Plasma	metabolic	5,728	6.72E-10	-0.08	0.013
	Chol	Cholesterol [Mass/volume] in Serum or Plasma	metabolic	5,746	4.81E-09	0.077	0.013
LDL	LDL.C	Cholesterol in LDL [Mass/volume] in Serum or Plasma by calculation	metabolic	5,503	5.71E-63	0.243	0.014
	Chol	Cholesterol [Mass/volume] in Serum or Plasma	metabolic	5,746	1.63E-53	0.212	0.014
TG	Trigs	Triglyceride [Mass/volume] in Serum or Plasma	metabolic	5,728	1.66E-53	0.19	0.012
	HDL.C	Cholesterol in HDL [Mass/volume] in Serum or Plasma	metabolic	5,607	6.08E-11	-0.08	0.012
	Gluc	Glucose lab	metabolic	11,174	2.33E-05	0.038	0.009
CAD covaried for CAD diagnosis	LDL_pre med	LDL pre-medication	metabolic	4,977	3.92E-05	0.061	0.004

LabWAS of a Polygenic Score for Coronary Artery Disease in VUMC

We next sought to recapitulate the risk biomarker profile for CAD through a LabWAS of a CAD PGS. The CAD PGS reproduced associations, in the direction of risk, with canonical risk factors for CAD (Figure 7a, Table 7) in the European ancestry population, including decreased HDL (p-value = 6.20×10^{-39}, beta = -0.07), increased TG (p-value = 3.98×10^{-25}, beta = 0.06), increased blood glucose (p-value = 1.18×10^{-21}, beta = 0.04) and glycated hemoglobin (p-value = 2.36×10^{-12}, beta = 0.05), and bedside glucose (p-value = 1.10×10^{-6}, beta = 0.03). The CAD PGS also associated with other known biomarkers of cardiovascular health such as increased troponin-I (p-value = 7.20×10^{-9}, beta = 0.04) and brain natriuretic peptide (p-value = 2.12×10^{-7}, beta = 0.05). CAD PGS associated with six blood composition markers, red blood cell distribution width (p-value = 1.60×10^{-11}, beta = 0.03), mean corpuscular hemoglobin (p-value = 6.73×10^{-10}, beta = -0.02), mean corpuscular volume (p-value = 1.17×10^{-9}, beta = -0.02), carbon dioxide (p-value = 3.36×10^{-9}, beta = -0.02), red blood cell sedimentation rate (p-value = 2.10×10^{-7}, beta = 0.05), and international normalized rate (p-value = 1.96×10^{-5}, beta = 0.03). Finally, CAD PGS associated with white blood cell count (p-value = 8.75×10^{-11}, beta = 0.02), creatinine (p-value = 2.13×10^{-6}, beta = 0.02), and blood urea nitrogen (p-value = 1.09×10^{-5}, beta = 0.02).

Notably, the CAD PGS was not initially associated with LDL values (p-value = 0.13, beta = 0.008). The lack of association, however, was attributable to lipid altering medication use and a significant association between the CAD PGS and LDL levels was detected when we restricted to pre-medication values (p = 6.19×10^{-9}, beta = 0.04).

To determine which biomarkers were explained by the clinical presence of CAD as opposed to the genetic risk for CAD, we adjusted the LabWAS of CAD PGS for the coronary

atherosclerosis phecode (411) (Figure 7b, Table 8). Four canonical biomarkers of CAD risk remained associated with CAD PGS including TG (p-value = 2.88 x 10^{-14}, beta = 0.05), pre-medication LDL (p-value = 2.40 x 10^{-13} beta = 0.05), HDL (p-value = 2.55 x 10^{-13}, beta = -0.04), LDL-C (p-value = 8.48 x 10^{-11}, beta = 0.04), blood glucose (p-value = 2.55 x 10^{-9}, beta = 0.02), total cholesterol (p-value = 4.16 x 10^{-9}, beta = 0.03), and glycated hemoglobin (p-value = 3.16 x 10^{-6}, beta = 0.03). The CAD PGS also remained associated with one immune marker, white blood cell count (p-value = 6.44 x 10^{-6}, beta = 0.02), and two other blood biomarkers, mean corpuscular volume (p-value = 3.23 x 10^{-7}, beta = -0.02) and mean corpuscular hemoglobin (p-value = 4.18 x 10^{-6}, beta = -0.02).

None of the associations in the initial LabWAS of CAD PGS among African ancestry individuals reached phenome-wide significance, however, three of the top four associations were canonical CAD risk factors including increased glycated hemoglobin A1c (p-value=9.56 x 10^{-4}, beta=0.04), increased glucose (p-value=0.002, beta=0.03), and increased LDL cholesterol (p-value=0.003, beta=0.04) (Figure 7c). When the LDL levels were restricted to pre-medication values, the top association with CAD PGS was pre-medication LDL (p-value = 8.50 x 10^{-5}, beta=0.06), however this association did not pass multiple testing correction. After controlling the analysis for CAD diagnosis, the association between CAD PGS and pre-medication LDL surpassed the Bonferroni correction for phenome-wide significance (p-value=3.92 x 10^{-5}, beta=0.06) (Figure 7d, Table 6).

Lastly, we ran a LabWAS of CAD diagnosis (i.e., using CAD cases/control status as the predictor variable) after adjusting for sex and median age across the EHR, which revealed the medical comorbidity pattern of CAD. CAD diagnosis was significantly associated with 136 out of

734 labs in our sample (Figure 8), including 34 immune, 32 blood, 24 metabolic, 17 cardiovascular, 8 urinary, 5 toxicology/pharmacology, 4 endocrine, 3 kidney, 3 liver, 1 cancer, and 5 other markers.

Figure 7. LabWAS of CAD PGS in VUMC. (A) European ancestry, (B) European ancestry controlling for CAD diagnosis, (C) African ancestry and (D) African ancestry controlling for CAD diagnosis. The red lines indicate the Bonferroni threshold for statistical significance and the blue line indicates a *p* value of 0.05. Upward triangles indicate that the PGS is associated with increased levels of the lab, while downward triangles indicate an association with reduced levels of the lab.

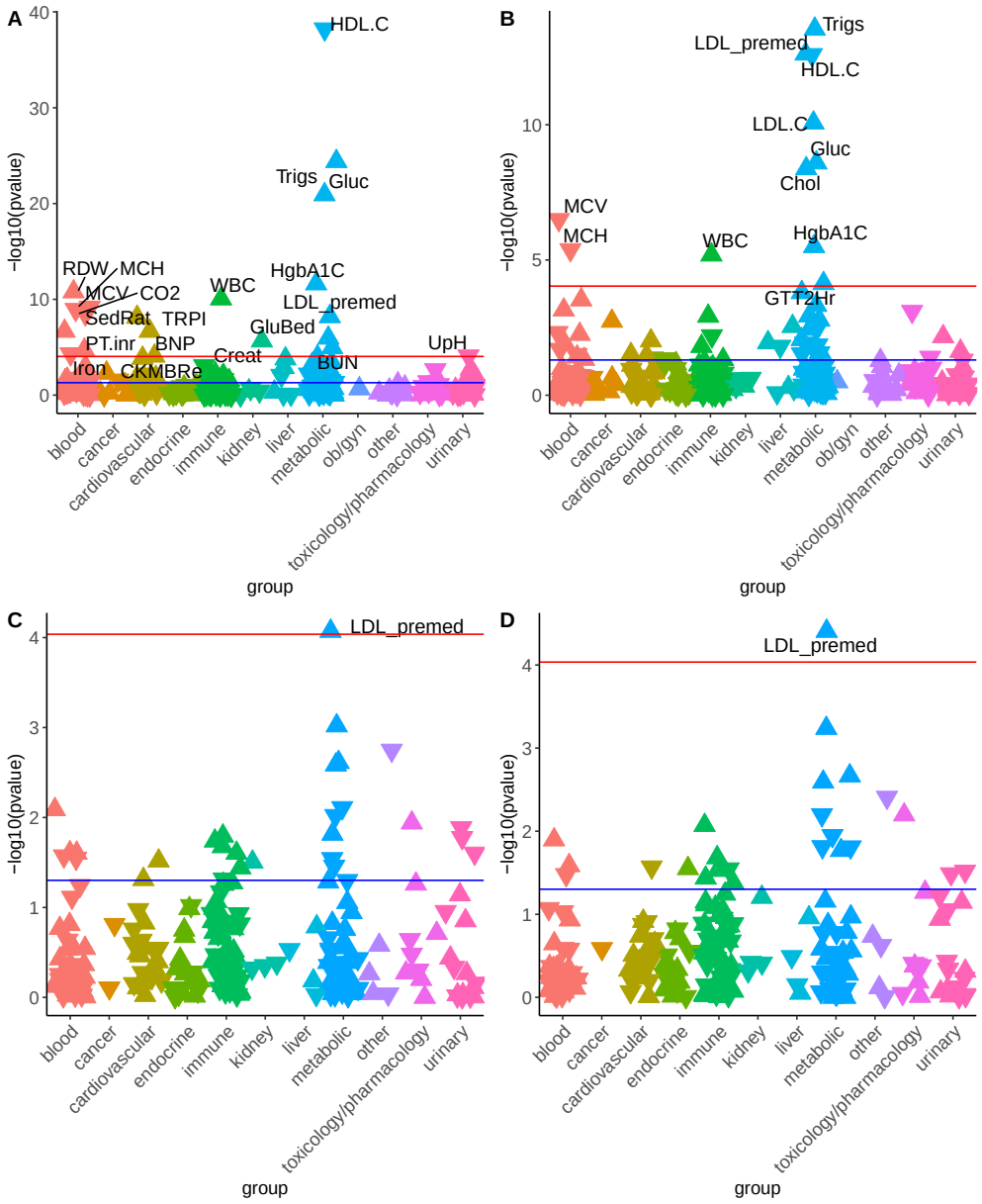


Figure 8. LabWAS of CAD diagnosis.

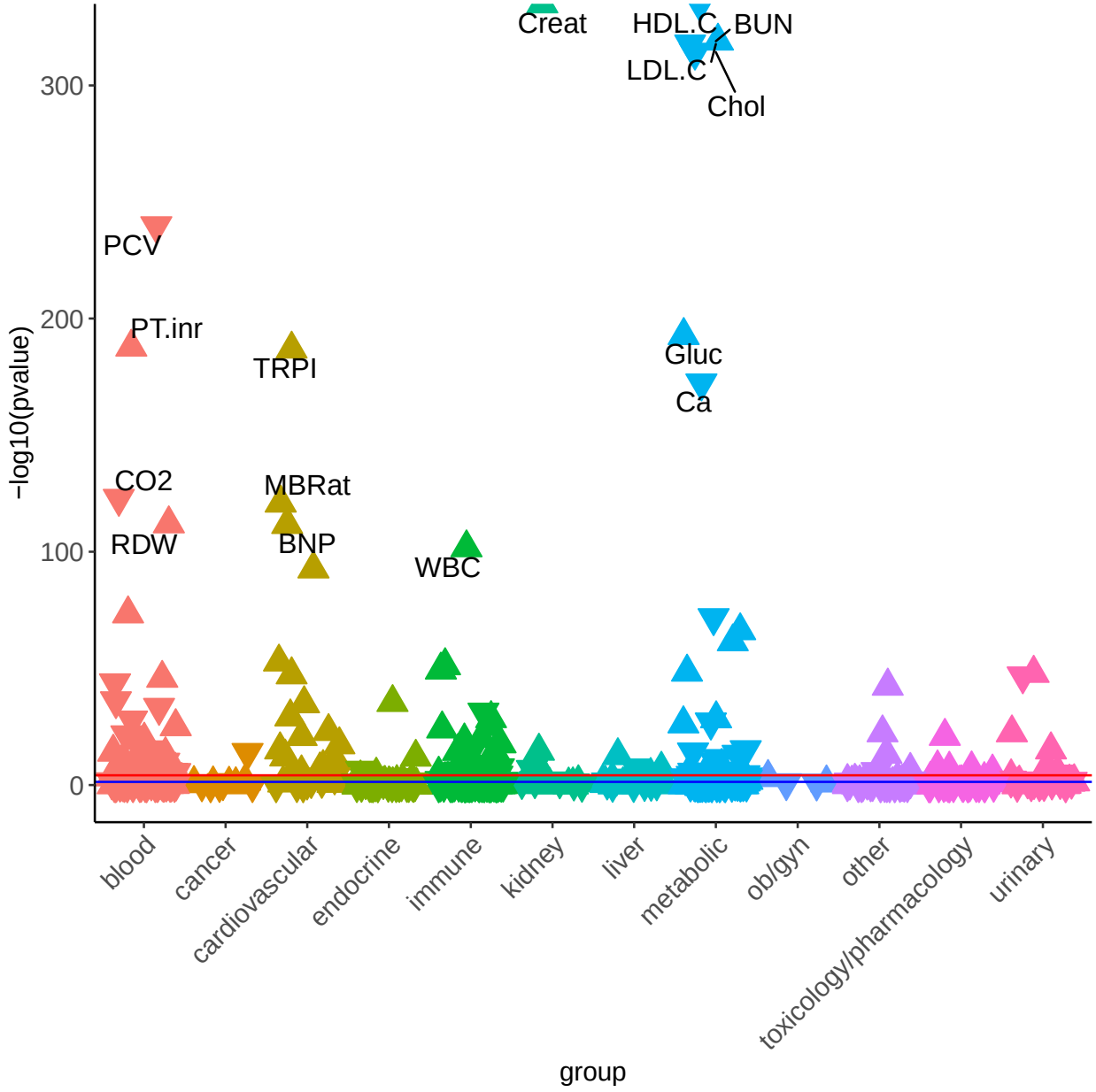


Table 7. Significant associations from the LabWAS of CAD PGS in individuals of European ancestry in VUMC.

Short Name	Full Name	Group	N	pvalue	beta	SE
HDL.C	Cholesterol in HDL [Mass/volume] in Serum or Plasma	metabolic	29,497	6.20E-39	-0.071	0.005
Trigs	Triglyceride [Mass/volume] in Serum or Plasma	metabolic	30,597	3.98E-25	0.059	0.006
Gluc	Glucose lab	metabolic	62,341	1.18E-21	0.036	0.004
HgbA1C	Hemoglobin A1c (Glycated)	metabolic	21,558	2.36E-12	0.046	0.007
RDW	Erythrocyte distribution width [Ratio] by Automated count	blood	64,874	1.60E-11	0.026	0.004
WBC	Leukocytes [#/volume] in Blood by Automated count	immune	64,836	8.75E-11	0.023	0.004
MCH	MCH [Entitic mass] by Automated count	blood	64,869	6.73E-10	-0.022	0.004
MCV	MCV [Entitic volume] by Automated count	blood	64,844	1.17E-09	-0.022	0.004
CO2	Carbon dioxide serum/plasma	blood	62,304	3.36E-09	-0.021	0.003
LDL_premed	LDL pre-medication	metabolic	21,756	6.19E-09	0.039	0.007
TRPI	Troponin I.cardiac [Mass/volume] in Serum or Plasma	cardiovascular	10,443	7.20E-09	0.045	0.008
SedRat	Erythrocyte sedimentation rate by Westergren method	blood	10,541	2.10E-07	0.052	0.010
BNP	Natriuretic peptide B [Mass/volume] in Serum or Plasma	cardiovascular	11,374	2.12E-07	0.046	0.009
GluBed	Glucose [Mass/volume] in Blood by Automated test strip	metabolic	25,193	1.10E-06	0.030	0.006
Creat	Creatinine [Mass/volume] in Blood	kidney	62,909	2.13E-06	0.018	0.004
BUN	Urea nitrogen serum/plasma	metabolic	62,403	1.09E-05	0.018	0.004
PT.inr	International normalized ratio	blood	34,919	1.96E-05	0.024	0.006

Table 8. Significant associations from the LabWAS of CAD PGS covaried for CAD diagnosis in individuals of European ancestry in VUMC.

Short Name	Long Name	Group	N	pvalue	beta	SE
Trigs	Triglyceride [Mass/volume] in Serum or Plasma	metabolic	30,523	2.88E-14	0.047	0.006
LDL_premed	LDL pre-medication	metabolic	21,689	2.40E-13	0.051	0.007
HDL.C	Cholesterol in HDL [Mass/volume] in Serum or Plasma	metabolic	29,430	2.55E-13	-0.042	0.006
LDL.C	Cholesterol in LDL [Mass/volume] in Serum or Plasma by calculation	metabolic	28,350	8.48E-11	0.040	0.006
Gluc	Glucose lab	metabolic	62,217	2.55E-09	0.023	0.004
Chol	Cholesterol [Mass/volume] in Serum or Plasma	metabolic	30,314	4.16E-09	0.034	0.006
MCV	MCV [Entitic volume] by Automated count	blood	64,687	3.23E-07	-0.019	0.004
HgbA1C	Hemoglobin A1c (Glycated)	metabolic	21,513	3.16E-06	0.034	0.007
MCH	MCH [Entitic mass] by Automated count	blood	64,712	4.18E-06	-0.017	0.004
WBC	Leukocytes [#/volume] in Blood by Automated count	immune	64,679	6.44E-06	0.017	0.004

Replication in Mass General Brigham Biobank

In the MGB, there were 21,499 individuals of European descent with genetic data available with recorded lab data. Slightly more than half of the sample was female (51.5%) and the average age was 56.1 years. The MGB patients contained 1,094 CAD cases and 20,405 CAD controls.

In MGB, the HDL PGS most strongly associated with HDL cholesterol (p-value < 2.23×10^{-308}, beta = 0.33), followed by decreased triglycerides (p-value = 2.77×10^{-109}, beta = -0.17), increased total cholesterol (p-value = 4.96×10^{-31}, beta = 0.09), and decreased very low density lipoprotein (p-value = 2.62×10^{-29}, beta = -0.14). HDL_{PGS} also associated with decreased values of glucose (p-value = 3.33×10^{-27}, beta = -0.07), hemoglobin A1c (p-value = 3.64×10^{-18}, beta = -0.07), and mean glucose value (p-value = 4.10×10^{-17}, beta = -0.07). Additional associations with HDL_{PGS} included cardiac relative risk (p-value = 5.72×10^{-17}, beta = -0.20), alanine aminotransferase (p-value = 2.45×10^{-10}, beta = -0.04), white blood cell count (p-value = 1.03×10^{-9}, beta = -0.04), mean corpuscular volume (p-value = 6.51×10^{-8}, beta = 0.03), non-HDL cholesterol (p-value = 1.41×10^{-7}, beta = -0.06), red blood cell distribution width (p-value = 2.60×10^{-7}, beta = -0.03), neutrophils (p-value = 2.95×10^{-7}, beta = -0.03), urate (p-value = 1.79×10^{-6}, beta = -0.05), and alkaline phosphatase (p-value = 2.01×10^{-6}, beta = -0.03) (Fig. 9a).

The LDL PGS associated with four metabolic labs including LDL-C (p-value = 1.78×10^{-158}, beta = 0.24), total cholesterol (p-value = 2.37×10^{-158}, beta = 0.20), calculated LDL cholesterol (p-value = 1.28×10^{-81}, beta = 0.23), and non-HDL cholesterol (p-value = 2.90×10^{-68}, beta = 0.19). The LDL PGS also associated with complement C4 (p-value = 1.85×10^{-5}, beta = 0.09), red

blood cell sedimentation rate (p-value = 2.60 x 10^{-5}, beta = 0.04), and increased cardiac relative risk (p-value = 3.80 x 10^{-5}, beta = 0.10) (Fig. 9b).

The TG PGS associated with twelve metabolic labs, including increased measured triglycerides (p-value < 2.23 x 10^{-308}, beta = 0.32), followed by increased very low density lipoprotein (p-value = 8.90 x 10^{-129}, beta = 0.30), decreased HDL (p-value = 1.33 x 10^{-123}, beta = -0.17), increased non-HDL cholesterol (p-value = 8.70 x 10^{-28}, beta = 0.12), increased glucose (p-value = 4.56 x 10^{-14}, beta = 0.05), average glucose (p-value = 4.16 x 10^{-10}, beta = 0.05), total cholesterol (p-value = 1.58 x 10^{-9}, beta = 0.05), anion gap (p-value = 1.52 x 10^{-7}, beta = 0.03), total protein (p-value = 4.63 x 10^{-7}, beta = 0.03), globulin in serum (p-value = 8.80 x 10^{-6}, beta = 0.03), aspartate aminotransferase (p-value = 1.26 x 10^{-5}, beta = 0.03), and sodium (p-value = 1.27 x 10^{-5}, beta = -0.03). TG_{PGS} also associated with seven immune labs, white blood cell count (p-value = 3.89 x 10^{-17}, beta = 0.05), lymphocytes (p-value = 7.86 x 10^{-11}, beta = 0.04), complement C4 (p-value = 1.58 x 10^{-9}, beta = 0.13), automated lymphocyte count (p-value = 2.14 x 10^{-9}, beta = 0.09), neutrophils (p-value = 3.09 x 10^{-7}, beta = 0.05), automated neutrophil count (p-value = 5.13 x 10^{-7}, beta = 0.03), and monocytes (p-value = 3.38 x 10^{-6}, beta = 0.05). Ten additional labs significantly associated with TG PGS, including increased cardiac relative risk (p-value = 5.49 x 10^{-15}, beta = 0.19), mean corpuscular volume (p-value = 3.02 x 10^{-14}, beta = -0.05), glycated hemoglobin A1c (p-value = 5.00 x 10^{-11}, beta = 0.05), urinary pH (p-value = 9.58 x 10^{-10}, beta = -0.04), red blood cell sedimentation rate (p-value = 2.22 x 10^{-8}, beta = 0.05), alanine aminotransferase (p-value = 3.88 x 10^{-8}, beta = 0.04), alkaline phosphatase (p-value = 3.17 x 10^{-7}, beta = 0.03), blood carbon dioxide (p-value = 5.63 x 10^{-7}, beta = -0.03), mean

corpuscular hemoglobin (p-value = 1.49 x 10^{-6}, beta = -0.03), and urate (p-value = 1.62 x 10^{-6}, beta = 0.05) (Fig. 9c).

Finally, the CAD PGS associated with several known CAD risk factors, including decreased HDL-C (p-value = 1.56 x 10^{-21}, beta = -0.07), increased glucose (p-value = 9.91 x 10^{-15}, beta = 0.05), increased glycated hemoglobin A1c (p-value = 4.44 x 10^{-14}, beta = 0.06), mean glucose (p-value = 1.75 x 10^{-12}, beta = 0.06), and increased triglycerides (p-value = 2.09 x 10^{-12}, beta = 0.05). The CAD PGS also associated with increased red blood cell distribution width (p-value = 2.42 x 10^{-14}, beta = 0.05), increased red blood cell sedimentation rate (p-value = 4.11 x 10^{-9}, beta = 0.05), increased alanine aminotransferase (p-value = 2.59 x 10^{-8}, beta = 0.04), decreased hemoglobin (p-value = 1.45 x 10^{-6}, beta = -0.03), increased alkaline phosphatase (p-value = 2.26 x 10^{-6}, beta = 0.03), increased white blood cell count (p-value = 6.77 x 10^{-6}, beta = 0.03), decreased albumin (p-value = 1.06 x 10^{-5}, beta = -0.03), increased globulin (p-value = 1.29 x 10^{-5}, beta = 0.03), decreased iron (p-value = 3.36 x 10^{-5}, beta = -0.04), and decreased hematocrit (p-value = 3.77 x 10^{-5}, beta = -0.03) (Fig. 10a).

After adjusting for CAD diagnosis, CAD PGS remained associated with several heart disease risk factors including decreased HDL-C (p-value = 8.23 x 10^{-16}, beta = -0.06), increased glucose (p-value = 6.80 x 10^{-11}, beta = 0.04), increased hemoglobin A1c (p-value = 2.08 x 10^{-10}, beta = 0.05), increased mean glucose (p-value = 2.29 x 10^{-9}, beta = 0.05), and increased triglycerides (p-value = 3.48 x 10^{-9}, beta = 0.05). Additionally, associations with red blood cell distribution width (p-value = 1.40 x 10^{-12}, beta = 0.04), alanine aminotransferase (p-value = 4.01 x 10^{-8}, beta = 0.04), red blood cell sedimentation rate (p-value = 5.50 x 10^{-8}, beta = 0.05),

alkaline phosphatase (p-value = 5.44×10^{-6}, beta = 0.03), serum globulin (p-value = 4.51×10^{-5}, beta = 0.03), and white blood cell count (p-value = 4.67×10^{-5}, beta = 0.03) remained (Fig. 10b). In MGB, the CAD PGS was not associated with levels of LDL-C (p-value = 0.06, beta = -0.03) and we were unable to investigate the effects of cholesterol lowering medications on the association.

Figure 9. Replication of Lipid PGS LabWAS in MGB. (A) HDL PGS, (B) LDL PGS, and (C) TG PGS. The red lines indicate the Bonferroni threshold for statistical significance and the blue line indicates a *p* value of 0.05. Upward triangles indicate that the PGS is associated with increased levels of the lab, while downward triangles indicate an association with reduced levels of the lab.

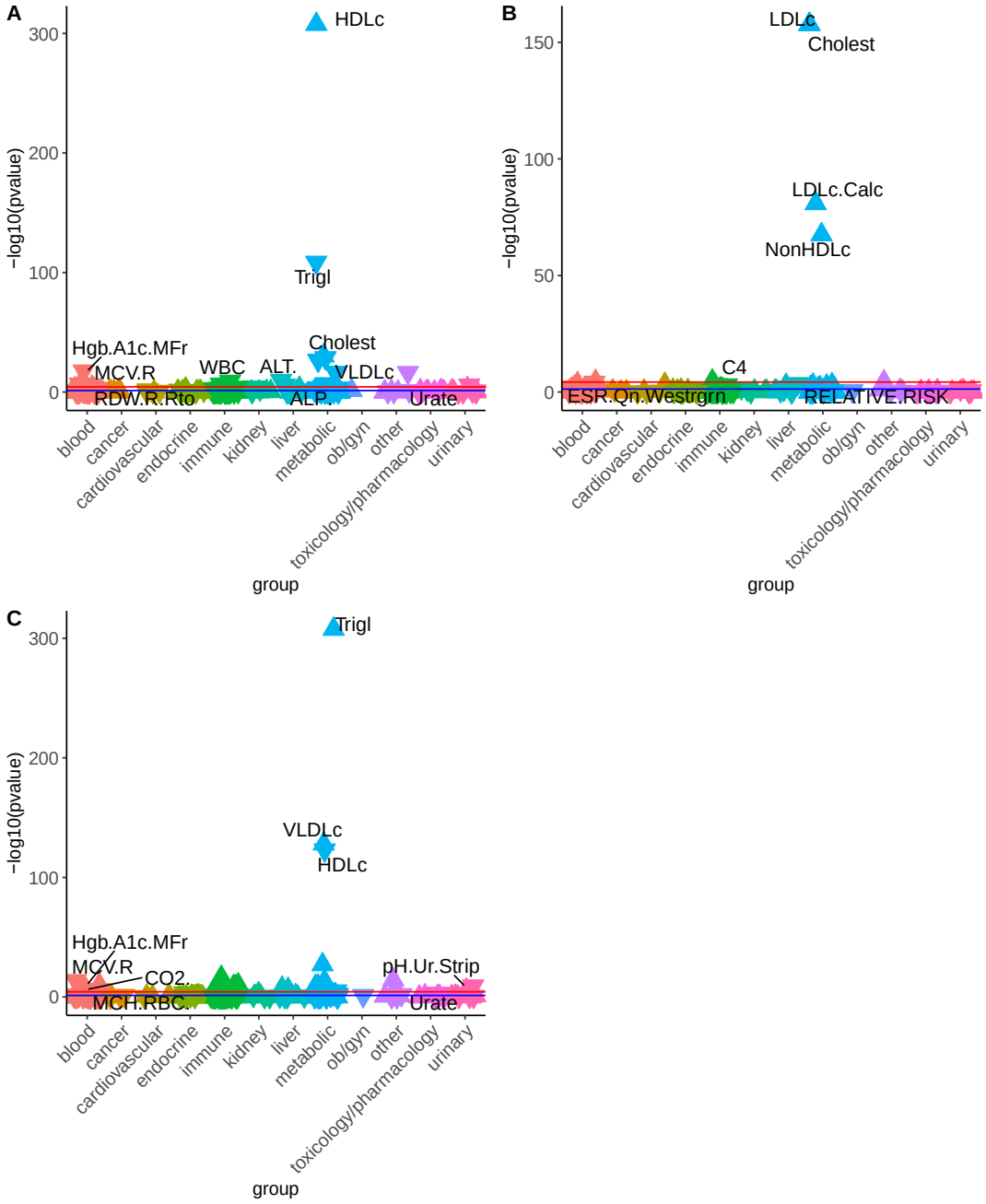
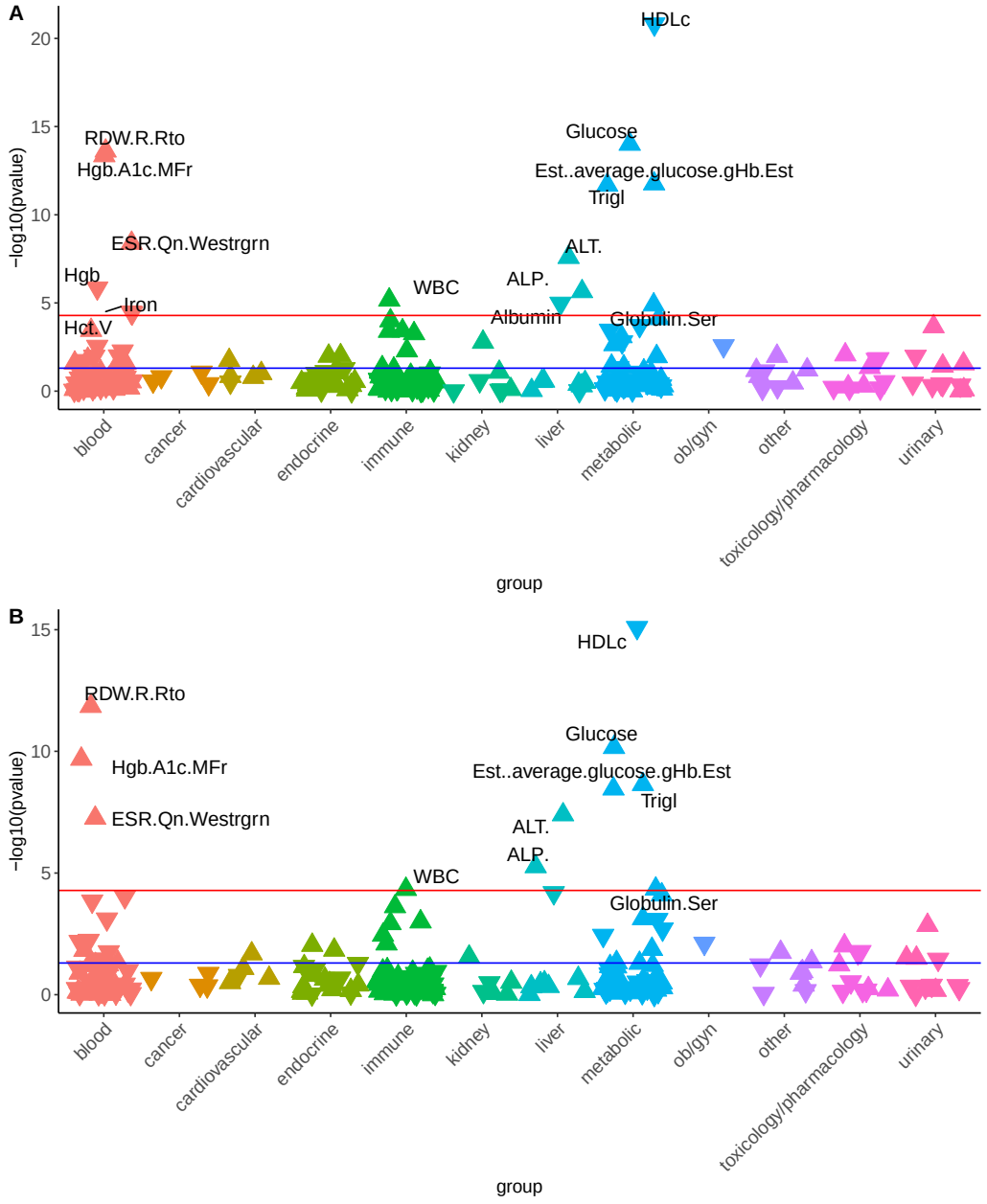


Figure 10. Replication of CAD PGS LabWAS in MGB. (A) CAD PGS (B) CAD PGS covaried for CAD diagnosis. The red lines indicate the Bonferroni threshold for statistical significance and the blue line indicates a *p* value of 0.05. Upward triangles indicate that the PGS is associated with increased levels of the lab, while downward triangles indicate an association with reduced levels of the lab.



Discussion

We show that EHR-derived lipids values are genetically similar to those in population-based studies, and that PGS for lipids robustly associate with their respective lab in a LabWAS. Additionally, LabWAS revealed that PGS for CAD associated with known lipid biomarkers, even in individuals without a history of CAD, and with potentially novel immune biomarkers. The results of our study add to a growing body of evidence indicating that lab values from EHRs with linked genetic data can be mined at scale to identify biomarkers for complex disease[66–70]. Our proof-of-principle analyses focused on lipids and CAD in 94,747 genotyped VUMC patients and revealed that EHR lipid values were genetically comparable to those measured in samples ascertained for research. We describe two proof-of-concept studies that demonstrate the power of our proposed discovery paradigm. First, we show that PGS for lipids (HDL, LDL, and triglycerides) associate robustly to their referent lipid across ancestries. Moreover, the CAD PGS recapitulated associations with known biomarkers in two biobanks. Importantly, the association between CAD PGS and canonical risk factors was significant even among those who did not have a CAD diagnosis. In analyses in MGB, several of the associations with CAD PGS replicated, helping to validate our approach to analyzing EHR laboratory data.

Furthermore, we show that treatments (in this example, lipid-altering medications) can influence the detection of risk biomarkers at the genetic level. For example, we found that the genetic correlation between LDL measurements in VUMC and MVP increased considerably when we restricted to pre-medication LDL measurements and controlled for CAD or diabetes diagnosis. Additionally, the CAD PGS was strongly associated with pre-medication median LDL values, but was not associated with combined pre- and post-medication median LDL values.

Though the results and approach presented provide an exciting path forward for genetic analysis of EHR-lab data, important limitations should be acknowledged. First, our analyses yielded more associations in patients of European ancestry compared to patients of African ancestry. This is likely to due to decreased power from both the discovery GWASs and the target sample. VUMC has considerably fewer patients of African ancestry than European ancestry, impacting our statistical power to find associations. The polygenic scores of lipids, which were trained on trans-ancestry GWAS summary statistics including individuals of African descent, strongly associated with the referent lipid in the African ancestry sample with effect estimates similar to those found in the European sample. However, the CAD polygenic score, which was trained on a trans-ancestry GWAS that did not include African ancestry samples yielded far fewer significant associations. These results highlight the critical importance of diversity in GWAS as the downstream applications of such studies are dramatically impacted by representation.

Second, high-throughput analysis of 939 lab traits in our LabWAS required us to prioritize statistical model performance over coefficient interpretability. In our primary analysis, we transformed lab values to fit the normal distribution to improve the performance of the linear regression models[90]. We applied the rank-based inverse normal quantile transformation to all labs, which ensured trait normality by replacing the value of each observation with its quantile from the standard normal distribution. The inverse normal quantile transformation thus preserved the rank ordering of observations, but not the values themselves, and model coefficients therefore are uninterpretable on the original scale. For example, based on our LabWAS results, we are unable to report the change in LDL levels in mg/dL per SD increase in

the CAD_{PGS}. Multiple testing correction was another statistical challenge inherent to the high-throughput analysis of lab traits. We used the Bonferroni threshold for statistical significance, but this threshold is likely to be overly strict because it ignores the correlation between lab tests.

In conclusion, we propose that PGS for complex disease can be used to discover genetically related biomarkers of disease by mining quantitative physiological measurements collected during routine clinical testing, but caution that mindful interpretation of correlational results is paramount to progress. We demonstrate the robustness of this discovery paradigm in a proof of principal analysis focused on CAD. As EHR resources grow in size, standardized analysis pipelines will be necessary to compare results across samples. LabWAS provides a starting point for consistent analysis of lab results stored in various EHR systems. Furthermore, we demonstrated that EHR-derived lipids are similar to measurements ascertained in traditional cohort studies, providing additional rationale for analyses of EHR labs[91]. QualityLab and LabWAS are scalable programs that can be used to confirm clinical paradigms and discover new genetic and environmental relationships between biomarkers and complex traits. We propose that future studies will leverage this discovery paradigm for analysis of rare or understudied complex traits with no known biomarker associations (e.g., psychiatric disorders).

CHAPTER III

INVESTIGATING THE ASSOCIATION BETWEEN DEPRESSION GENETICS AND WHITE BLOOD CELL COUNT ACROSS THE PSYCHEMERGE NETWORK*[2]

Introduction

After validating the LabWAS approach, we next applied the method to depression polygenic scores (PGS). While independent biobanks can be used to discover associations, combining multiple health record systems through consortia can validate those discoveries in broader populations. The PsycheMERGE Network consists of investigators from institutions across the United States with the common goal of using EHRs and biobanks to advance the identification, biology, and treatment of psychiatric disorders[92]. Here, we investigate the effect of polygenic risk for depression on clinically measured lab values leveraging data from healthcare systems participating in the PsycheMERGE Network. Four biobanks from the PsycheMERGE Network were included, Vanderbilt University Medical Center (VUMC), Mass General Brigham (MGB), Mount Sinai Icahn School of Medicine (MSSM), and the Million Veterans Program (MVP), resulting in a total sample size of 382,452 individuals.

Methods

Sample Description

[2] *Adapted with permission from Sealock JM et al., JAMA Psych 2021

Electronic health record and genotype information were extracted for individuals of European descent across four biobanks in the PsycheMERGE Network: Vanderbilt University Medical Center (VUMC), Massachusetts General Brigham (MGB), Million Veteran Program (MVP), and Mount Sinai Icahn School of Medicine (MSSM).

Vanderbilt University Medical Center (VUMC) EHR system and biobank are described in Chapter 2. The VUMC Institutional Review Board oversees BioVU and approved this project (IRB#172020). In VUMC, primary analyses were conducted in individuals of European and African ancestry. Lab results were extracted from the EHRs of 70,704 individuals of primarily European ancestry and 12,384 individuals of primarily African ancestry and cleaned using QualityLab.

The Million Veteran Program is an observational cohort study and mega-biobank in the Department of Veterans Affairs (VA)[93]. Participants are active users of the Veterans Health Administration and provide a blood sample, responses to questionnaires and consent to allow access to clinical data from the VA electronic health records[93]. The MVP v3.0 data release used in this study includes genotyping data from 455,789 individuals; DNA was extracted from whole blood (which was collected during enrollment to the MVP) and genotyping was performed with the MVP 1.0 Genotyping array[93]. For this study, we only considered samples with a European Ancestry (EUR) as determined by HARE (Harmonized ancestry and race/ethnicity) analysis[94] (N=289,880).All relevant ethical regulations for work with human subjects were followed in the conduct of the study and written informed consent was obtained from all participants.

The Bio*Me* Biobank (N=9,255), at the Icahn School of Medicine at Mount Sinai (MSSM), is an EHR-linked biobank of participants from the Mount Sinai Health System in New York, NY.

Participant recruitment into Bio*Me* has been ongoing since 2007, predominantly recruited from general medicine and primary care clinics, and the rest from specialty practices and recruitment events. Bio*Me* participants consent to provide DNA and plasma samples linked to their de-identified EHRs, and then provide additional information on self-reported ancestry, health behaviors, and medical history through questionnaires administered upon enrollment. All participants in the study provided informed consent and study procedures followed guidelines for human subjects research.

The Massachusetts General Brigham Biobank (MGBB) (N=25,331), formerly known as the Partners Healthcare Biobank, is an ongoing virtual cohort study of patients across the MGB General Brigham hospital system (including Brigham and Women's Hospital, Massachusetts General Hospital, and other affiliated hospitals), which provides a large-scale resource of linked longitudinal electronic health records (EHR) data, genomic data, and self-reported survey data[89]. All patients provided informed consent before enrollment, and all study procedures were approved by the Massachusetts General Brigham Institutional Review Board.

Depression Polygenic Scoring

Depression polygenic scores were generated using PRS-CS[86] using SNP weights from the largest available depression meta-analysis[17]. The LD reference panel was constructed from 503 European samples in the 1000 Genomes Project phase 3[76]. PGS were scaled to have a mean of zero and a unit standard deviation (SD) so that effect estimates in subsequent analyses are interpreted per 1 SD increase in depression PGS. In VUMC data, the depression PGS explained 0.8% of the variance in MDD diagnosis (p-value=3.85×10^{-55}).

LabWAS of Depression PGS in VUMC

Associations between the depression PGS and labs were estimated with a lab-wide association scan (LabWAS) approach[95] controlled for sex and top 10 genetic principal components. In conditional analyses, the LabWAS of depression PGS was covaried for potential confounders, including BMI (median across each individual's EHR), and for depression, anxiety, adjustment reaction, and tobacco use disorder (as a proxy for smoking status) diagnoses, defined by phecodes 296.2, 300.1, 304, and 318, respectively. We controlled for BMI because of the reported relationship between obesity with inflammation[96], changes in metabolic markers[97], and risk of depression[98]. In sensitivity analyses focused on smoking, we mined smoking data from the social history forms within the EHR and extracted an ever/never smoking variable which indicates whether an individual has ever smoked. We tested the ever/never smoking variable as a covariate in the LabWAS of depression PGS. Primary analyses were restricted to individuals of European descent and repeated in individuals of African ancestry (n=12,384).

Sensitivity Analyses in VUMC

A series of conditional and sensitivity analyses were performed to ensure the association between depression PGS and WBC was not due to a common comorbid confounder phenotype present in individuals with both increased depression PGS and WBC. To find phenotypes associated with both depression PGS and WBC, separate phenome-wide association scans (PheWAS) were conducted of depression PGS and of the median, age-

adjusted, INT normalized WBC measurement. Next, phenotypes that were significantly associated with both depression PGS and WBC at Bonferroni significance ($p_{WBC} < 3.64 \times 10^{-5}$, $p_{depression\ PGS} < 3.72 \times 10^{-5}$) were selected and binned into seven categories based on phenotypic similarity. Group-based case-control variables were constructed, in which an individual was considered a case if they were a case for *any* of the group's phecodes. Controls were required to be a control for *all* phecodes. To assess the effect of the comorbid phenotypes on the association between depression PGS and WBC, a series of linear regression analyses were conducted controlling for each of the groups separately and all common phenotype groups together. All analyses were controlled for sex, top 10 genetic principal components, and median age across the medical record.

Replication in the PsycheMERGE Network

Targeted replication analyses focused on depression PGS and WBC counts were conducted in three external biobanks. Depression PGS were constructed and WBC count quality controlled as in VUMC. The depression PGS and age-adjusted WBC counts were fitted in a linear regression model controlling for sex and top 10 genetic principal components. The associations controlling for depression and anxiety diagnoses were also replicated using the same phenotype definition as described in the discovery LabWAS at VUMC. The effect estimates from each analysis were meta-analyzed across all four sites using a fixed-effect inverse variance weighted model in the meta[99] R package.

Depression PGS and WBC Mediation Analysis

Two mediation models were investigated using the mediation[100] R package. First, WBC count was modeled as the mediator between depression PGS (exposure) and depression diagnosis (outcome). Second, depression diagnosis was modeled as the mediator between depression PGS (exposure) and WBC (outcome).

While mediation analysis can be easily performed with continuous exposures (in this case the MDD-PGS), the calculation of the "proportion of variance mediated" cannot be interpreted on a continuous scale. Instead, we have to specify two discrete levels of the exposure in order to make the contrast (i.e., average MDD-PGS and high MDD-PGS). Therefore, the reference level (average MDD-PGS) and the comparison level (high MDD-PGS) must be defined by two distinct levels of the exposure variable. We selected individuals in the 50th percentile to represent the average MDD-PGS and tested three different comparison levels including individuals at the 85th, 90th, and 95th percentiles. There was no meaningful difference in the proportion mediated between the three comparison levels, thus we chose the 90th percentile as representative of the "high MDD-PGS" in the main text and have provided all results in Tables 15-16.

The proportion mediated estimates from all four sites were meta-analyzed using a fixed-effect inverse variance weighted model in the meta[99] R package. Due to the uniqueness of MVP (i.e., combat exposed, primarily male, etc.) compared to the other sites, we also conducted meta-analyses excluding MVP (Tables 15-16).

<u>Depression PGS and WBC-differential Mediation Analysis</u>

To determine which WBC cell types contributed to the association between depression PGS and depression diagnosis, a series of multiple mediator analyses were conducted using the mediation[100] R package. Each WBC subtype count was analyzed as the main mediator between depression PGS (exposure) and depression diagnosis (outcome) with the remaining subtypes as the alternative mediators. In a multiple mediator analysis, a single main mediator and additional alternative mediators are specified. A structural equation modeling approach is used to assess the effect of the main mediator between the exposure and outcome after controlling for the correlation structure between the alternative mediators and the outcome[100]. All measurements were required to be recorded on the same date for each individual to ensure they were from the same WBC-differential (N=24,383). For individuals with multiple WBC-differentials recorded in their EHR, median WBC count values and the corresponding subtype absolute values were selected. All measurements were adjusted for cubic splines of age at observation and normalized using a rank-based inverse normal transformation[90,101].

Mendelian Randomization

We conducted bidirectional Mendelian Randomization (MR) between depression and WBC count using generalized summary-based MR (GSMR)[102] in the GCTA package. Index SNPs were selected using the default settings in GCTA: p-value threshold of $5x10^{-8}$, linkage disequilibrium r^2 clumping threshold of 0.05, and a HEIDI-outlier threshold of 0.01 to remove SNPs that have pleiotropic effects on both risk factor and disease. From the depression[17] and WBC count[103] summary statistics, 47 and 203 SNPs were selected as index SNPs, respectively.

LabWAS, PheWAS, conditional, replication, and mediation analyses were conducted using R, version 3.4.3. Mendelian Randomization was conducted using GCTA version 1.92.4. Code for each analysis can be found here: https://bitbucket.org/davislabteam/mdd-pgs-labwas

Results

LabWAS of Depression PGS

Depression PGS were screened for associations with 315 clinical lab measurements using a lab-wide association scan (LabWAS)[95] in VUMC's biobank (N=70,704). After multiple testing correction, the LabWAS of depression PGS revealed significant associations with four elevated immune markers, white blood cell (WBC) count (p-value=$1.07x10^{-17}$, beta=0.03, SE=0.004), urinary WBC (p-value=$1.45x10^{-5}$, beta=0.03, SE=0.007), absolute monocyte count (p-value=$2.54x10^{-5}$, beta=0.02, SE=0.005), and absolute neutrophil count (p-value=$5.91x10^{-5}$, beta=0.02, SE=0.005). Significant associations also included several metabolic markers including increased triglycerides (p-value=$3.14x10^{-18}$, beta=0.05, SE=0.006), decreased HDL-C (p-value=$1.23x10^{-11}$, beta=-0.04, SE=0.005), decreased calcitriol (p-value=$2.83x10^{-8}$, beta=-0.04, SE=0.007), increased glucose (p-value=$2.84x10^{-7}$, beta=0.02, SE=0.004), decreased blood urea nitrogen (BUN) (p-value=$5.19x10^{-7}$, beta=-0.02, SE=0.004), decreased calcium (p-value=$9.74x10^{-7}$, beta=-0.02, SE=0.004), and decreased calcidiol (p-value=$7.03x10^{-5}$, beta=-0.04, SE=0.01). Depression PGS also associated with decreased troponin-I (p-value=$1.09x10^{-6}$, beta=-0.05, SE=0.009), decreased urinary red blood cells (p-value=$1.37x10^{-5}$, beta=-0.03, SE=0.006), decreased thyroxine (p-value=$1.72x10^{-5}$, beta=-0.03, SE=0.006), and decreased blood carbon dioxide (p-value=$4.06x10^{-6}$, beta=-0.02, SE=0.003) (Figure 11a, Table 9).

In a conditional analysis, we sequentially controlled for diagnoses for depression, anxiety, adjustment reaction, and tobacco use disorder, and median BMI across the EHR (Figure 12, Table 10). In the analysis with all covariates, the most significant association remained WBC count (p-value=1.11×10^{-10}, beta=0.03, SE=0.005), followed by triglycerides (p-value=1.91×10^{-5}, beta=0.04, SE=0.008) (Figure 11b).

While depression PGS remained robustly associated with WBC count across all analyses, the magnitude of the effect was modest (beta=0.03). Stratification of individuals in the discovery cohort (VUMC) showed even at the highest decile of depression PGS, WBC count measurements are elevated but remain within the clinical reference range (4-11 thousand cells/uL) (Figure 13). Individuals in the first decile of depression PGS had a mean WBC count measurement of 7.94 compared to a mean WBC count of 8.34 for individuals in the top decile of depression PGS, equating to a 5% increase in WBC count from the first to tenth decile.

No labs were significantly associated in the LabWAS of depression PGS in individuals of African descent, likely due to the smaller sample size of the African ancestry sample (n=12,383), and the low generalizability of polygenic scores built using European summary statistics in African decent populations[104]. However, the association with WBC count was in the same direction, with a similar magnitude, as in the European sample (p-value=0.058, beta=0.02, SE=0.01) (Figure 14).

Figure 11. Lab-wide association scan of depression polygenic score in individuals of European ancestry. In analysis A) associations were controlled for sex and top 10 principal components of ancestry. In analysis B) associations were controlled for sex, top 10 principal components of ancestry, diagnoses for depression, anxiety, adjustment disorder, tobacco use disorder, and median BMI across EHR.

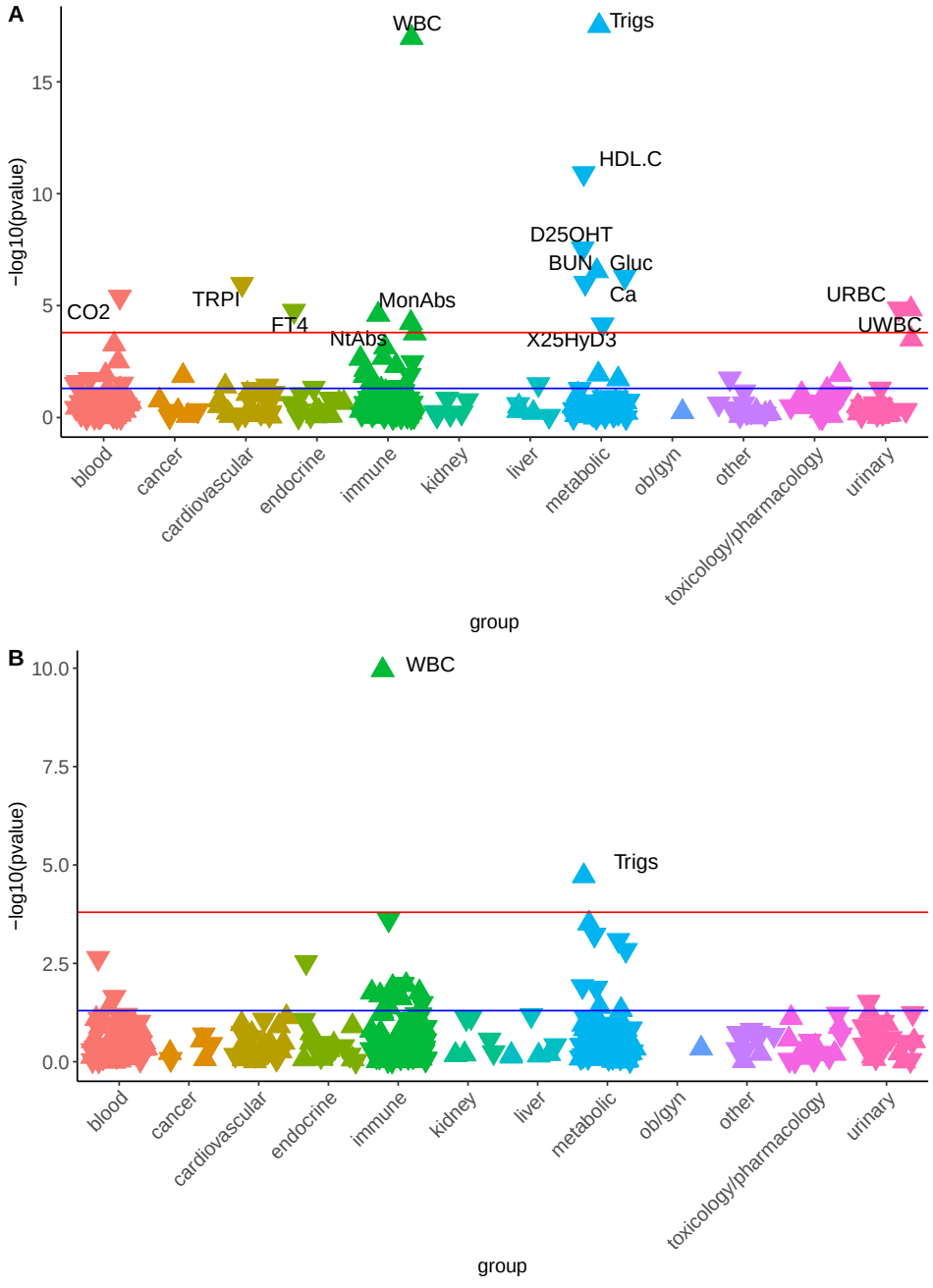


Figure 12. LabWAS of depression PGS in VUMC controlled for a) sex and top 10 principal components of ancestry, b) depression diagnosis, c) depression and anxiety diagnoses, d) depression, anxiety, and adjustment reaction, e) diagnoses for depression, anxiety, adjustment reaction, and median BMI, f) diagnoses for depression, anxiety, adjustment reaction, tobacco use disorder and median BMI, and g) diagnoses for depression, anxiety, adjustment reaction, median BMI, and smoking ever documented in the EHR.

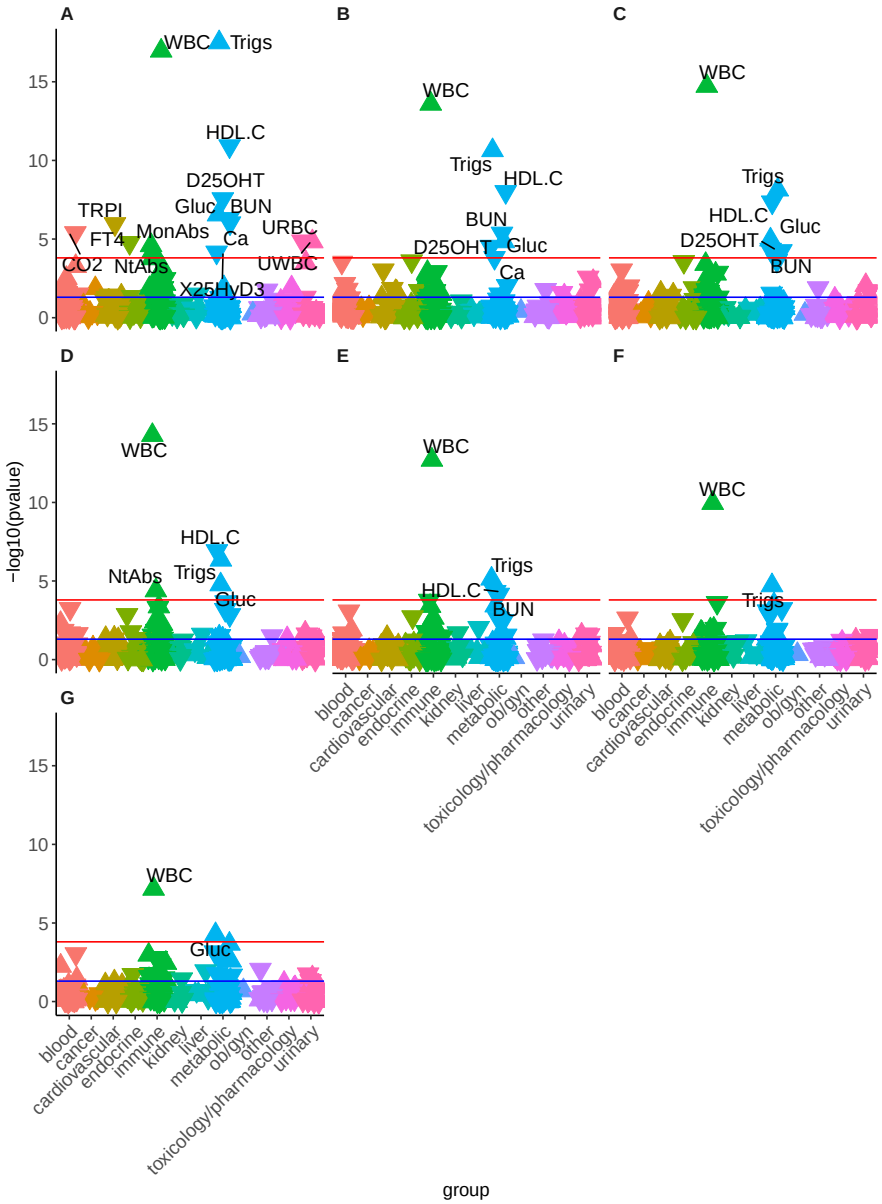


Figure 13. Median WBC measurements stratified by depression PGS decile in VUMC. Individuals were divided into deciles based on their depression PGS. Each individual's median untransformed WBC measurement is plotted based on depression PGS decile. Blue lines indicate the normal clinical range for WBC (4-11thou cells/uL). The dotted line in between boxes connects median WBC values between deciles.

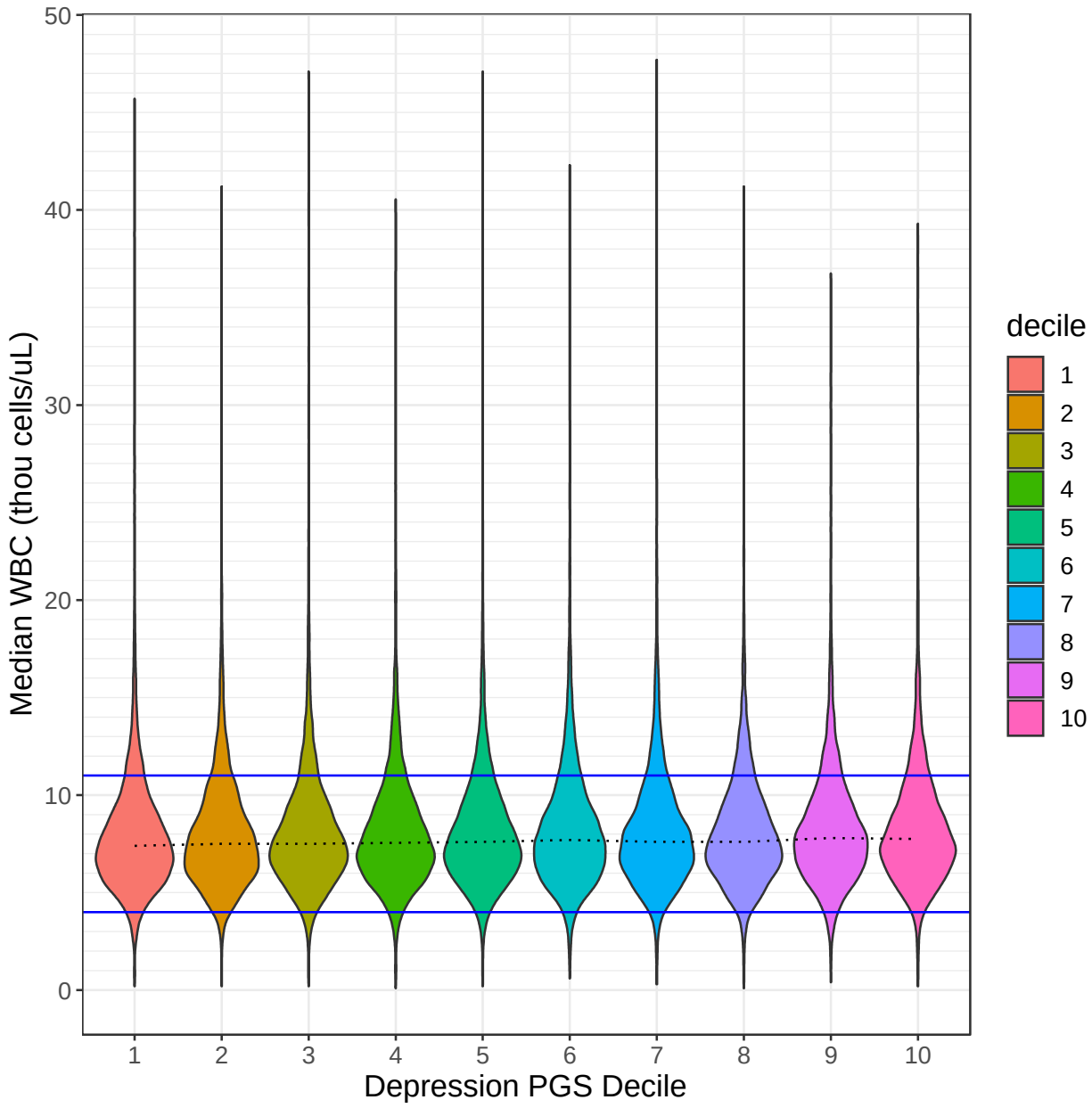


Figure 14. Lab-wide association scan of depression PGS in individuals of African ancestry in VUMC. Associations were controlled for sex and top 1o principal components of ancestry. The blue line represents p-value = 0.05, and the red line represents Bonferroni significance (p-value=2.21 x 10^{-4}).

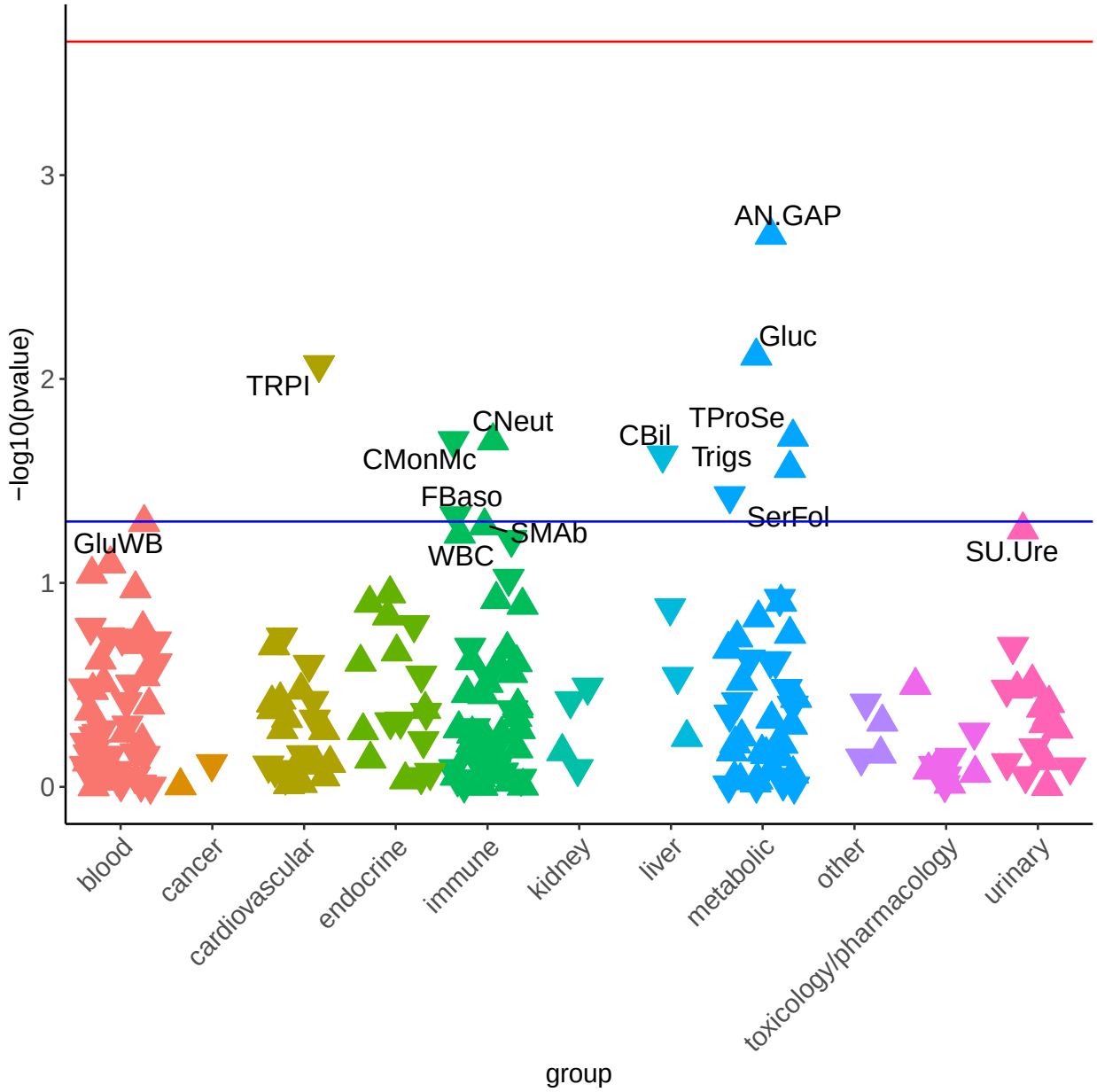


Table 9. Significant associations from the LabWAS of depression polygenic scores in VUMC.

Associations were controlled for sex and top 10 principal components of ancestry.

Lab	Full Name	Group	N	p-value	Beta	SE
Trigs	Triglyceride [Mass/volume] in Serum or Plasma	metabolic	30,703	3.14E-18	0.049	0.006
WBC	Leukocytes [#/volume] in Blood by Automated count	immune	65,120	1.07E-17	0.030	0.004
HDL.C	Cholesterol in HDL [Mass/volume] in Serum or Plasma	metabolic	29,598	1.23E-11	-0.037	0.005
D25OHT	Calcitriol [Mass/volume] in Serum or Plasma	metabolic	18,325	2.83E-08	-0.039	0.007
Gluc	Glucose lab	metabolic	62,555	2.84E-07	0.019	0.004
BUN	Urea nitrogen serum/plasma	metabolic	62,617	5.19E-07	-0.020	0.004
Ca	Calcium serum/plasma serum/plasma	metabolic	62,357	9.74E-07	-0.017	0.004
TRPI	Troponin I.cardiac [Mass/volume] in Serum or Plasma	cardiovascular	10,443	1.09E-06	-0.045	0.009
CO2	Carbon dioxide serum/plasma	blood	62,518	4.06E-06	-0.016	0.003
URBC	Erythrocytes [#/area] in Urine sediment by Microscopy high power field	urinary	23,022	1.37E-05	-0.025	0.006
UWBC	Leukocytes [#/area] in Urine sediment by Microscopy high power field	urinary	24,388	1.45E-05	0.029	0.007
FT4	Thyroxine (T4) free [Mass/volume] in Serum or Plasma	endocrine	26,261	1.72E-05	-0.026	0.006
MonAbs	MonAbs	immune	25,188	2.54E-05	0.022	0.005
NtAbs	NtAbs	immune	25,176	5.91E-05	0.021	0.005
X25HyD3	Calcidiol [Mass/volume] in Serum or Plasma	metabolic	9,525	7.03E-05	-0.040	0.010

Table 10. Association between WBC and depression PGS from conditional LabWAS analyses. All analyses were controlled for sex and top 10 principal components of ancestry.

Covariates	**N**	**pvalue**	**Beta**	**SE**
Original	65,120	1.07E-17	0.030	0.004
Depression Diagnosis	64,679	2.54E-14	0.031	0.004
Depression + Anxiety Diagnosis	64,679	1.88E-15	0.035	0.004
Depression + Anxiety + Adjustment Reaction Diagnosis	64,679	5.29E-15	0.036	0.005
Depression + Anxiety + Adjustment Reaction Diagnosis + BMI	61,793	1.99E-13	0.034	0.005
Depression + Anxiety + Adjustment Reaction Diagnosis + Tobacco Use Disorder + BMI	61,793	1.11E-10	0.031	0.005
Depression + Anxiety + Adjustment Reaction Diagnosis + BMI + Ever/Never Smoking	61,793	6.84E-08	0.028	0.005

Conditional Analyses of WBC

In separate PheWAS analyses, depression PGS and median WBC count were significantly associated with 66 and 469 phecodes, respectively. Of these significantly associated phecodes, 32 were common to both depression PGS and median WBC count and were binned into seven categories based on phenotypic similarity: cardiovascular, psychiatric, obesity, respiratory, hepatic, pain, and autoimmune conditions (Figure 15a, Table 11).

The association between depression PGS and WBC count remained significant after controlling for each group separately and controlling for all phenotype groups together (p-value=4.19×10^{-3}, beta=0.02) with effect estimates similar to the original association despite the reduced sample size (N = 13,269) (Figure 15b, Table 12).

Figure 15. Controlling for common phenotypes between depression PGS and WBC. a) Phenotypes associated with both depression PGS and WBC divided into groups based on phenotypic similarity in VUMC. b) The association between depression PGS and WBC controlling for each “confounder” phenotype group in VUMC. Group-based cases were any individual who was a case for a any of a group’s phecodes and controls were individuals who were controls for all of a group’s phecodes. Associations were found using linear regressions controlled for each group. In the ”all” analysis, all groups were controlled for in one regression.

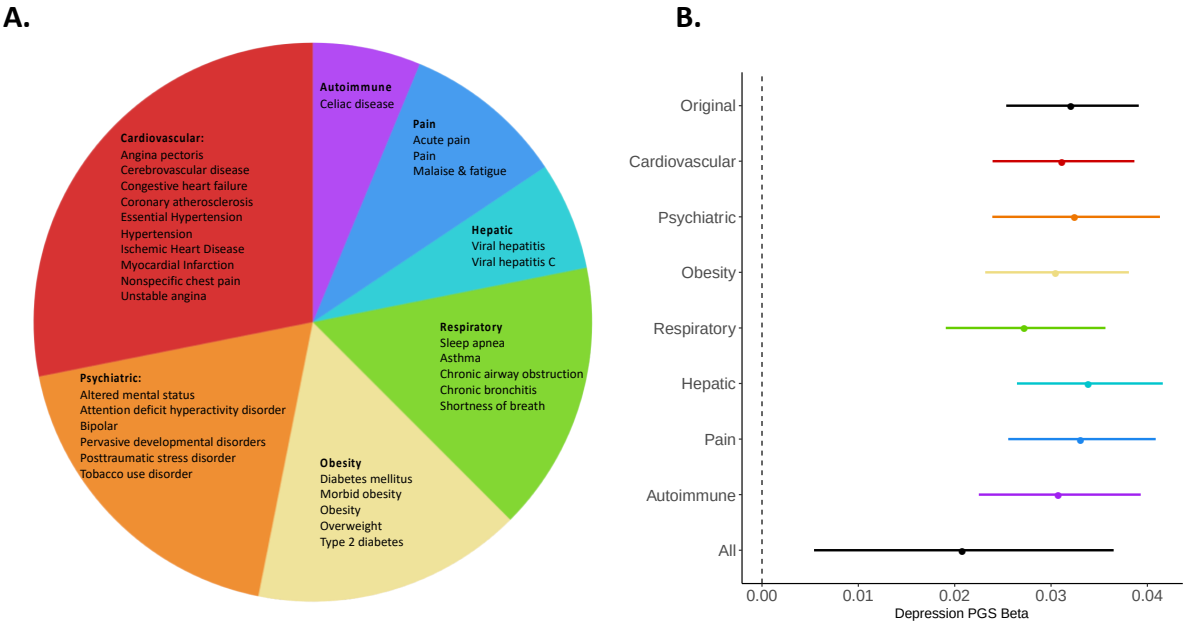


Table 11. Phenotypes associated with depression PGS and median WBC measurements using PheWAS in VUMC binned into categories based on similarity.

Group	Phecode	Phenotype	WBC p-value	WBC beta	Depression PGS p-value	Depression PGS beta
Cardiovascular	411.3	Angina pectoris	5.14E-20	0.0265	1.54E-06	0.0244
	433	Cerebrovascular disease	1.09E-29	0.0176	6.37E-06	0.0157
	428.1	Congestive heart failure (CHF) NOS	3.90E-69	0.017	2.38E-05	0.0152
	411.4	Coronary atherosclerosis	2.34E-98	0.0151	1.12E-08	0.0131
	401.1	Essential hypertension	3.26E-143	0.0115	4.90E-06	0.0099
	401	Hypertension	6.38E-142	0.0114	9.73E-07	0.0098
	411	Ischemic Heart Disease	5.21E-104	0.0144	7.38E-09	0.0125
	411.2	Myocardial infarction	1.17E-78	0.0213	1.22E-05	0.0192
	418	Nonspecific chest pain	5.14E-18	0.0116	1.97E-14	0.0105
	411.1	Unstable angina (intermediate coronary syndrome)	9.82E-21	0.0274	1.95E-08	0.0252
Psychiatric	292.4	Altered mental status	1.29E-14	0.0194	1.91E-05	0.0176
	313.1	Attention deficit hyperactivity disorder	1.22E-12	0.0316	2.27E-07	0.0324
	296.1	Bipolar	9.90E-06	0.0263	3.91E-23	0.0261
	313	Pervasive developmental disorders	2.63E-15	0.0265	2.40E-08	0.0274
	300.9	Posttraumatic stress disorder	1.96E-06	0.0324	2.76E-25	0.0327
	318	Tobacco use disorder	4.77E-175	0.0152	1.89E-25	0.0137
Obesity	250	Diabetes mellitus	3.02E-100	0.013	1.62E-06	0.0116
	278.11	Morbid obesity	5.06E-90	0.021	5.71E-09	0.0197
	278.1	Obesity	4.41E-94	0.0151	9.65E-09	0.0139
	278	Overweight, obesity and other hyperalimentation	1.11E-81	0.0141	9.54E-08	0.0130
	250.2	Type 2 diabetes	7.15E-110	0.0134	7.60E-08	0.0119
Respiratory	327.3	Sleep apnea	1.36E-14	0.0161	9.64E-06	0.0149
	495	Asthma	9.18E-25	0.0181	3.86E-06	0.0171
	496	Chronic airway obstruction	5.69E-144	0.018	1.68E-08	0.0160
	496.2	Chronic bronchitis	6.60E-55	0.0325	1.02E-05	0.0300
	512.7	Shortness of breath	1.59E-91	0.0132	6.99E-07	0.0117
Hepatic	70	Viral hepatitis	7.14E-53	0.0301	4.40E-07	0.0279
	70.3	Viral hepatitis C	6.85E-52	0.0319	1.35E-06	0.0296
Pain	338.1	Acute pain	8.70E-23	0.0151	2.38E-06	0.0138
	338	Pain	1.20E-20	0.0133	4.09E-09	0.0121
	798	Malaise and fatigue	1.45E-63	0.0108	3.49E-07	0.0098
Autoimmune	557.1	Celiac disease	1.52E-15	0.0568	1.78E-09	0.0545

Table 12. Association between depression PGS and WBC levels controlled for common phenotype groups in VUMC. The association was controlled for each phenotype group separately and all groups in one analysis in the "All" phenotype. Associations were found using a linear regression controlling for sex and top 10 principal components of ancestry.

Phenotype	**N**	**Depression PGS P-value**	**Depression PGS beta**	**SE**	**Lower 95% CI**	**Upper 95% CI**
Original	65,120	1.07E-17	0.030	0.003	0.023	0.037
Cardiovascular	55,184	8.1E-17	0.031	0.004	0.024	0.039
Psychiatric	41,213	2.1E-13	0.033	0.004	0.024	0.041
Obesity	54,28	8.4E-16	0.031	0.004	0.023	0.038
Respiratory	44,274	9.7E-11	0.027	0.004	0.019	0.036
Hepatic	54,016	1.2E-18	0.034	0.004	0.027	0.042
Pain	52,649	1.9E-17	0.033	0.004	0.026	0.041
Autoimmune	44,504	5.5E-12	0.029	0.004	0.021	0.037
All	13,269	0.0082	0.021	0.008	0.005	0.037

Replication in the PsycheMERGE Network

Given the robustness of the association with WBC and the history of associations between depression status and pro-inflammatory markers, we focused on WBC count for replication and further investigation. Findings were replicated in three external biobanks, the Million Veteran Program (MVP), Mount Sinai Icahn School of Medicine (MSSM), and Massachusetts General Brigham Biobank (MGBB) (Table 13). In both MVP (N=289,880) and MGBB (N=20,828), the association between depression PGS and WBC count remained significant with effect estimates replicating those observed at VUMC (Figure 16). In MSSM, the effect size point estimate was similar to those observed in the three other sites, but did not reach statistical significance, likely due to the smaller sample size (n=823). The meta-analyzed effect estimate from the four sites was robust and highly significant (p-value=1.03×10^{-136}, beta=0.03, SE=0.002), even after controlling for depression diagnosis (p-value=9.52×10^{-102}, beta=0.03, SE=0.002), and after controlling for depression and anxiety diagnoses (p-value=8.23×10^{-100}, beta=0.03, SE=0.002) (Figure 16, Table 14).

Figure 16. Replication within the PsycheMERGE Network. The association between depression PGS and median WBC levels was replicated across the PsycheMERGE Network with sensitivity analyses controlling for depression and anxiety diagnoses.

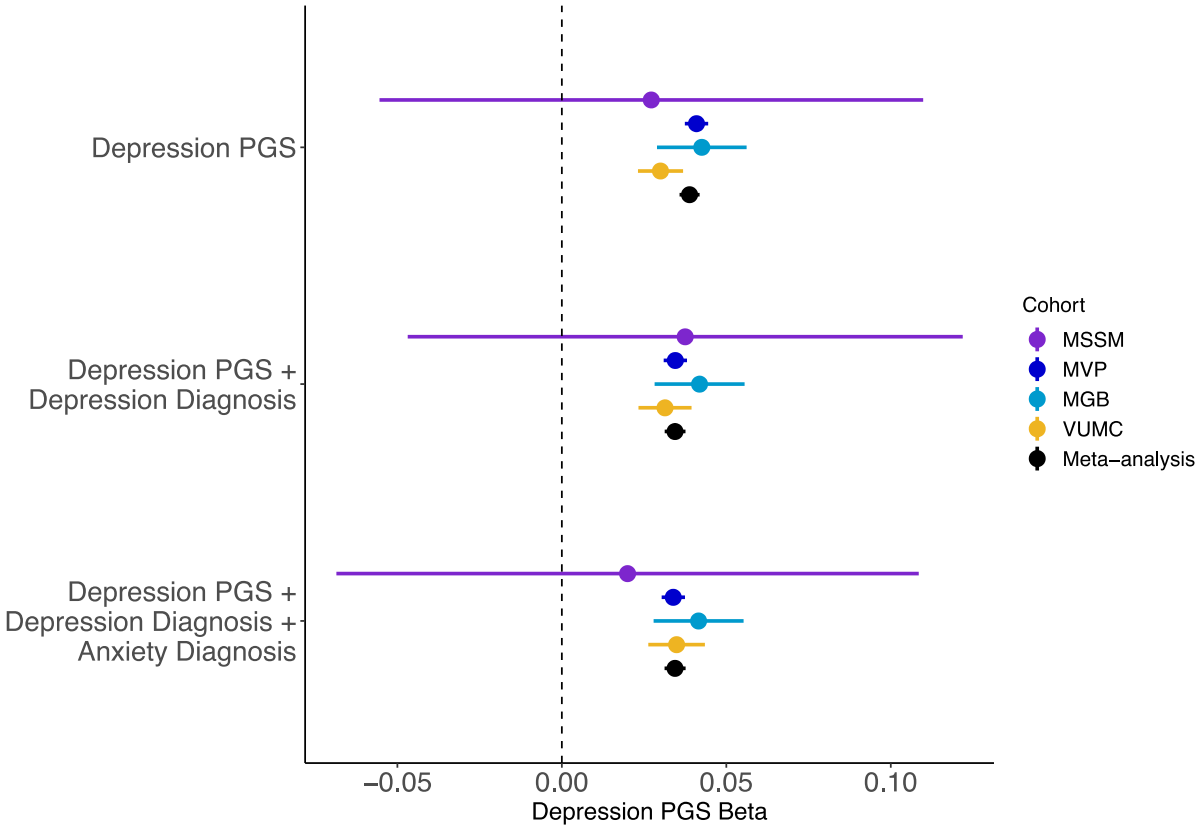


Table 13. Characteristics of PsycheMERGE Network samples.

Site	Group	N genotyped (European)	N WBC measurement	N genotyped & WBC measurement	% Female	Average age in years, (SD)	Average Length of record in years, (SD)
MSSM	All	9,255	3,668	823	52.10%	59.7 (16.0)	11.2 (4.4)
	Depression or Anxiety Controls	6,722	2,499	578	51.4%	59.3 (16.4)	10.7 (4.4)
	Depression or Anxiety Cases	1,622	1,169	245	53.9%	60.5 (15.0)	12.5 (3.9)
VUMC	All	72,828	948,590	70,921	55.9%	48.1 (22.3)	8.7 (6.3)
	Depression or Anxiety Controls	59,520	301,982	57,161	52.6%	46.8 (23.7)	7.6 (6.1)
	Depression or Anxiety Cases	15,985	71,692	15,951	64.4%	50.9 (18.8)	11.3 (6.1)
MVP	All	289,880	289,880	289,880	7.2%	64.3 (12.0)	12.0
	Depression or Anxiety Controls	150,328	150,328	150,328	4.1%	67.7 (11.2)	11.2
	Depression or Anxiety Cases	129,552	129,552	129,552	10.9%	61.6 (11.9)	12.9
MGB	All	25,331	72,329	20,828	51.5%	56.1	13.8 (8.3)
	Depression or Anxiety Controls	17,879	51,612	17,098	52.0%	59.8 (16.7)	11.3 (7.1)
	Depression or Anxiety Cases	7,452	20,717	3,730	64.2%	56.7 (16.9)	14.0 (6.7)

Table 14. Results of PsycheMERGE replication between depression PGS and WBC levels. Associations were controlled for sex and top 10 principal components. Beta estimates were combined for meta-analysis using a fixed-effects inverse weighted method.

Analysis	Cohort	P-value	Beta	SE	Lower 95% CI	Upper 95% CI
Depression PGS	MSSM	0.519	0.027	0.042	-0.055	0.11
	MVP	3.84E-114	0.041	0.002	0.037	0.044
	MGB	9.11E-10	0.043	0.007	0.029	0.056
	VUMC	1.07E-17	0.03	0.004	0.023	0.037
	Meta-analysis	1.03E-136	0.034	0.002	0.031	0.038
Depression PGS + Depression Diagnosis	MSSM	0.384	0.038	0.043	-0.047	0.122
	MVP	2.12E-81	0.035	0.002	0.031	0.038
	MGB	1.92E-9	0.042	0.007	0.028	0.056
	VUMC	2.54E-14	0.031	0.004	0.023	0.039
	Meta-analysis	9.52E-102	0.034	0.002	0.031	0.038
Depression PGS + Depression Diagnosis + Anxiety Diagnosis	MSSM	0.658	0.02	0.045	-0.069	0.109
	MVP	1.71E-78	0.034	0.002	0.03	0.037
	MGB	2.55E-9	0.042	0.007	0.028	0.055
	VUMC	1.88E-15	0.035	0.004	0.026	0.044
	Meta-analysis	8.23E-100	0.034	0.002	0.031	0.038

Mediation Analysis

Two potential pathways between depression PGS, WBC count, and depression diagnosis were assessed using mediation analyses. In the first analysis, median WBC was modeled as a mediator of the relationship between depression PGS (exposure) and depression diagnosis (outcome). Meta-analysis across all sites revealed WBC mediated 2.5% of the association between depression PGS and depression diagnosis (95% CI=2.2-20.8%, p-value=2.84×10^{-70}) (Table 15). When excluding MVP from the meta-analysis, WBC count mediated 0.5% of the association, although this association was not statistically significant (95% CI=-0.03-0.9%, p-value=0.06) (Table 15).

In the second analysis, depression diagnosis was modeled as a mediator of the association between the depression PGS (exposure) and median WBC (outcome). Meta-analysis across all sites indicated depression diagnosis mediated 9.8% of the association between depression PGS and WBC count (95% CI=8.4-11.1%, p-value=1.78×10^{-44}) (Table 17). MDD diagnosis mediated 1.4% of the association when excluding MVP from the meta-analysis (95% CI=-0.6-3.4%, p-value=0.17) (Table 16).

Depression PGS and WBC-differential Mediation Analysis

WBC counts are calculated from the sum of five different cell subtypes: neutrophils, lymphocytes, monocytes, basophils, and eosinophils. These cell subtypes can be measured along with the total WBC using a complete blood count differential (CBC-differential) lab. To determine whether specific WBC components accounted for the relationships between depression PGS and depression diagnosis, we performed a series of multiple mediator analyses.

When depression PGS was modeled as the exposure and depression diagnosis as the outcome, neutrophils were the only cell type that explained a significant proportion (1.9%; 95% CI=0.2–3.1%) of the association between depression PGS and depression diagnosis (Table 17).

Table 15. Mediation results with WBC as the mediator across PsycheMERGE sites. Proportion mediated estimates were estimated using three different treatment percentiles of depression PGS (85%, 90%, and 95%) compared to the 50% percentile of PGS for controls.

Cohort	MDD PGS Percentile	Control Percentile Value	Treatment Percentile Value	Proportion Mediated pvalue	Proportion Mediated (SE)	Lower 95% CI
VUMC	0.85	0.011	1.033	0.138	0.003 (0.003)	-0.001 – 0.008
	0.90		1.279	0.138	0.003 (0.003)	-0.001 – 0.008
	0.95		1.634	0.138	0.003 (0.003)	-0.001 – 0.008
MVP	0.85	0.024	1.026	<2.23e-308	0.035 (0.002)	0.031 – 0.038
	0.90		1.258	<2.23e-308	0.035 (0.002)	0.031 – 0.038
	0.95		1.6	<2.23e-308	0.035 (0.002)	0.031 – 0.038
MGB	0.85	0.002	1.038	0.014	0.012 (0.006)	0.003 – 0.024
	0.90		1.264	0.014	0.012 (0.006)	0.003 – 0.024
	0.95		1.62	0.014	0.012 (0.006)	0.003 – 0.024
MSSM	0.85	-0.009	1.049	0.86	-0.016 (0.06)	-0.240 – 0.100
	0.90		1.331	0.868	-0.016 (0.069)	-0.242 – 0.118
	0.95		1.738	0.862	-0.016 (0.062)	-0.240 – 0.105
Meta-analysis	0.85	-	-	3.20E-70	0.025 (0.001)	0.022 – 0.208
	0.90	-	-	2.84E-70	0.025 (0.001)	0.022 – 0.208
	0.95	-	-	2.57E-70	0.025 (0.001)	0.022 – 0.208
Meta-analysis excluding MVP	0.85	-	-	0.066	0.005 (0.002)	-0.0003 – 0.009
	0.90	-	-	0.066	0.005 (0.002)	-0.0003– 0.009
	0.95	-	-	0.066	0.005 (0.002)	-0.0003– 0.009

Table 16. Mediation results with MDD diagnosis as the mediator across PsycheMERGE sites. Proportion mediated estimates were estimated using three different treatment percentiles of depression PGS (85%, 90%, and 95%) compared to the 50% percentile of PGS for controls.

Cohort	MDD PGS Percentile	Control Percentile Value	Treatment Percentile Value	Proportion Mediated p-value	Proportion Mediated (SE)	Lower 95% CI
VUMC	0.85	0.011	1.033	0.152	0.01 (0.011)	-0.004 – 0.032
	0.90		1.279	0.152	0.01 (0.011)	-0.004 – 0.032
	0.95		1.634	0.152	0.011 (0.011)	-0.004 – 0.032
MVP	0.85	0.024	1.026	<2.23e-308	0.162 (0.01)	0.143 – 0.181
	0.90		1.258	<2.23e-308	0.162 (0.009)	0.144 – 0.180
	0.95		1.6	<2.23e-308	0.162 (0.009)	0.145 – 0.180
MGB	0.85	0.002	1.038	0.012	0.044 (0.033)	0.011 – 0.108
	0.90		1.264	0.012	0.044 (0.033)	0.011 – 0.108
	0.95		1.62	0.012	0.045 (0.033)	0.011 – 0.109
MSSM	0.85	-0.009	1.049	0.784	-0.113 (0.57)	-1.409 – 1.003
	0.90		1.331	0.732	-0.104 (0.517)	-1.511 – 0.910
	0.95		1.738	0.73	-0.084 (0.56)	-1.042 – 1.014
Meta-analysis	0.85	-	-	5.91E-40	0.095 (0.007)	0.081 – 0.109
	0.90	-	-	1.78E-44	0.098 (0.007)	0.084 – 0.111
	0.95	-	-	9.73E-45	0.097 (0.007)	0.083 - 0.110
Meta-analysis excluding MVP	0.85	-	-	0.203	0.014 (0.011)	-0.007 – 0.035
	0.90	-	-	0.197	0.014 (0.011)	-0.007 – 0.034
	0.95	-	-	0.170	0.014 (0.010)	-0.006 – 0.034

Table 17. Immune subpopulation and depression diagnosis mediation analysis. Using a multiple mediator analysis, each subpopulation was modeled as the main mediator between the exposure and the outcome with the remaining subpopulations as alternative mediators.

Outcome	**Cell Type**	**Proportion Mediated**	**Lower 95% CI**	**Upper 95% CI**
MDD diagnosis	Basophils	0.005	-0.009	0.015
	Eosinophils	0.002	-0.015	0.014
	Lymphocytes	0.008	-0.007	0.018
	Monocytes	-0.001	-0.018	0.011
	Neutrophils	0.019	0.002	0.031

Mendelian Randomization

When modeling WBC count as the exposure and depression as the outcome, MR analysis provided additional evidence for an increase in depression risk with an increase in WBC count (p-value=0.014, b_{xy}=0.27) (Figure 17, Table 18). However, depression modeled as the exposure showed no evidence of causal influence on the WBC count outcome (p-value=0.302, b_{xy}=0.022).

Figure 17. Results of bidirectional Mendelian Randomization with A) MDD as the exposure and WBC as the outcome, and B) WBC as the exposure and MDD as the outcome.

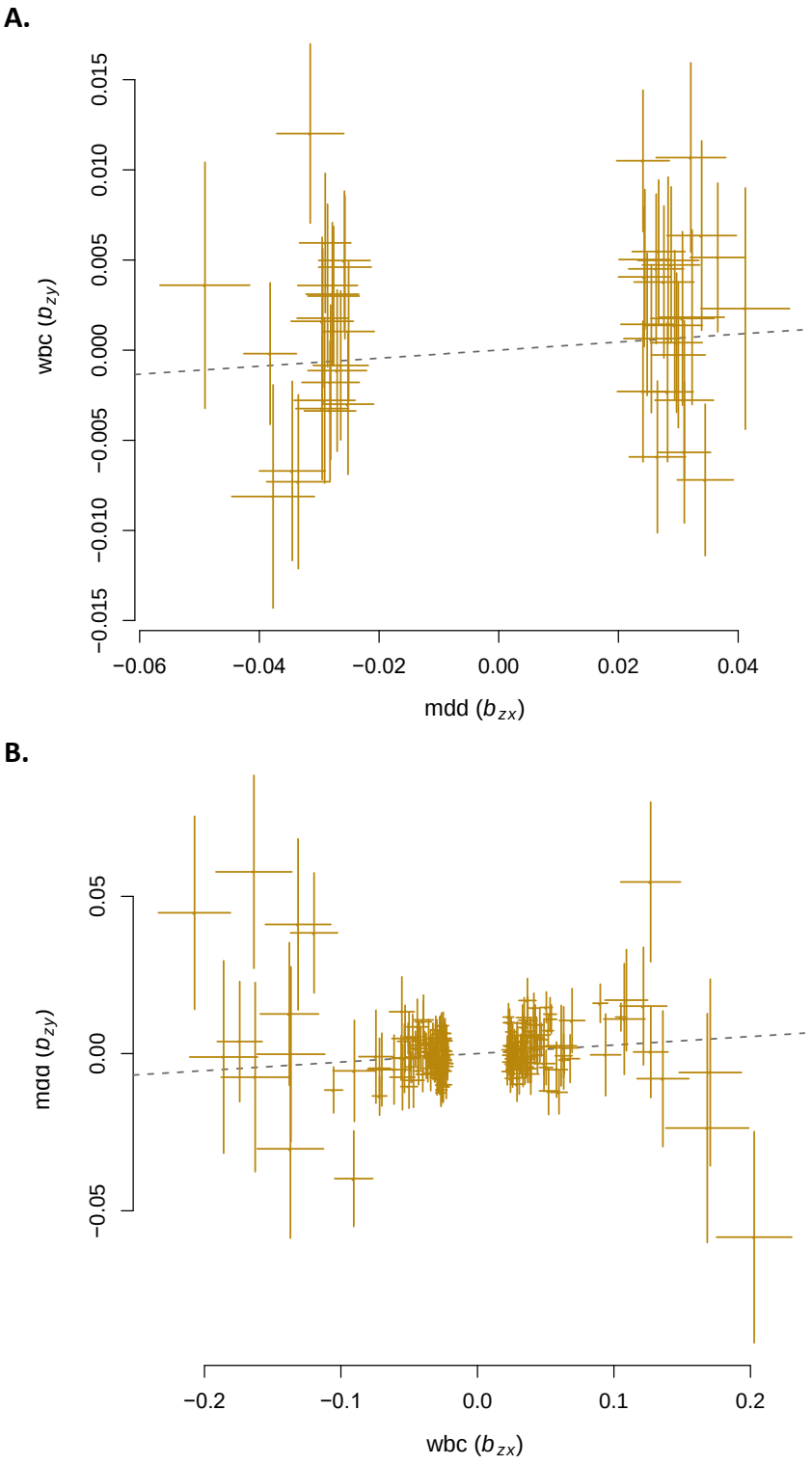


Table 18. Results of bidirectional Mendelian Randomization between depression and WBC.

Exposure	**Outcome**	**b_{xy}**	**SE**	**P-value**	**N SNPs**	**Multi SNP based HEIDI Outlier**
MDD	WBC	0.0223	0.0216	0.302	47	0.639
WBC	MDD	0.0272	0.0110	0.014	203	0.021

Discussion

Depression is consistently associated with increased pro-inflammatory biomarkers, however, the mechanisms underlying these associations remain unclear. In this study, analysis of EHR-linked biobanks within the PsycheMERGE Network were utilized to examine the effect of depression polygenic scores on a variety of clinical lab traits, revealing a robustly replicated association with increased white blood cell count. Notably, several other lab traits associated with depression polygenic scores, including lipids, blood glucose, and blood urea nitrogen. The variety of associations with depression PGS indicate multiple areas of biology are affected by depression genetics, including metabolism[105,106] and inflammation[106–108]. We chose to further investigate the relationship with WBC count given the existing literature and the robustness of the observed association to clinical confounders.

In a lab-wide screen, increased polygenic depression risk was associated with increased inflammatory markers including WBC count, even after controlling for depression, anxiety, multiple comorbid phenotypes, BMI, and smoking, thus highlighting depression PGS as an important risk factor for the pro-inflammatory state observed in depression. These results indicated that genetic risk for depression, independent of depressive symptoms, is linked to a pro-inflammatory biomarker. The effect of the depression PGS on WBC count was modest across all biobanks, suggesting that individuals with high depression genetic liability may have an activated, but not abnormal immune system. Nonetheless, sustained activation of the immune system could have important implications for the risk of developing depression.

The results of our mediation analyses did not distinguish a singular causal pathway. Instead, depression diagnosis mediated 2.5% of the association between depression PGS and

WBC count, while WBC count mediated 9.8% of the association between depression PGS and depression diagnosis. However, mendelian randomization results do provide further evidence for a causal path from increased WBC count levels to increased depression risk, but do not support a model of depression leading to increased WBC levels. It is important to note that only 47 SNPs met criteria to be included as MDD instrument variables, limiting the statistical power of the analysis.

The notable difference in the proportions mediated between MVP and the other sites could be due to phenotypic uniqueness of the MVP sample. For example, MVP is overwhelmingly male (92.8%), which could contribute to residual confounding by sex that is not fully accounted for in the model. Additionally, the mediated pathways could be particularly strong in MVP due to the high prevalence of depression in the sample (MVP=44.7%, others=23.3%). A sensitivity analysis excluding MVP yielded marginally significant results, and indicate that additional analysis in larger sample size is warranted.

In the clinic, WBC measurements can be broken down into measurements of each WBC subtype. Abnormal levels of different WBC subtypes can index different immune processes. Understanding which cell types underlie the relationship between depression polygenic score and depression diagnosis through WBC can help narrow a specific immune process involved in depression. Neutrophil counts explained 1.9% of the association between depression polygenic score and depression diagnosis, and no other subtypes contributed significantly to the association. Neutrophils are well known as responders to acute bacterial infection[109] and are the most abundant WBC subtype in circulation (40-60%)[109]. Although neutrophils are typically

restricted from crossing the blood-brain barrier, neutrophils are known to infiltrate the central nervous system during infection, trauma, or neurodegeneration[110].

Our study should be interpreted in light of its limitations. First, the WBC measurements used in the study were clinically derived, with measurements reflecting a range of health states. To address this, we limited to observations within 4 standard deviations (described in supplement) and noted that WBC was measured on nearly everyone in our primary and replication sample populations. However, it remains possible that individuals with clinical orders for WBC differential panels may represent a clinically different sample than those with only the total WBC measurement. Additionally, EHRs often contain multiple WBC measurements for the same individual. In this study, only the median values per individual were utilized, leaving unanswered questions about the effect of depression PGS on WBC over time and in response to antidepressant treatment. Finally, even though the relationship between depression PGS and WBC was robust, the effect sizes are small, making WBC an unlikely candidate for use as a diagnostic biomarker of depression.

Polygenic scores for depression are associated with increased inflammatory markers, specifically WBC count, even in the absence of depressive symptoms. Inflammatory markers may play a causal role in the etiology of depression and subsequent inflammation. The associations described in this study highlight the importance of WBC biology in depression and demonstrate the use of EHR-based genomics as a tool for discovery of physiological markers in psychiatric traits.

CHAPTER IV

EVALUATING THE LONGITUDINAL EFFECT OF ANTIDEPRESSANT USE ON WHITE BLOOD CELL COUNT

Introduction

In the United States, antidepressants are routinely prescribed in clinical care, with eight antidepressants ranked in the top 50 most prescribed drugs in 2018[111,112]. In addition to anti-depressive effects, antidepressants show anti-inflammatory effects on circulating biomarkers such as IL-6, CRP, and IL-2B[44,45]. Previous studies of immune markers and antidepressants were largely limited to examining common SSRIs, while the effects of other antidepressant classes on immune markers remain less explored. Additionally, only short follow-up times were used, leaving unknown whether anti-inflammatory effects persist throughout a longer period of time in an antidepressant trial.

In this chapter, we utilize electronic health records (EHRs) to examine the longitudinal effects of several antidepressant classes on WBC count. Patient records typically include multiple time points for laboratory values, providing a valuable resource to examine the effects of medication over time on a large sample size. In order to increase generalizability of our findings across all antidepressant users, we conducted primary analyses on all patients with antidepressant records with sensitivity analyses stratified by antidepressant indication. In order to maximize sample sizes and improve generalizability we assessed the effects of antidepressant classes, rather than individual drugs, on WBC count.

Methods

Study Sample

Vanderbilt University Medical Center (VUMC) is a tertiary care center that provides inpatient and outpatient care in Nashville, TN. The VUMC EHR was established in 1990 and includes data on billing codes from the International Classification of Diseases, 9th and 10th editions (ICD-9 and ICD-10), Current Procedural Terminology (CPT) codes, medications, laboratory values, reports, and clinical documentation. The de-identified mirror of the EHR, numbers more than 3.2 million patient records. For this study, we only included data recorded before January 1, 2017. This protocol was approved by the Vanderbilt University Medical Center institutional review board (#172020), and was deemed non-human subjects research because all information is de-identified.

Extracting and Cleaning Medication Names from EHRs

Antidepressant medications and dates associated with medication mentions were extracted from the "Problem list" document type within the EHR and mapped to generic names and classes (Table 19). Due to small sample size monoamine oxidase inhibitors (MAOIs) were excluded (N individuals = 1,127).

To validate our longitudinal modeling approach, we extracted two known anti-inflammatory medications, biologic immunosuppressants and chemotherapies. Biologic immunosuppressant names were extracted from the EHR "Problem list". To extract chemotherapy medications, medications were mapped to The United States Pharmacopeial

Convention (USP) terms and we extracted medications from the “Problem list” that mapped to the anti-neoplastic category.

As a negative control, we also evaluated the effect of contraceptives on WBC which has no known anti-inflammatory properties. All medications with “ethinyl estradiol” were extracted from the EHR documents “Problem list”, “Medications Known to be Prescribed For or Used by the Patient”, “Outpatient Rx Order Summary”, “Outpatient Visit – Obstetrics/Gynecology”, and “Gynecology Clinic Visit”.

Table 19. List of generic antidepressant medications by class extracted from electronic health records.

Drug Class	Generic Name
SSRI	citalopram
	escitalopram
	fluoxetine
	fluvoxamine
	paroxetine
	sertraline
SNRI	desvenlafaxine
	venlafaxine
	milnacipran
	levomilnacipran
	duloxetine
TCA	amitriptyline
	desipramine
	imipramine
	nortriptyline
	amoxapine
	trimipramine
	doxepin
	clomipramine
	maprotiline
	protriptyline
	mirtazapine
Atypical	bupropion
	nefazodone
	vilazodone
	vortioxetine
	trazodone
MAOI	selegiline
	tranylcypromine
	phenelzine

Extracting WBC Measurements

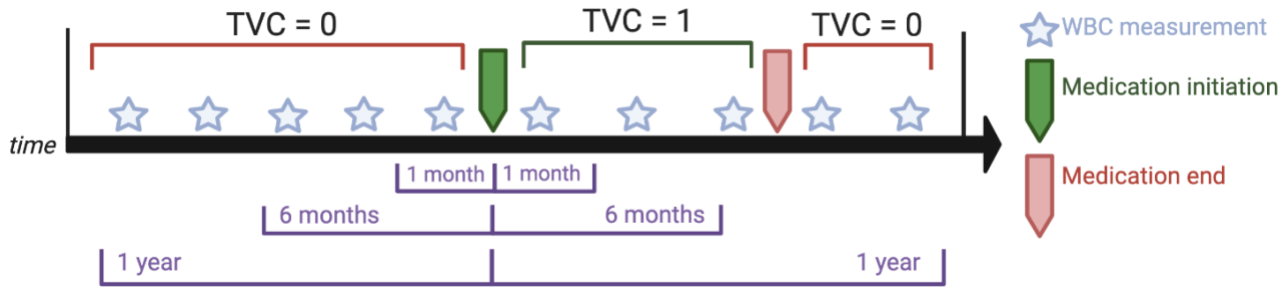
Laboratory values were extracted from the EHR and cleaned as previously described[113]. For analysis, lab values were required to occur on out-patient visit dates to ensure values were not associated with acute illness during emergency visits or in-patient hospitalizations. For statistical modeling, laboratory values were normalized using a rank-based inverse normal transformation[90].

Longitudinal Cohort Construction

In order to examine the acute and long-term effects of medications on WBC count, we constructed three versions of the longitudinal model. We examined acute effects using WBC measurements recorded 30 days before and after medication initiation. Long-term effects were examined using WBC measurements recorded 6 months and 1 year before and after medication initiation.

To construct each time cohort, we determined the time period an individual was on an medication by extracting medication initiation and end for each individual defined as the first and last medication date, respectively. Individuals were required to have WBC measurements at least 30, 180, or 365 days before and after medication initiation to be included in the 1-month, 6-month, or 1-year cohorts, respectively (Figure 18). Longitudinal cohorts were constructed separately for each antidepressant class, biologic immunosuppressants, chemotherapy, and contraceptives.

Figure 18. Creation of longitudinal cohorts and medication time-varying covariates.



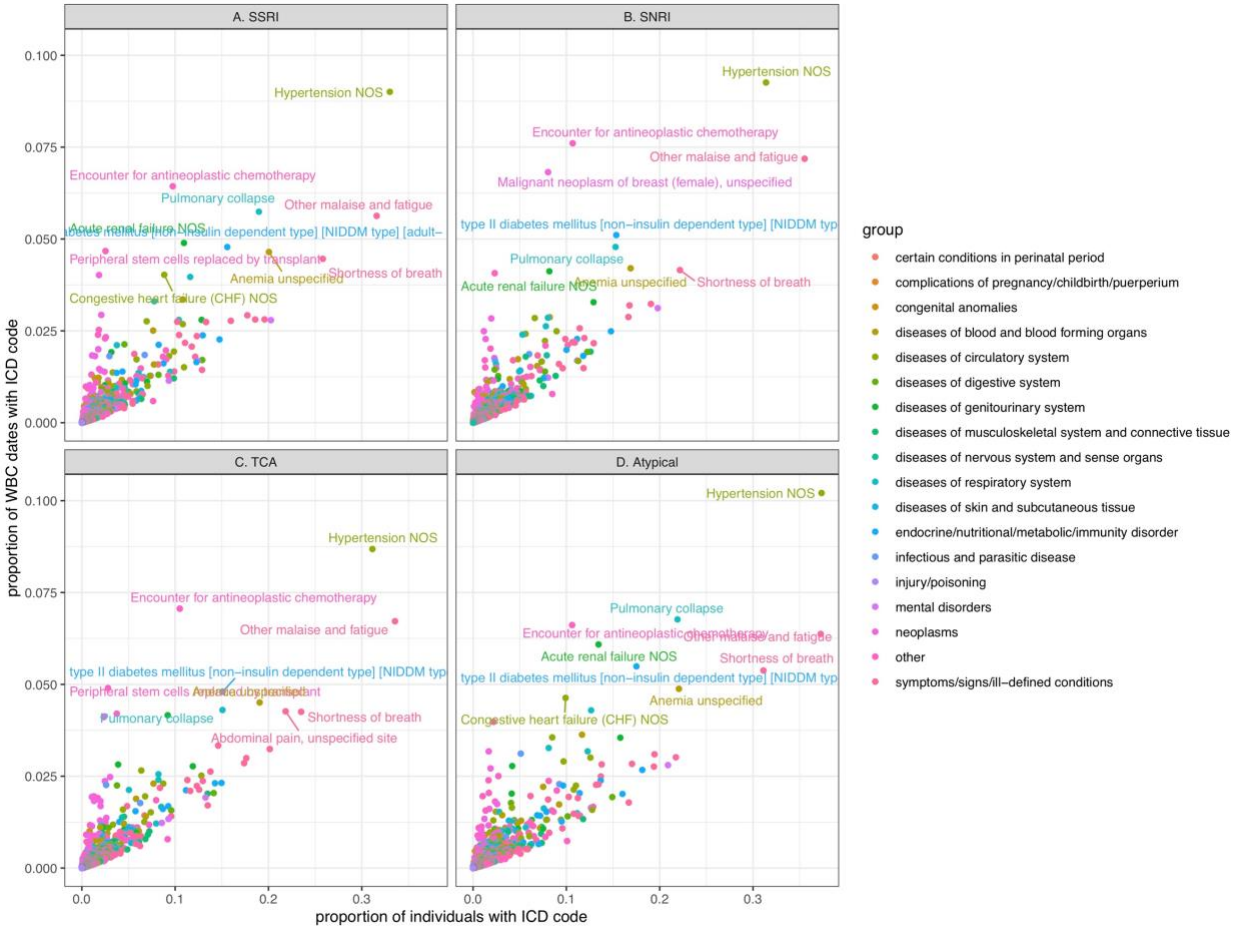
Diagnostic Co-Occurrence with WBC Measurements

WBC count is not measured uniformly on all patients and could be confounded by diagnosis. In order to identify potential confounders, we used the date of the WBC count test to identify the corresponding ICD code associated with the test order. We evaluated all co-occurrences within the 1-year cohort (inclusive of the 1-month and 6-month cohorts). We then calculated the proportion of WBC count laboratory tests that co-occurred with each observed ICD code and the proportion of individuals for which the WBC count laboratory test and ICD code co-occurred (Figure 19).

The co-occurrences proportions revealed an enrichment of hypertension, cancer, and chemotherapy related codes, indicating that hypertension, cancer, and chemotherapy were clear reasons for multiple WBC count measurements. To control for the effects of cancer and chemotherapy treatment on the WBC value and subsequent associations with antidepressants, we excluded individuals with any cancer or chemotherapy code in their EHR (CPT codes: 96360 – 96361; 96365 – 96379; 96401 – 96549; ICD9 codes: 140 – 239.9; ICD10 codes: C00 – D49.9).

The effect of a hypertension diagnosis on the association between antidepressants and WBC count was assessed in a separate analysis described below.

Figure 19. Co-occurrences between out-patient WBC measurements and ICD codes for A) SSRI, B) SNRI, C) TCA, and D) Atypical 1-year longitudinal cohorts. The x-axis represents the proportion of individuals in the cohort with a particular ICD codes and the y-axis represents the proportion of WBC measurements in the cohort with the ICD code.



Longitudinal Effects of Medications on WBC Count

A time-varying covariate (TVC) for medication use was constructed depending on whether the WBC count was measured before medication initiation (TVC=0), between the first and last medication dates (TVC=1), or after the last medication mention date (TVC=0) (Figure 18).

We evaluated the effect of the medication TVC on WBC count for each time cohort using a linear mixed model controlled for the main effects of sex, race, and age at measurement, and the random effect of age at measurement[99]. Models were assessed separately for each antidepressant class, biologic immunosuppressants, chemotherapy, and contraceptives. Because the contraceptive cohorts only contained females, sex was not included in the model.

To assess the effect of hypertension on the association between antidepressants and WBC count, we repeated the longitudinal models for antidepressants and added a time-varying covariate for first hypertension diagnosis. Hypertension diagnosis was defined by the presence of at least two hypertension ICD codes (401, 401.1, 401.9, 416.0, 459.3, 459.30, 459.31, 459.32, 459.39, 572.3, 997.91, I10, I27.0, K76.6, 010, R03.0). The time-varying covariate was based on the date of the first hypertension code.

For longitudinal models, statistical significance was determined using Bonferroni correction for multiple testing of $p=9.09 \times 10^{-4}$ (0.05/(55).We corrected for each lab and medication pair, but not for each cohort because the 6-month and 1-month cohorts were nested within the 1-year cohort, making them non-independent tests.

Longitudinal Effects of Antidepressants on the Complete Blood Count Panel

In the clinic, WBC count is measured as part of a complete blood count panel (CBC). To test the hypothesis that antidepressant longitudinal effects on WBC count are specific to this lab, we extracted CBC laboratory values including red blood cell distribution width (RDW), packed cell volume (PCV), mean platelet volume (MPV), mean corpuscular volume (MCV), mean corpuscular hemoglobin concentration (MCHC), and mean corpuscular hemoglobin (MCH). To assess whether antidepressants associated with labs outside of the CBC, we also extracted a non-CBC lab, creatinine. We created longitudinal cohorts for each lab as described for WBC count and repeated the longitudinal analyses using the CBC labs and creatinine.

Effect of Antidepressants on WBC Count Stratified by Indication

Antidepressants can be prescribed for a variety of indications including depression, anxiety, chronic pain, and insomnia[114]. We evaluated the effect of antidepressants on WBC count across three indications: depression, anxiety, and chronic pain. Insomnia was not assessed due to small sample size. To determine if the effects of antidepressants on WBC count persisted across all indications or were specific to a subset of patients, we repeated the longitudinal analyses between antidepressants in WBC count stratified by indication.

Depression cases were defined by the presence of a phecode for depression, major depressive disorder, adjustment reaction, or dysthymic disorder (296.2, 296.22, 304, and 300.4, respectively). Anxiety cases were defined the by presence of phecode for anxiety, 300. Chronic pain cases were defined as the presence of a phecode for migraine, chronic pain, myalgia,

osteoarthritis, rheumatoid arthritis, or pain in joint (340, 338.2, 770, 740, 714, and 745, respectively).

Effect of Antidepressants on WBC subtypes

WBC count is a measurement of five cellular subtypes, neutrophils, lymphocytes, monocytes, basophils, and eosinophils. The subtypes can be clinically measured using a WBC-differential test. To determine if a particular WBC subtype was driving the association between antidepressants and WBC, we repeated the longitudinal analyses on each WBC subtype. In order to more accurately reflect a WBC-differential clinical measurement, we required all differential measurements to occur on the same date. We extracted the absolute count values for each subtype, created longitudinal cohorts for each cell type and antidepressant, and evaluated the effect of antidepressant use on each cell type as described for the total WBC count.

Results

Study Sample

Across the entire VUMC EHR system, 377,611 (16.9%) individuals had documentation of any antidepressant. The most common class was SSRIs (N=177,101), followed by Atypicals (N=76,937), SNRIs (N=61,234), TCAs (N=47,918), and MAOIs (N=1,127). Due to the small sample size, MAOIs were not analyzed in this study. The majority of individuals on an antidepressant were female (67.1%) and EHR-reported white (85.0%). Among individuals with an antidepressant documented, 16% had a diagnosis of either major depressive disorder,

depression, or adjustment reaction, 14% had a diagnosis of anxiety, 4.7% had a diagnosis of chronic pain, and 4.2% had a diagnosis of migraine. Among individuals with an antidepressant recorded in their record, 59.6% also had at least one WBC measurement. Descriptions of the longitudinal samples stratified by antidepressant class and time cohorts can be found in Tables 20 & 21.

Table 20. Sample characteristics of antidepressant users in the VUMC EHR stratified by time cohort for out-patient WBC measurements.

Class	Time Cohort	N	N Female (%)	N White (%)	Length of record, years (IQR)	Number of ICD codes (IQR)	Age at first antidepressant, years (IQR)	Length of antidepressant trial, years (IQR)
SSRI	All	177,101	120,266 (67.9)	151,716 (85.7)	8.49 (2.7 - 13.39)	179.56 (25 - 196)	47.73 (33.78 - 61.63)	2.89 (0.34 - 4.18)
	1 year	4,390	2,843 (64.8)	3,680 (83.8)	7.75 (2.64 - 11.66)	165.7 (45.75 - 209.25)	46.39 (31.67 - 59.73)	2.18 (0.39 - 2.86)
	6 months	2,763	1,732 (62.7)	2,302 (83.3)	7.17 (2.16 - 10.72)	178.59 (46 - 221)	47.49 (33.16 - 60.89)	1.98 (0.31 - 2.57)
	1 month	521	310 (59.5)	436 (83.7)	6.46 (1.47 - 10.43)	229.7 (63 - 278)	49.32 (36.11 - 62.77)	1.68 (0.17 - 2.23)
SNRI	All	61,234	45,434 (74.2)	53,677 (87.7)	8.91 (3.2 - 13.82)	214.26 (29 - 240)	51.14 (41.12 - 61.53)	2.47 (0.31 - 3.51)
	1 year	1,483	1042 (70.3)	1,309 (88.3)	8.61 (3.5 - 12.99)	179.7 (61.25 - 223.5)	49.91 (39 - 61.16)	2.06 (0.37 - 2.67)
	6 months	840	564 (67.1)	733 (87.3)	7.88 (3.06 - 11.84)	187.91 (60 - 229.5)	50.14 (39.73 - 61.17)	1.95 (0.33 - 2.47)
	1 month	114	65 (57)	102 (89.5)	6.02 (1.43 - 9.5)	216.11 (68.25 - 266.75)	50.06 (39.7 - 62.86)	1.28 (0.13 - 1.62)
TCA	All	47,918	32,869 (68.6)	39,953 (83.4)	9.41 (3.58 - 14.48)	241.43 (31 - 274)	48.2 (35.37 - 62.3)	2.31 (0.25 - 3.05)
	1 year	1,252	852 (68.1)	1,028 (82.1)	8.43 (3.44 - 12.47)	176.5 (52 - 217.25)	47.62 (34.84 - 60.32)	1.84 (0.3 - 2.3)
	6 months	750	507 (67.6)	613 (81.7)	8.03 (2.95 - 11.89)	185.04 (54 - 225)	48.17 (36.97 - 60.23)	1.74 (0.26 - 2.05)
	1 month	120	71 (59.2)	100 (83.3)	7.04 (1.72 - 11.66)	219.22 (58.25 - 231.75)	48.7 (35.3 - 62.44)	1.38 (0.14 - 1.31)
Atypical	All	76,937	49,636 (64.5)	66,245 (86.1)	9.14 (3.39 - 14.12)	240.44 (35 - 277)	50.56 (38.58 - 62.86)	2.32 (0.27 - 3.13)
	1 year	2,213	1,336 (60.4)	1,850 (83.6)	8.08 (3.21 - 12.11)	211.27 (63 - 261.75)	49.86 (36.37 - 62.01)	1.72 (0.29 - 2.26)
	6 months	1,375	797 (58)	1,131 (82.3)	7.34 (2.57 - 10.99)	230.11 (65 - 292)	50.18 (36.49 - 62.03)	1.54 (0.25 - 1.79)
	1 month	237	110 (46.4)	202 (85.2)	6.15 (1.64 - 9.35)	335.44 (99 - 437)	51.94 (39.65 - 64.86)	1.35 (0.12 - 1.51)

Table 21. Description of indications for antidepressant users in VUMC EHR stratified by time cohort for out-patient WBC measurements.

Class	Time	N Depression Cases (%)	N Anxiety Cases (%)	N Chronic Pain Cases (%)	N Insomnia Cases (%)	N PTSD Cases (%)	N Tobacco Use Disorder Cases (%)	N Any Indication (%)
SSRI	All	39,124 (22.1)	34,184 (19.3)	47,956 (27.1)	10,137 (5.7)	3,843 (2.2)	14,190 (8)	77,718 (43.9)
	1 year	1,023 (23.3)	799 (18.2)	1,128 (25.7)	162 (3.7)	114 (2.6)	324 (7.4)	2,037 (46.4)
	6 months	605 (21.9)	433 (15.7)	670 (24.2)	83 (3)	69 (2.5)	214 (7.7)	1,181 (42.7)
	1 month	114 (21.9)	82 (15.7)	93 (17.9)	10 (1.9)	12 (2.3)	38 (7.3)	191 (36.7)
SNRI	All	15,525 (25.4)	11,980 (19.6)	23,270 (38)	4,960 (8.1)	1,604 (2.6)	5,855 (9.6)	3,1275 (51.1)
	1 year	374 (25.2)	280 (18.9)	645 (43.5)	99 (6.7)	34 (2.3)	129 (8.7)	876 (59.1)
	6 months	212 (25.2)	147 (17.5)	343 (40.8)	49 (5.8)	18 (2.1)	74 (8.8)	476 (56.7)
	1 month	21 (18.4)	19 (16.7)	30 (26.3)	8 (7)	0 (0)	12 (10.5)	46 (40.4)
TCA	All	9,218 (19.2)	7,948 (16.6)	19,803 (41.3)	4,260 (8.9)	1026 (2.1)	4,434 (9.3)	24,422 (51)
	1 year	226 (18.1)	184 (14.7)	524 (41.9)	83 (6.6)	24 (1.9)	96 (7.7)	659 (52.6)
	6 months	132 (17.6)	113 (15.1)	291 (38.8)	44 (5.9)	16 (2.1)	59 (7.9)	373 (49.7)
	1 month	22 (18.3)	15 (12.5)	30 (25)	5 (4.2)	4 (3.3)	7 (5.8)	43 (35.8)
Atypical	All	21,419 (27.8)	16,780 (21,.8)	26,325 (34.2)	7,959 (10.3)	2472 (3.2)	9,801 (12.7)	40,318 (52.4)
	1 year	629 (28.4)	474 (21.4)	738 (33.3)	171 (7.7)	64 (2.9)	271 (12.2)	1,212 (54.8)
	6 months	379 (27.6)	278 (20.2)	423 (30.8)	89 (6.5)	45 (3.3)	183 (13.3)	716 (52.1)
	1 month	57 (24.1)	43 (18.1)	67 (28.3)	12 (5.1)	10 (4.2)	40 (16.9)	108 (45.6)

Longitudinal Effects of Positive and Negative Control Medications on WBC

Biologic immunosuppressants and chemotherapy associated with decreases in WBC count in all longitudinal cohorts (Biologics: 1-month: p-value = 1.64×10^{-3}, beta = -0.27, SE = 0.09; 6-months: p-value = 4.01×10^{-22}, beta = -0.18, SE = 0.02; 1-year: p-value = 2.71×10^{-37}, beta = -0.19, SE = 0.01; Chemotherapy: 1-month: p-value = 3.60×10^{-5}, beta = -0.18, SE = 0.04; 6-months: p-value = 6.12×10^{-34}, beta = -0.24, SE = 0.02; 1-year: p-value = 3.31×10^{-47}, beta = -0.22, SE = 0.02). Contraceptive use did not associate with WBC count in any longitudinal cohort (Figure 20, Table 22).

Figure 20. Longitudinal associations between WBC and A) biologic immunosuppressants, B) chemotherapy, and C) oral contraceptives. Asterisks (*) indicate associations passing multiple testing correction (<9.09 x 10^{-4}).

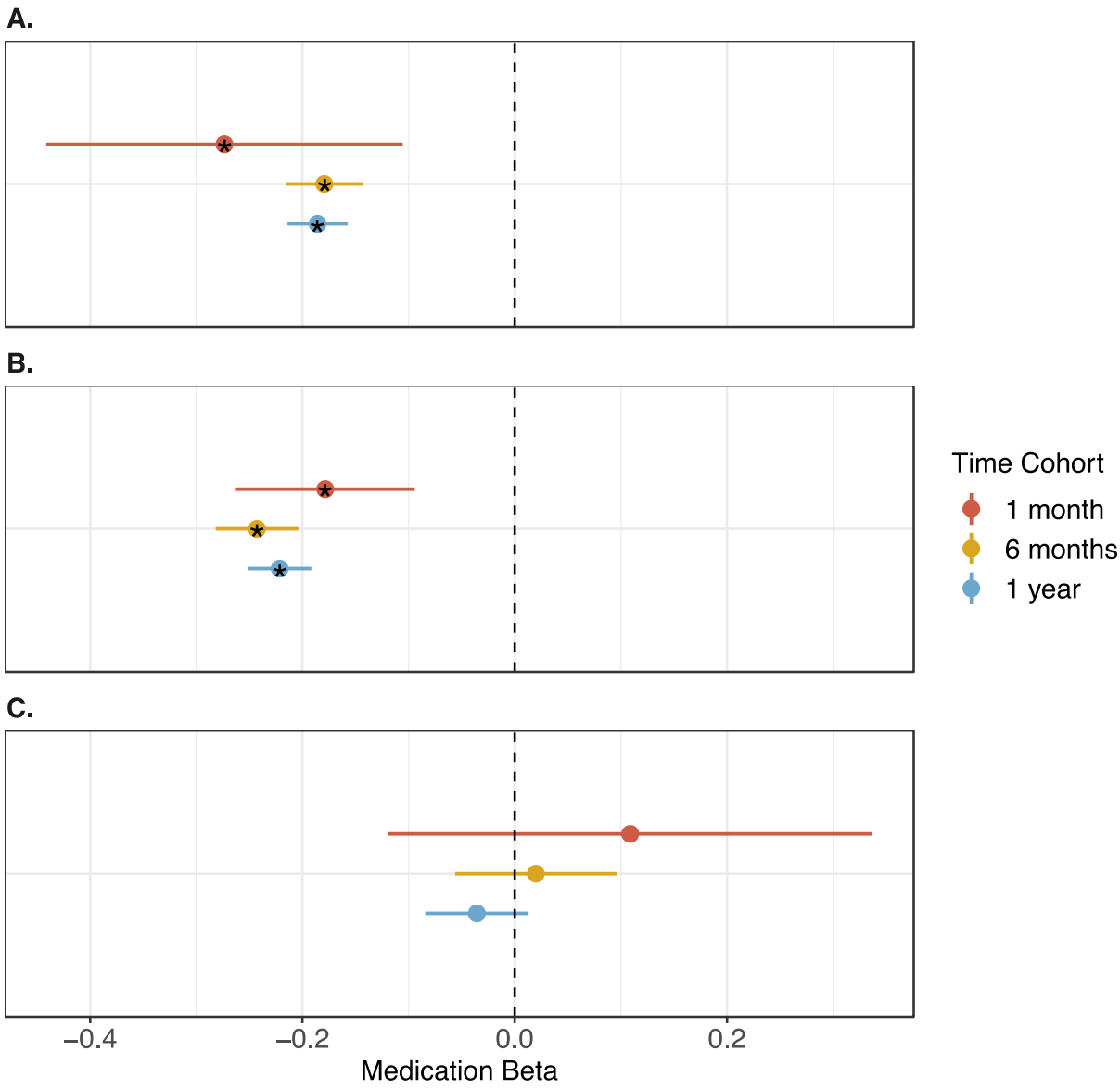


Table 22. Longitudinal effects of biologic immunosuppressants, chemotherapy, and contraceptives on white blood cell count.

Medication	Time	p-value	Beta	SE	N Individuals	N Observations
Biologic Immunosuppressants	1 month	1.65E-03	-0.273	0.086	77	285
	6 months	4.01E-22	-0.179	0.018	1,032	4,917
	1 year	2.71E-37	-0.186	0.014	1,269	8,449
Chemotherapy	1 month	3.60E-05	-0.178	0.043	211	895
	6 months	6.12E-34	-0.243	0.020	957	5,427
	1 year	3.31E-47	-0.221	0.015	1,328	9,343
Contraceptives	1 month	0.365	0.109	0.116	38	150
	6 months	0.609	0.020	0.039	311	1,382
	1 year	0.151	-0.036	0.025	713	3,365

Longitudinal Effects of Antidepressants on WBC

In the out-patient antidepressant exposed sample, only SSRI use and Atypical use associated with decreases in WBC levels in the 1-month cohorts (SSRI: p-value = 5.28×10^{-5}, beta = -0.11, SE = 0.03; Atypical: p-value = 7.77×10^{-8}, beta = -0.21, SE = 0.04). In the 6-month cohorts, all antidepressant classes associated with decreases in WBC count (SSRI: p-value = 4.43×10^{-19}, beta = -0.10, SE = 0.01; SNRI: p-value = 4.61×10^{-5}, beta = -0.09, SE = 0.02; TCA: p-value = 3.39×10^{-5}, beta = -0.09, SE = 0.02; Atypical: p-value = 4.13×10^{-12}, beta = -0.11, SE = 0.02). All antidepressants also associated with decreases in WBC count in the 1-year cohorts (SSRI: p-value = 2.19×10^{-36}, beta = -0.11, SE = 0.01; SNRI: p-value = 6.88×10^{-7}, beta = -0.08, SE = 0.02; TCA: p-value = 7.75×10^{-6}, beta = -0.08, SE = 0.02; Atypical: p-value = 1.12×10^{-17}, beta = -0.10, SE = 0.01) (Figure 21A, Table 23).

After controlling for hypertension diagnosis in the longitudinal models the pattern of significant results remained largely unchanged with the exception of an association between SNRI use and WBC count in the 6-month cohort that fell just shy of the multiple testing threshold (Figure 21). In the 1-month cohorts, SSRI and Atypical use associated with decreased WBC count (SSRI: p-value: 1.99×10^{-4}, beta = -0.11, SE = 0.03; Atypical: p-value: 1.92×10^{-7}, beta = -0.20, SE = 0.04). SSRI, TCA, and Atypical use associated with decreased WBC count in the 6-month cohorts (SSRI: p-value: 3.73×10^{-17}, beta = -0.10, SE = 0.01; TCA: p-value: 5.29×10^{-6}, beta = -0.10, SE = 0.02; Atypical: p-value: 1.12×10^{-11}, beta = -0.11, SE = 0.02). All antidepressant classes associated with decreased WBC count in the 1-year cohort (SSRI: p-value: 3.00×10^{-32}, beta = -0.11, SE = 0.01; SNRI: p-value: 4.99×10^{-6}, beta = 0.07, SE = 0.02; TCA: p-value: 4.89×10^{-6}, beta = -0.08, SE = 0.02; Atypical: p-value: 1.34×10^{-17}, beta = -0.10, SE = 0.01) (Figure 21B).

Figure 21. Longitudinal associations between antidepressant use and WBC values. A) Associations were controlled for fixed effects of sex, race, age, and the random effect of age. B) Associations were additionally controlled for a time-varying covariate for hypertension diagnosis. Asterisks (*) indicate associations passing multiple testing correction (<9.09 x 10^{-4}).

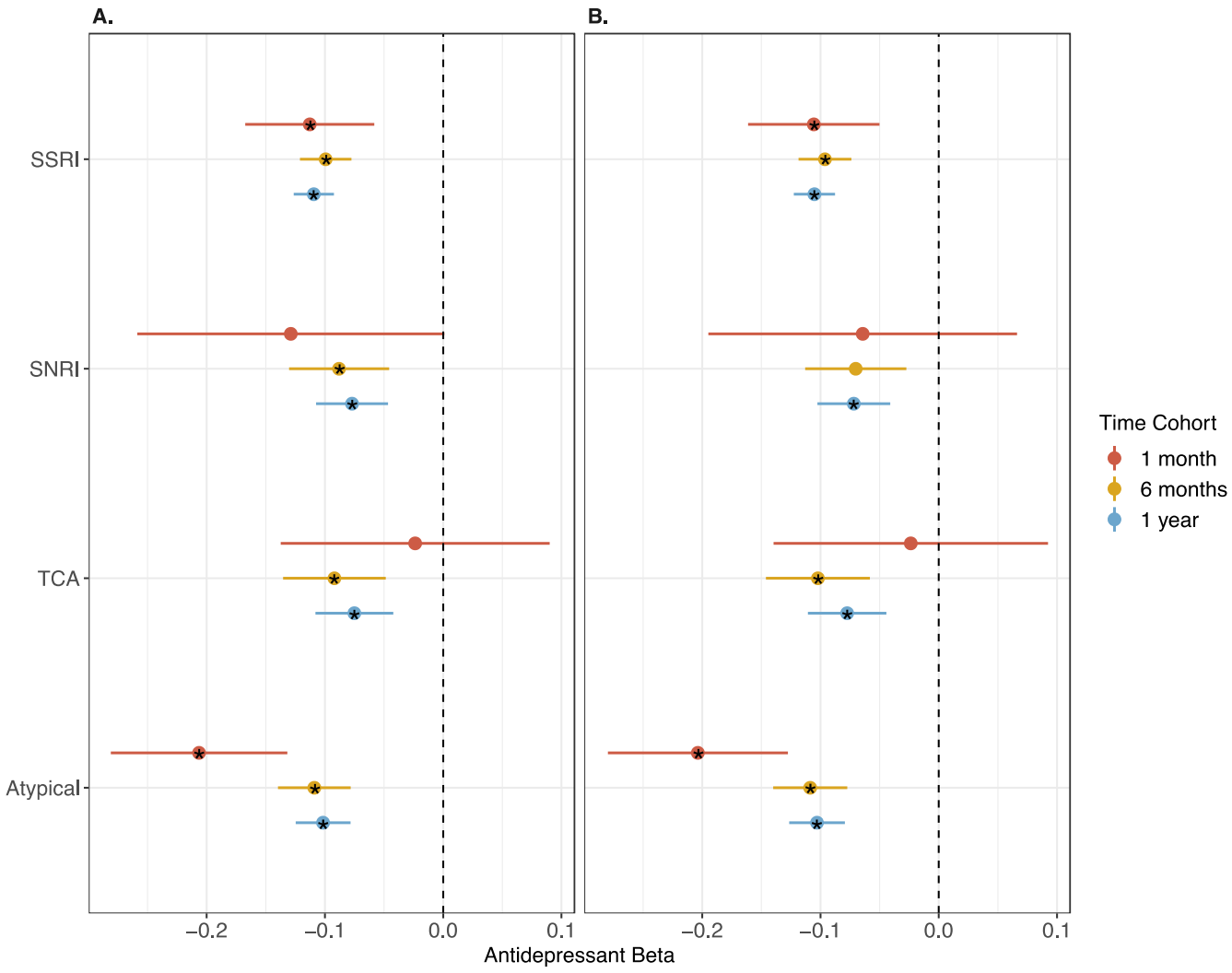


Table 23. Longitudinal effect of antidepressants on WBC count stratified by antidepressant class.

Class	Time Cohort	p-value	Beta	SE	N Observations	N Individuals
SSRI	1 month	5.28E-05	-0.113	0.028	2,080	521
	6 months	4.43E-19	-0.099	0.011	14,503	2,763
	1 year	2.19E-36	-0.109	0.009	26,040	4,390
SNRI	1 month	0.052	-0.129	0.066	443	114
	6 months	4.61E-05	-0.088	0.022	4,027	840
	1 year	6.88E-07	-0.077	0.016	8,404	1,483
TCA	1 month	0.683	-0.024	0.058	456	120
	6 months	3.39E-05	-0.092	0.022	3,669	750
	1 year	7.75E-06	-0.075	0.017	7,162	1,252
Atypical	1 month	7.77E-08	-0.206	0.038	1,249	237
	6 months	4.13E-12	-0.109	0.016	8,333	1,375
	1 year	1.12E-17	-0.102	0.012	14,892	2,213

Longitudinal Effects of Antidepressants on Complete Blood Count Panel Labs

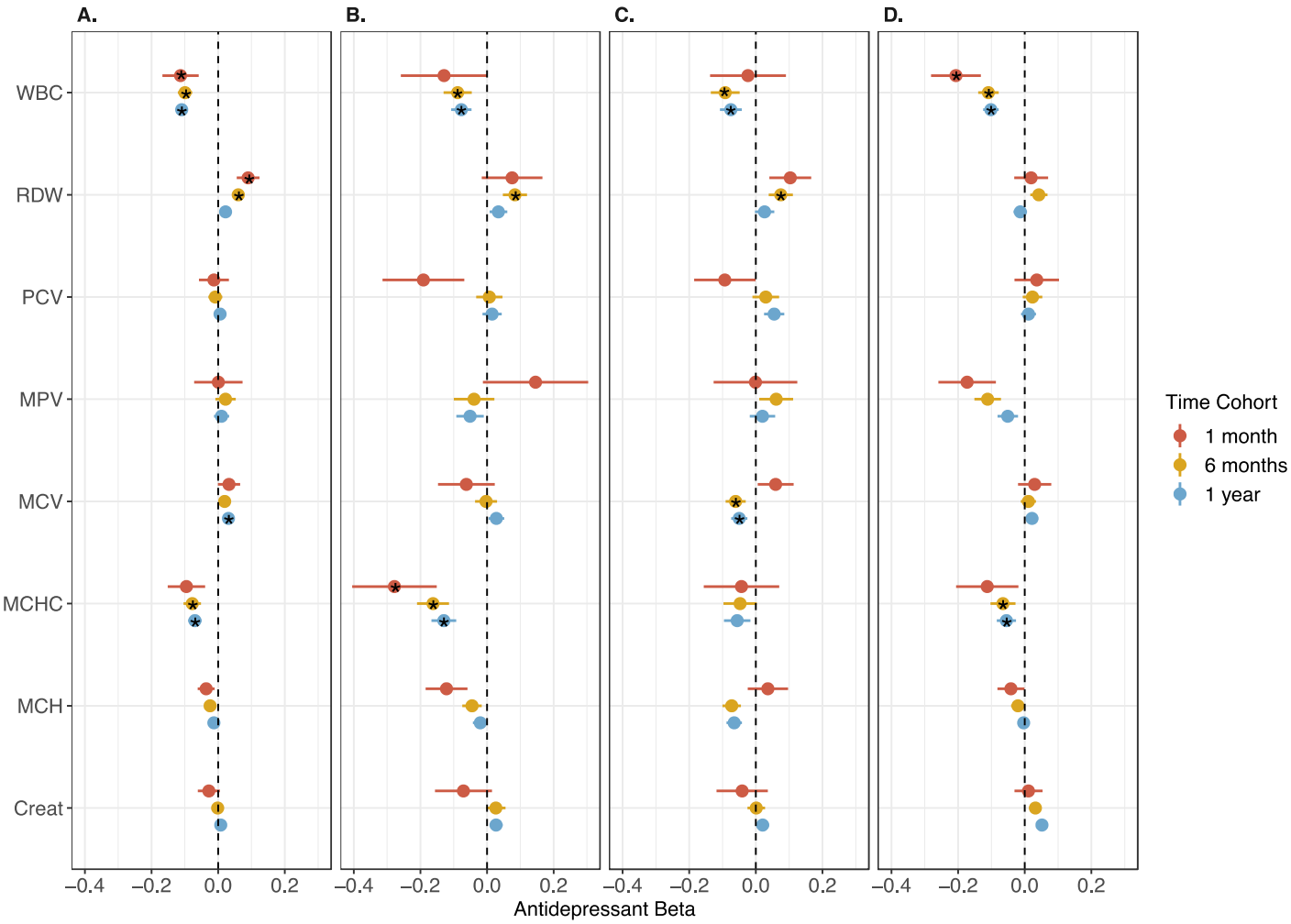
In the 1-month cohort SSRIs associated with increased RDW (p-value = 2.35×10^{-7}, beta = 0.09, SE = 0.02). In the 6-month cohort, SSRIs associated with decreased MCHC (p-value = 5.56×10^{-9}, beta = -0.08, SE = 0.01), and increased RDW (p-value = 2.34×10^{-10}, beta = 0.06, SE = 0.01). In the 1-year cohort, SSRIs associated with decreased MCHC (p-value = 1.10^{-11}, beta = -0.07, SE = 0.01) and increased MCV (p-value = 2.91×10^{-6}, beta = 0.03, SE = 0.01) (Figure 22A).

SNRI use associated with decreased MCHC in the 1-month, 6-months, and 1-year cohorts (1-month: p-value = 2.40×10^{-5}, beta = -0.28, SE = 0.06; 6-months: p-value = 4.63×10^{-11}, beta = -0.16, SE = 0.02; 1-year: p-value = 1.02×10^{-11}, beta = 0.13, SE = 0.02). Additionally, SNRI use associated with increased RDW in the 6-month cohort (p-value = 5.88×10^{-6}, beta = 0.08, SE = 0.02) (Figure 22B).

TCA use associated with increased in RDW in the 6-month cohort (p-value = 5.73×10^{-5}, beta = 0.07, SE = 0.02) and decreased MCV (p-value = 7.83×10^{-5}, beta = -0.06, SE = 0.02). In the 1-year cohort, TCA use associated with decreased MCV (p-value = 6.09×10^{-5}, beta = -0.05, SE = 0.01) (Figure 22C).

Atypical use associated with decreased MCHC in the 6-month cohort (p-value = 6.66×10^{-4}, beta = -0.07, SE = 0.02) and 1-year cohort (p-value = 1.99×10^{-4}, beta = -0.06, SE = 0.01) (Figure 22D).

Figure 22. Longitudinal associations between A) SSRI, B) SNRI, C) TCA and A) Atypical use and complete blood count and creatinine lab values. Asterisks (*) indicate associations passing multiple testing correction (<9.09 x 10^{-4}).



Effect of Antidepressants on WBC Count Stratified by Indication

In depression cases, SSRI and SNRI use associated with decreases in WBC count in the 1-year cohorts (SSRI: p-value = 1.88×10^{-6}, beta = -0.08, SE = 0.02; SNRI: p-value = 3.42×10^{-4}, beta = -0.10, SE = 0.03) (Figure 23A, Table 24).

In anxiety cases, SNRI use associated with decreases in WBC count in the 1-year cohort (p-value = 6.10×10^{-4}, beta = -0.11, SE = 0.03) and 6-month cohort (p-value = 2.94×10^{-4}, beta = -0.19, SE = 0.05). TCA use in the 1-month cohort associated with decreases in WBC count in anxiety cases (p-value = 4.09×10^{-4}, beta = -0.56, SE = 0.15) (Figure 23B, Table 24).

Among chronic pain cases, SSRI use associated with decreases in WBC count in the 1-year cohort (p-value = 8.42×10^{-5}, beta = -0.07, SE = 0.02). TCA use associated with decreases in WBC count in the 1-month and 6-month cohorts of chronic pain cases (1-month: p-value = 2.08×10^{-5}, beta = -0.49, SE = 0.11; 6-months: p-value = 1.25×10^{-5}, beta = -0.14, SE = 0.03). Additionally, Atypical use associated with decreases in WBC count in all time cohorts within chronic pain cases (1-month: p-value = 6.24×10^{-4}, beta = -0.27, SE = 0.08; 6-months: p-value = 6.30×10^{-11}, beta = -0.19, SE = 0.03; 1-year: p-value = 3.64×10^{-12}, beta = -0.14, SE = 0.02) (Figure 23C, Table 24).

TCA use in the 1-month cohort showed significant associations with decreased WBC count among anxiety and chronic pain cases and a nominally significant association among depression cases (p-value = 0.04) that was not present in the overall sample. To determine if there was a specific effect of TCAs in the 1-month cohort among individuals with documented psychiatric indications, we re-ran the overall analysis excluding individuals with an indication. In

this sample, TCA use was nominally associated with increased WBC count (p-value = 0.04, beta = 0.14, SE = 0.07), however, this association did not survive multiple testing correction.

Figure 23. Longitudinal associations between antidepressant use and WBC values stratified by antidepressant indications A) depression, B) anxiety, and C) chronic pain. Asterisks (*) indicate associations passing multiple testing correction (<9.09 x 10^{-4}).

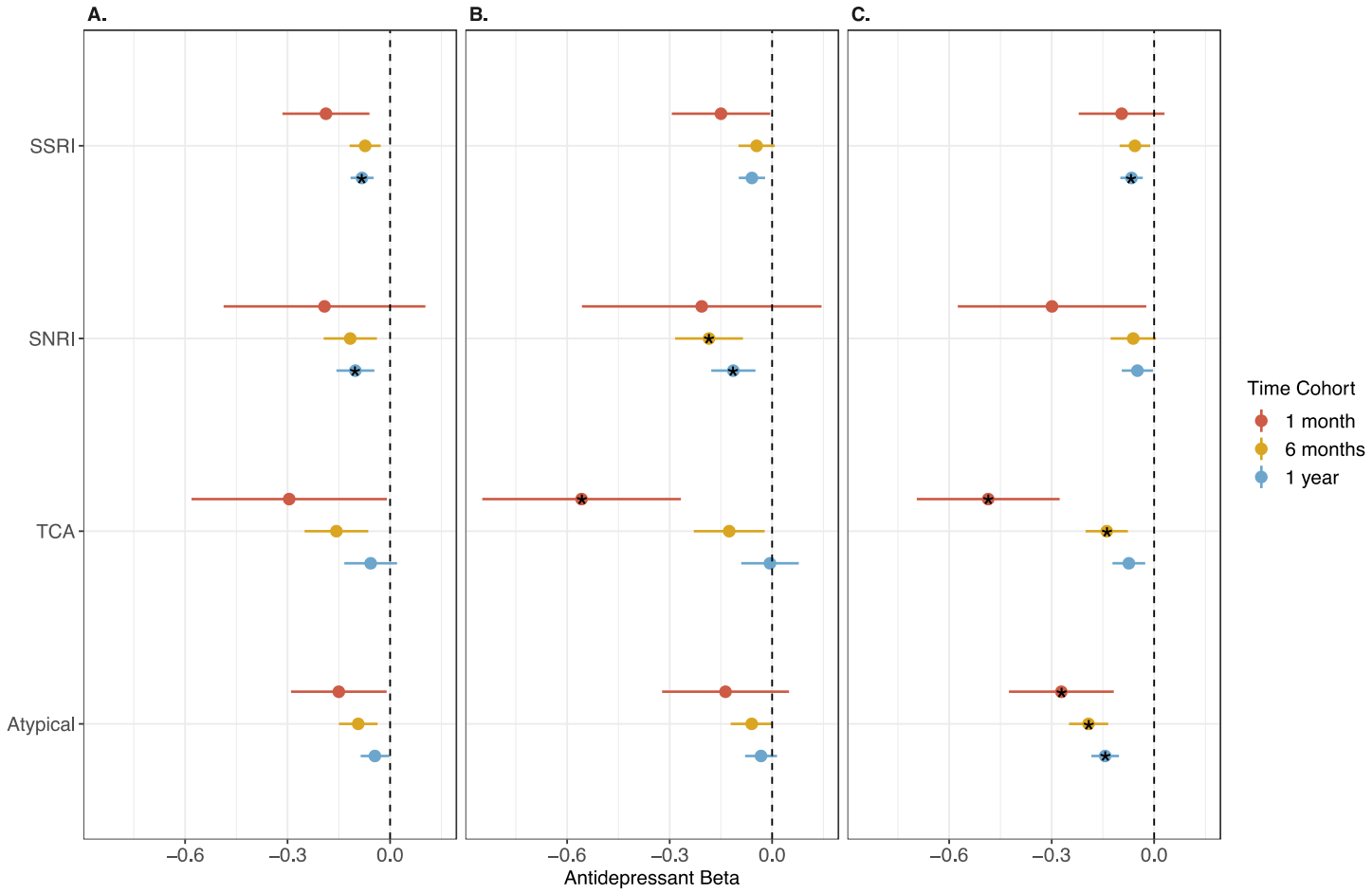


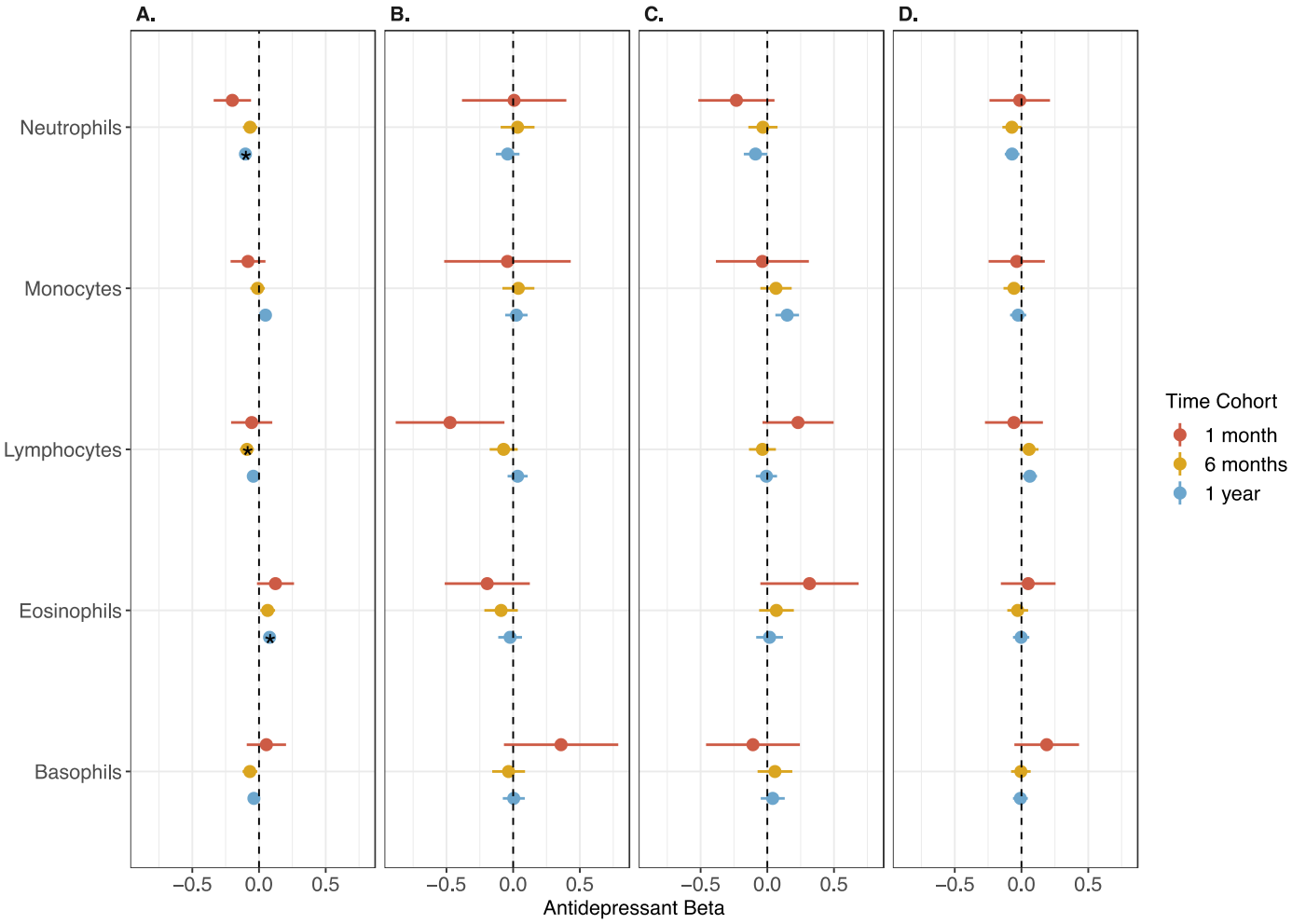
Table 24. Longitudinal associations between antidepressants and WBC count stratified by indication.

Indication	Class	Time Cohort	p-value	Beta	SE	N Observations	N Individuals
Depression	SSRI	1 month	4.12E-03	-0.188	0.065	488	114
		6 months	1.60E-03	-0.073	0.023	3,531	605
		1 year	1.88E-06	-0.082	0.017	6,616	1,023
	SNRI	1 month	0.209	-0.192	0.151	87	21
		6 months	3.38E-03	-0.117	0.040	1,287	212
		1 year	3.42E-04	-0.102	0.028	2,675	374
	TCA	1 month	0.045	-0.295	0.146	100	22
		6 months	9.87E-04	-0.157	0.048	854	132
		1 year	0.149	-0.057	0.039	1,615	226
	Atypical	1 month	0.037	-0.150	0.071	289	57
		6 months	1.25E-03	-0.093	0.029	2,473	379
		1 year	0.036	-0.045	0.021	4,704	629
Anxiety	SSRI	1 month	0.042	-0.150	0.073	369	82
		6 months	0.095	-0.045	0.027	2,501	433
		1 year	2.55E-03	-0.059	0.020	5,023	799
	SNRI	1 month	0.252	-0.206	0.179	70	19
		6 months	2.94E-04	-0.185	0.051	834	147
		1 year	6.10E-04	-0.114	0.033	1,836	280
	TCA	1 month	4.09E-04	-0.558	0.148	77	15
		6 months	0.018	-0.126	0.053	708	113
		1 year	0.885	-0.006	0.043	1,266	184
	Atypical	1 month	0.154	-0.136	0.095	193	43
		6 months	0.057	-0.060	0.031	1,750	278
		1 year	0.173	-0.032	0.024	3,351	474
Chronic Pain	SSRI	1 month	0.138	-0.095	0.064	350	93
		6 months	0.013	-0.056	0.023	3,383	670
		1 year	8.42E-05	-0.066	0.017	6,804	1,128
	SNRI	1 month	0.043	-0.299	0.141	102	30
		6 months	0.071	-0.061	0.034	1,509	343
		1 year	0.035	-0.049	0.023	3,467	645
	TCA	1 month	2.08E-05	-0.486	0.107	111	30
		6 months	1.25E-05	-0.139	0.032	1,458	291
		1 year	2.52E-03	-0.074	0.024	3,084	524
	Atypical	1 month	6.24E-04	-0.272	0.078	299	67
		6 months	6.30E-11	-0.192	0.029	2,360	423
		1 year	3.64E-12	-0.143	0.021	4,692	738

Effect of Antidepressants on WBC subtypes

SSRI use associated with decreases in neutrophil count in the 1-year cohort (p-value = 8.61×10^{-7}, beta = -0.10, SE = 0.02), increased eosinophil count in the 1-year cohort (p-value = 2.51×10^{-4}, beta = 0.08, SE = 0.02), and decreased lymphocyte count in the 6-month cohort (p-value = 7.78×10^{-4}, beta = -0.09, SE = 0.03). No other associations passed multiple testing correction (Figure 24).

Figure 24. Longitudinal associations between A) SSRI, B) SNRI, C) TCA, and D) Atypical use and absolute counts of WBC subtype values. Asterisks (*) indicate associations passing multiple testing correction (<9.09 x 10^{-4}).



Discussion

In this study, we utilized EHR data to conduct a large-scale investigation of the short-term and long-term effects of antidepressants on a clinical immune marker, WBC count. To validate our longitudinal modeling approach, we first assessed the effects of known anti-inflammatory medications, biologic immunosuppressants and chemotherapy, on WBC count. Additionally, we examined the effect of contraceptive use on WBC count, which has no known immunomodulatory effects. As expected, biologic immunosuppressants and chemotherapy associated with decreases in WBC count over time and contraceptives did not associate with a change in WBC count over time. These results confirm that our longitudinal modeling approach can detect known anti-inflammatory effects and does not result in associations with any tested medication.

All antidepressant classes exhibited anti-inflammatory effects on WBC count, consistent with previous studies of antidepressants and other immune markers[44]. The effect sizes of biologic immunosuppressants and chemotherapy on WBC count were roughly double the effect sizes of antidepressants on WBC count, suggesting antidepressants are not strong anti-inflammatories. Most previous studies on antidepressants and immune markers were limited to depression cases. However, because antidepressants are used to treat a variety of conditions, we included all individuals with evidence of antidepressants in our initial analyses. Our results indicate the anti-inflammatory effect of antidepressants extend to all users and is not specific to depression cases.

We also conducted indication stratified analyses in depression cases, anxiety cases, and chronic pain cases separately. The pattern of associations in the indication stratified cohorts

were largely similar to the entire sample, further suggesting an anti-inflammatory effect of antidepressants independent of diagnostic status. However, TCA use associated with a decrease in WBC count in the 1-month cohorts of the indication stratified samples and was not associated in the entire sample. After excluding individuals with documented psychiatric indications, TCA use in the 1-month cohort was nominally associated with increased WBC count, but did not pass multiple testing correction. Future analysis with larger sample sizes are required to determine if there is an indication-specific acute effect of TCAs on WBC count.

The associations between antidepressants and WBC count could be due to a specific effect on immune cells, or they could be a part of a larger array of effects of antidepressants on biomarkers. To distinguish these possibilities, we evaluated the effects of antidepressants on other biomarkers commonly measured alongside WBC count. Antidepressants showed a few associations with other labs, including increased RDW and decreased MCHC. Interestingly, RDW is closely correlated with other inflammatory markers, such as CRP[115] and can discriminate between inflammatory and non-inflammatory joint diseases[116]. RDW is also reported to increase with depression symptoms[117,118]. However, rather than balancing increased RDW levels, our results indicate that some antidepressant classes (SSRIs, SNRIs, and TCAs) associate with increased RDW levels. Additionally, MCHC was previously reported to be decreased with increased inflammation ("anemia of inflammation") and depression[119]. Even though little data is available for the association between antidepressants and MCHC, antidepressant use associates with decreased hemoglobin levels[120], consistent with our findings that antidepressant use associates with decreased MCHC. Our results raise the hypothesis that,

antidepressants may counteract the effects of inflammation through WBC count but have less of an impact on other inflammatory markers like RDW and MCHC.

WBC measurements are calculated from a sum of five cell subtypes. In the clinic, WBC subtypes can be measured using a WBC-differential. In order to determine if a particular subtype was driving the association between antidepressants and WBC count, we conducted longitudinal analyses between antidepressants and each WBC subtype. The only associations that emerged were between SSRIs and neutrophils in the 1-month and 1-year cohorts. However, WBC differentials are not measured as frequently as overall WBC count, limiting our power in the analyses.

Although EHRs provide a large-scale resource for investigating longitudinal effects, there are several notable limitations. First, medications were extracted from clinical records using a natural language processing system, MedEx[121]. Extracting medications from clinical notes rather than pharmacy records or patient-report introduces a degree of uncertainty on medication start and stop dates. Second, our cohort construction does not take into consideration medication adherence. However, both of these limitations would decrease our statistical power to see an association, making it possible the effect estimates are underestimated in our study. Third, antidepressants are often used as combinations between drug classes, which we did not take into consideration. Fourth, it is possible that the decrease in WBC count with antidepressant use is dose-dependent, however, we did not evaluate antidepressant dose in our study. Finally, lab tests used in this study were clinically derived and represent a range of health states. To account for this, we removed values greater than four standard deviations from the sample mean to remove implausible values and restricted to values measured in the

out-patient setting to remove values due to severe illness or hospitalization. We also conducted phenotypic filtering to remove patients with cancer or chemotherapy, which commonly co-occurred with antidepressant use and frequent WBC measurements in our sample.

Overall, this work contributes to the immunomodulatory knowledge of antidepressants and lays the foundation for understanding alternative therapeutic routes for antidepressants and biomarkers of antidepressant response.

CHAPTER V

CONCLUSIONS

Depression is a common psychiatric disorder that is consistently linked to an increased levels of circulating pro-inflammatory biomarkers. However, the biology underlying an activated immune system in depression remains unknown. Electronic health records linked to genetic information provide a powerful resource to investigate the genetic effects of complex disease on other traits. In this thesis, we develop a method to scan for associations across lab values, validate our findings using polygenic scores (PGS) of lipids and coronary artery disease, apply our method to depression PGS, and investigate the effects of antidepressant treatment on an associated immune marker.

Our method, lab-wide association scan (LabWAS), is a hypothesis-generating approach to find associations between a trait of interest and the full breadth of clinical lab values. In proof-of-principle analyses, lipid PGS strongly associated with their referent lipid and CAD PGS associated with known risk factors, including lipids and blood glucose. Interestingly, CAD PGS also associated with white blood cell count, an inflammatory marker not currently used to diagnose or monitor heart disease. This association remained after controlling for CAD diagnosis, indicating that CAD genetics could play a role in increasing inflammation. These results highlight the usefulness of analyzing polygenic scores which takes advantage of the entire patient population regardless of disease status. This approach offers a potential path

forward for the detection of novel biomarkers and for improved understanding of biomarker activity during the prodromal phase of disease.

We also show lipid-altering medications can influence the detection of risk biomarkers at the genetic level. This finding has important and complex implications for the clinical use of PGS recently discussed in the literature[122,123]. For example, as preventative treatments for complex diseases are adopted, the risk factors targeted by those treatments are less likely to play a role in the development of disease in current and future treated populations. Thus, today's PGS will no longer identify at-risk individuals in future generations who are routinely treated for risk factors which are only now being discovered. PGS, while incredibly valuable, provide only a snapshot of the human genetic profile of complex disease and thus are highly susceptible to these types of cohort effects in addition to other known sources of technical and experimental artifacts[124,125].

Unlike CAD, many complex diseases do not yet have *bona fide* biomarkers, but do have well-powered GWAS that can be used to mine large biobanks and identify quantitative labs which may be correlated, even weakly, with genetic risk for disease. In Chapter 3, we conducted a LabWAS of depression PGS to identify lab values associated with increased genetic liability for depression. We identified an association between depression PGS and WBC count that survived controlling for various comorbid phenotypes and replicated in a meta-analysis across three other independent biobanks in the PsycheMERGE Network. Our results suggest the pro-inflammatory state previously reported in depression is at least partially due to genetic factors, however, the pathway remains unclear.

There are two main models that connect depression to a pro-inflammatory state, the neuroinflammation model and the stress response model. The neuroinflammation model hypothesizes that an activated immune system contributes to risk of developing depression[126,127]. The stress response model proposes the stress of depression symptoms leads to a pro-inflammatory state[128,129]. Importantly, these two models are not mutually exclusive and some have suggested they form a feedback loop[130,131]. In support of this hypothesis, our mediation results do not distinguish either the neuroinflammation model or the stress response model as the exclusive pathway between depression and WBC.

While PGS are powerful tools for identifying associations with genetic liability, they are based on GWAS summary statistics that are typically unadjusted for phenotypic comorbidities. This approach is optimal in GWAS for many reasons, however, it introduces the possibility of "phenotypic hitchhiking" in which a comorbid trait is unintentionally selected during the ascertainment of the index trait. Thus, two heritable phenotypes that might share common environmental risk factors but no genetic risk factors can subsequently appear genetically correlated in PGS analysis, even in independent samples. We therefore emphasize that the genetic approach presented in this thesis is still fundamentally correlational.

Antidepressant treatment has previously been associated with decreasing circulating immune markers, suggesting antidepressant therapy balances the pro-inflammatory state seen in depression. Our results indicate antidepressant therapy associates with decreases in WBC count up to a year after antidepressant initiation, which is a notable increase compared to the 6-8 weeks previously described with other immune markers. The association between antidepressant use and decreased WBC persisted across all antidepressant users as well as

when stratified to individuals with known indications, suggesting a common anti-inflammatory mechanism of antidepressants rather than a specific action among depression cases.

Therapeutic response to antidepressants is thought to be due to an increase in neurotransmitter signaling in the brain. The decrease in pro-inflammatory markers with antidepressant treatment suggests anti-inflammatory action may offer an alternative therapeutic pathway of antidepressants, however, the previous studies on antidepressant response and inflammatory markers are mixed. Interestingly, our results show decreases in WBC count begin shortly after antidepressant initiation (within 1 month), indicating anti-inflammatory action may precede mood changes from antidepressants which typically occur 4-6 weeks after initiation. Although not examined here, EHRs provide extensive information on treatment response through clinical questionnaires such as the patient health questionnaire (PHQ). Future studies aimed at extracting the PHQ from EHRs would provide substantially larger sample sizes and longer time frames to assess the association between inflammatory markers and antidepressant response.

EHR-linked biobanks have several strengths for genetic research, including large sample sizes, long timeframes, and recruitment of participants across diagnostic status. However, there are notable limitations. First, lab values are not uniformly measured on all patients which can create confounding by indication with lab measurements. However, the lab of interest in this thesis, WBC count, is commonly measured on a wide variety of patients, decreasing the impact of indication bias. Additionally, medications were extracted from clinical records using a natural language processing system, MedEx[121]. Extracting medications from clinical notes rather than pharmacy records or patient-report introduces a degree of uncertainty on medication start and

stop dates. Additionally, our cohort construction does not take into consideration medication adherence. However, both of these limitations would decrease our statistical power to observe an association, resulting in underestimation of true effect sizes in our study.

In summary, our results provide a basis for future investigation of genes contributing to the pro-inflammatory state in depression, identification of a pro-inflammatory subtype of depression, and clinical panels for to inform antidepressant response.

REFERENCES

1. James, S. L. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1789–1858 (2018).
2. Brown, D. W., Giles, W. H. & Croft, J. B. White blood cell count: An independent predictor of coronary heart disease mortality among a national cohort. *J. Clin. Epidemiol.* **54**, 316–322 (2001).
3. Hare, D. L., Toukhsati, S. R., Johansson, P. & Jaarsma, T. Depression and cardiovascular disease: A clinical review. *Eur. Heart J.* **35**, 1365–1372 (2014).
4. Musselman, D. L., Evans, D. L. & Nemeroff, C. B. The relationship of depression to cardiovascular disease: Epidemiology, biology, and treatment. *Arch. Gen. Psychiatry* **55**, 580–592 (1998).
5. Benros, M. E. *et al.* Autoimmune diseases and severe infections as risk factors for mood disorders a nationwide study. *JAMA Psychiatry* **70**, 812–820 (2013).
6. Anderson, R., Freedland, K., RE, C. & Lustman, P. J. The Prevalence of Comorbid Depression. *Diabetes Care* **24**, 1069–1078 (2001).
7. Gkrania-Klotsas, E. *et al.* Differential white blood cell count and type 2 diabetes: Systematic review and meta-analysis of cross-sectional and prospective studies. *PLoS One* **5**, (2010).
8. Mezuk, B., Eaton, W. W., Albrecht, S. & Golden, S. H. Depression and type 2 diabetes

over the lifespan: A meta-analysis. *Diabetes Care* **31**, 2383–2390 (2008).

9. Shim, W. S. *et al.* The association of total and differential white blood cell count with metabolic syndrome in type 2 diabetic patients. *Diabetes Res. Clin. Pract.* **73**, 284–291 (2006).

10. Horwitz, A. V., Wakefield, J. C. & Lorenzo-Luaces, L. History of depression. *Oxford Handb. Mood Disord.* 11–23 (2015). doi:10.1093/oxfordhb/9780199973965.013.2

11. Telles-Correia, D. & Marques, J. G. Melancholia before the twentieth century: Fear and sorrow or partial insanity? *Front. Psychol.* **6**, 1–4 (2015).

12. Ventriglio, A. & Bhugra, D. Descartes' dogma and damage to Western psychiatry. *Epidemiol. Psychiatr. Sci.* **24**, 368–370 (2014).

13. Chang, C. K. *et al.* Life expectancy at birth for people with serious mental illness and other major disorders from a secondary mental health care case register in London. *PLoS One* **6**, (2011).

14. Nordentoft, M. *et al.* Excess Mortality, Causes of Death and Life Expectancy in 270,770 Patients with Recent Onset of Mental Disorders in Denmark, Finland and Sweden. *PLoS One* **8**, (2013).

15. Iwata, M., Ota, K. T. & Duman, R. S. The inflammasome: Pathways linking psychological stress, depression, and systemic illnesses. *Brain. Behav. Immun.* **31**, 105–114 (2013).

16. Andersson, N. W. *et al.* Depression and the risk of autoimmune disease: a nationally representative, prospective longitudinal study. *Psychol. Med.* **45**, 3559–3569 (2015).

17. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat.*

Neurosci. **22**, 343–352 (2019).

18. Eijsbouts, C. *et al.* Genome-wide analysis of 53,400 people with irritable bowel syndrome highlights shared genetic pathways with mood and anxiety disorders. *Nat. Genet.* **53**, 1543–1552 (2021).
19. Strimbu, K. & Tavel, J. A. What are biomarkers? *Curr. Opin. HIV AIDS* **5**, (2010).
20. Luan, Y. Y. & Yao, Y. M. The clinical significance and potential role of C-reactive protein in chronic inflammatory and neurodegenerative diseases. *Front. Immunol.* **9**, 1–8 (2018).
21. Maes, M. Evidence for an immune response in major depression: A review and hypothesis. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **19**, 11–38 (1995).
22. Beydoun, M. A. *et al.* White blood cell inflammatory markers are associated with depressive symptoms in a longitudinal study of urban adults. *Transl. Psychiatry* **6**, (2016).
23. Wium-Andersen, M. K., Ørsted, D. D., Nielsen, S. F. & Nordestgaard, B. G. Elevated C-reactive protein levels, psychological distress, and depression in 73131 individuals. *JAMA Psychiatry* **70**, 176–184 (2013).
24. Zorrilla, E. P. *et al.* The relationship of depression and stressors to immunological assays: A meta-analytic review. *Brain. Behav. Immun.* **15**, 199–226 (2001).
25. Haapakoski, R., Mathieu, J., Ebmeier, K. P., Alenius, H. & Kivimäki, M. Cumulative meta-analysis of interleukins 6 and 1β, tumour necrosis factor α and C-reactive protein in patients with major depressive disorder. *Brain. Behav. Immun.* **49**, 206–215 (2015).
26. Valkanova, V., Ebmeier, K. P. & Allan, C. L. CRP, IL-6 and depression: A systematic review and meta-analysis of longitudinal studies. *J. Affect. Disord.* **150**, 736–744 (2013).
27. Dowlati, Y. *et al.* A Meta-Analysis of Cytokines in Major Depression. *Biol. Psychiatry* **67**,

446–457 (2010).

28. Liu, Y., Ho, R. C. M. & Mak, A. Interleukin (IL)-6, tumour necrosis factor alpha (TNF-α) and soluble interleukin-2 receptors (sIL-2R) are elevated in patients with major depressive disorder: A meta-analysis and meta-regression. *J. Affect. Disord.* **139**, 230–239 (2012).

29. Bell, J. A. *et al.* Repeated exposure to systemic inflammation and risk of new depressive symptoms among older adults. *Transl. Psychiatry* **7**, 1–8 (2017).

30. Khandaker, G. M., Pearson, R. M., Zammit, S., Lewis, G. & Jones, P. B. Association of serum interleukin 6 and C-reactive protein in childhood with depression and psychosis in young adult life a population-based longitudinal study. *JAMA Psychiatry* **71**, 1121–1128 (2014).

31. Kiecolt-Glaser, J. K., Derry, H. M. & Fagundes, C. P. Inflammation: Depression fans the flames and feasts on the heat. *Am. J. Psychiatry* **172**, 1075–1091 (2015).

32. Copeland, W. E., Shanahan, L., Worthman, C., Angold, A. & Costello, E. J. Cumulative depression episodes predict later C-reactive protein levels: A prospective analysis. *Biol. Psychiatry* **71**, 15–21 (2012).

33. Harrison, N. A. *et al.* Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity. *Biol. Psychiatry* **66**, 407–414 (2009).

34. Strike, P. C., Wardle, J. & Steptoe, A. Mild acute inflammatory stimulation induces transient negative mood. *Journal of Psychosomatic Research* **57**, 189–194 (2004).

35. Capuron, L., Ravaud, A. & Dantzer, R. Early depressive symptoms in cancer patients receiving interleukin 2 and/or interferon alfa-2b therapy. *J. Clin. Oncol.* **18**, 2143–2151

(2000).

36. Denicoff, K. *et al.* The Neuropsychiatric Effects of Treatment with Interleukin-2 and Lymphokine-Activated Killer Cells. *Ann. Intern. Med.* **107**, 293–300 (1987).

37. Renault, P. F. *et al.* Psychiatric Complications of Long-term Interferon Alfa Therapy. *Arch. Intern. Med.* **147**, 1577–1580 (1987).

38. Köhler, O. *et al.* Effect of Anti-inflammatory Treatment on Depression, Depressive Symptoms, and Adverse Effects: A Systematic Review and Meta-analysis of Randomized Clinical Trials. *JAMA Psychiatry* **71**, 1381–1391 (2014).

39. Kappelmann, N., Lewis, G., Dantzer, R., Jones, P. B. & Khandaker, G. M. Antidepressant activity of anti-cytokine treatment: A systematic review and meta-analysis of clinical trials of chronic inflammatory conditions. *Mol. Psychiatry* **23**, 335–343 (2018).

40. Wittenberg, G. M. *et al.* Effects of immunomodulatory drugs on depressive symptoms: A mega-analysis of randomized, placebo-controlled clinical trials in inflammatory disorders. *Mol. Psychiatry* **25**, 1275–1285 (2020).

41. Raison, C. L. *et al.* A Randomized Controlled Trial of the Tumor Necrosis Factor Antagonist Infliximab for Treatment-Resistant Depression: The Role of Baseline Inflammatory Biomarkers. *JAMA Psychiatry* **70**, 31–41 (2013).

42. Lopez-Munoz, F. & Alamo, C. Monoaminergic Neurotransmission: The History of the Discovery of Antidepressants from 1950s Until Today. *Curr. Pharm. Des.* **15**, 1563–1586 (2009).

43. Wong, D. T., Perry, K. W. & Bymaster, F. P. The Discovery of Fluoxetine Hydrochloride (Prozac). *Nat. Rev. Drug Discov.* **4**, 764–774 (2005).

44. Hannestad, J., Dellagioia, N. & Bloch, M. The effect of antidepressant medication treatment on serum levels of inflammatory cytokines: A meta-analysis. *Neuropsychopharmacology* **36**, 2452–2459 (2011).

45. Hiles, S. A., Baker, A. L., De Malmanche, T. & Attia, J. Interleukin-6, C-reactive protein and interleukin-10 after antidepressant treatment in people with depression: A meta-analysis. *Psychological Medicine* (2012). doi:10.1017/S0033291712000128

46. Canan, F. Effect of Escitalopram on White Blood Cells in Patients With Major Depression. *J. Clin. Med. Res.* (2009). doi:10.4021/jocmr2009.12.1275

47. Strawbridge, R. *et al.* Inflammation and clinical response to treatment in depression: A meta-analysis. *Eur. Neuropsychopharmacol.* **25**, 1532–1543 (2015).

48. McIntosh, A. M., Sullivan, P. F. & Lewis, C. M. Uncovering the Genetic Architecture of Major Depression. *Neuron* **102**, 91–103 (2019).

49. Shadrina, M., Bondarenko, E. A. & Slominsky, P. A. Genetics Factors in Major Depression Disease . *Frontiers in Psychiatry* **9**, 334 (2018).

50. Camp, N. J. & Cannon-Albright, L. A. Dissecting the genetic etiology of major depressive disorder using linkage analysis. *Trends Mol. Med.* **11**, 138–144 (2005).

51. Levinson, D. F. The Genetics of Depression: A Review. *Biol. Psychiatry* **60**, 84–92 (2006).

52. Border, R. *et al.* No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *Am. J. Psychiatry* **176**, 376–387 (2019).

53. Cai, N. *et al.* Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).

54. Levey, D. F. *et al.* Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat. Neurosci.* **24**, (2021).

55. Levey, D. F. *et al.* Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat. Neurosci.* (2021). doi:10.1038/s41593-021-00860-2

56. Gaspar, H. A. *et al.* Using genetic drug-target networks to develop new drug hypotheses for major depressive disorder. *Transl. Psychiatry* **9**, (2019).

57. Wray, N. R. *et al.* From Basic Science to Clinical Application of Polygenic Risk Scores. *JAMA Psychiatry* (2020). doi:10.1001/jamapsychiatry.2020.3049

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

59. Andersen, A. M. *et al.* Polygenic Scores for Major Depressive Disorder and Risk of Alcohol Dependence. *JAMA psychiatry* **74**, 1153–1160 (2017).

60. Mullins, N. *et al.* GWAS of suicide attempt in psychiatric disorders and association with major depression polygenic risk scores. *Am. J. Psychiatry* **176**, 651–660 (2019).

61. Amare, A. T. *et al.* Association of polygenic score for major depression with response to lithium in patients with bipolar disorder. *Mol. Psychiatry* **26**, 2457–2470 (2021).

62. Dennis, J. *et al.* Genetic risk for major depressive disorder and loneliness in sex-specific associations with coronary artery disease. *Mol. Psychiatry* **26**, 4254–4264 (2021).

63. McCoy, T. H. *et al.* Polygenic loading for major depression is associated with specific

medical comorbidity. *Transl. Psychiatry* **7**, 1–6 (2017).

64. Kappelmann, N. *et al.* Dissecting the association between inflammation, metabolic dysregulation, and specific depressive symptoms: A genetic correlation and 2-sample mendelian randomization study. *JAMA Psychiatry* **78**, 161–170 (2020).

65. Smoller, J. W. The use of electronic health records for psychiatric phenotyping and genomics. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **177**, 601–612 (2018).

66. Shameer, K. *et al.* A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. Genet.* **133**, 95–109 (2014).

67. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0064-5

68. Verma, A. *et al.* PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *Am. J. Hum. Genet.* (2018). doi:10.1016/j.ajhg.2018.02.017

69. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).

70. Verma, A. *et al.* Integrating clinical laboratory measures and ICD-9 code diagnoses in phenome-wide association studies. in *Pacific Symposium on Biocomputing* (2016). doi:10.1142/9789814749411_0016

71. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* (2008). doi:10.1038/clpt.2008.89

72. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based

linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

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

74. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).

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

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

77. Davis, L. K. *et al.* Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture. *PLoS Genet.* (2013). doi:10.1371/journal.pgen.1003864

78. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

79. Zeng, P. *et al.* Statistical analysis for genome-wide association study. *Journal of Biomedical Research* (2015). doi:10.7555/JBR.29.20140007

80. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *bioRxiv* 598110 (2019). doi:10.1101/598110

81. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1285 (2013).

82. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* (2015). doi:10.1038/ng.3211

83. Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0108-x

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

85. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across human complex traits. *Nat. Genet.* (2020). doi:10.1038/s41588-020-0653-y

86. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1–10 (2019).

87. Nikpay, M. *et al.* A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).

88. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, 1–6 (2019).

89. Karlson, E. W., Boutin, N. T., Hoffnagle, A. G. & Allen, N. L. Building the partners healthcare biobank at partners personalized medicine: Informed consent, return of research results, recruitment lessons and operational considerations. *J. Pers. Med.* **6**, 1–11 (2016).

90. McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* (2019). doi:10.1111/biom.13214

91. Casey, J. A., Schwartz, B. S., Stewart, W. F. & Adler, N. E. Using Electronic Health Records

for Population Health Research: A Review of Methods and Applications. *Annual Review of Public Health* (2016). doi:10.1146/annurev-publhealth-032315-021353

92. Zheutlin, A. *et al.* VALIDATION OF PSYCHIATRIC POLYGENIC RISK SCORES ACROSS THREE HEALTHCARE SYSTEMS USING ELECTRONIC HEALTH RECORDS. *Eur. Neuropsychopharmacol.* (2019). doi:10.1016/j.euroneuro.2018.07.068

93. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

94. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).

95. Dennis, J. K. *et al.* Lab-wide association scan of polygenic scores identifies biomarkers of complex disease. *medRxiv* 2020.01.24.20018713 (2020). doi:10.1101/2020.01.24.20018713

96. Karczewski, J. *et al.* Obesity and inflammation. *Eur. Cytokine Netw.* **29**, 83–94 (2018).

97. Shamai, L. *et al.* Association of body mass index and lipid profiles: Evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obes. Surg.* **21**, 42–47 (2011).

98. Luppino, F. S. *et al.* Overweight, Obesity, and Depression: A Systematic Review and Meta-analysis of Longitudinal Studies. *Arch. Gen. Psychiatry* **67**, 220–229 (2010).

99. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, (2015).

100. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. Mediation: R package for causal mediation analysis. *J. Stat. Softw.* **59**, 1–38 (2014).

101. Beasley, T. M., Erickson, S. & Allison, D. B. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav. Genet.* (2009). doi:10.1007/s10519-009-9281-0

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

103. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).

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

105. Bot, M. *et al.* Metabolomics Profile in Depression: A Pooled Analysis of 230 Metabolic Markers in 5283 Cases With Depression and 10,145 Controls. *Biol. Psychiatry* (2020). doi:10.1016/j.biopsych.2019.08.016

106. Otte, C. *et al.* Major depressive disorder. *Nat. Rev. Dis. Prim.* **2**, 1–21 (2016).

107. Osimo, E. F., Baxter, L. J., Lewis, G., Jones, P. B. & Khandaker, G. M. Prevalence of low-grade inflammation in depression: A systematic review and meta-Analysis of CRP levels. *Psychological Medicine* (2019). doi:10.1017/S0033291719001454

108. McNally, L., Bhagwagar, Z. & Hannestad, J. Inflammation, glutamate, and glia in depression: A literature review. *CNS Spectrums* (2008). doi:10.1017/S1092852900016734

109. Amulic, B., Cazalet, C., Hayes, G. L., Metzler, K. D. & Zychlinsky, A. Neutrophil function: From mechanisms to disease. *Annual Review of Immunology* (2012). doi:10.1146/annurev-immunol-020711-074942

110. Manda-Handzlik, A. & Demkow, U. The Brain Entangled: The Contribution of Neutrophil

Extracellular Traps to the Diseases of the Central Nervous System. *Cells* **8**, 1477 (2019).

111. Mojtabai, R. & Olfson, M. Proportion of antidepressants prescribed without A psychiatric diagnosis is growing. *Health Aff.* (2011). doi:10.1377/hlthaff.2010.1024

112. Fuentes, A., Pineda, M. & Venkata, K. Comprehension of Top 200 Prescribed Drugs in the US as a Resource for Pharmacy Teaching, Training and Practice. *Pharmacy* **6**, 43 (2018).

113. Dennis, J. K. *et al.* Clinical laboratory test-wide association scan of polygenic scores identifies biomarkers of complex disease. *Genome Med.* (2021). doi:10.1186/s13073-020-00820-8

114. Wong, J. *et al.* Treatment Indications for Antidepressants Prescribed in Primary Care in Quebec, Canada, 2006-2015. *Jama* **315**, 2230 (2016).

115. Lippi, G. *et al.* Relation Between Red Blood Cell Distribution Width and Inflammatory Biomarkers in a Large Cohort of Unselected Outpatients. *Arch. Pathol. Lab. Med.* **133**, 628–632 (2009).

116. Horta-Baas, G. & Romero-Figueroa, M. del S. Clinical utility of red blood cell distribution width in inflammatory and non-inflammatory joint diseases. *Int. J. Rheum. Dis.* **22**, 47–54 (2019).

117. May, H. T. *et al.* Abstract 11979: Red Cell Distribution Width and Depression Among Patients Undergoing Angiography. *Circulation* **128**, A11979–A11979 (2013).

118. Shafiee, M. *et al.* Depression and anxiety symptoms are associated with white blood cell count and red cell distribution width: A sex-stratified analysis in a population-based study. *Psychoneuroendocrinology* **84**, 101–108 (2017).

119. Lee, J.-M. *et al.* Association between Mean Corpuscular Hemoglobin Concentration and

Future Depressive Symptoms in Women. *Tohoku J. Exp. Med.* **241**, 209–217 (2017).

120. Vulser, H. *et al.* Depression, antidepressants and low hemoglobin level in the Paris Prospective Study III: A cross-sectional analysis. *Prev. Med. (Baltim).* **135**, 106050 (2020).

121. Xu, H. *et al.* MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Informatics Assoc.* **17**, 19–24 (2010).

122. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* (2018). doi:10.1038/s41588-018-0183-z

123. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).

124. Janssens, A. C. J. W. Validity of polygenic risk scores: are we measuring what we think we are? *Human Molecular Genetics* (2019). doi:10.1093/hmg/ddz205

125. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* (2018). doi:10.1097/YPG.0000000000000206

126. Troubat, R. *et al.* Neuroinflammation and depression: A review. *Eur. J. Neurosci.* 1–21 (2020). doi:10.1111/ejn.14720

127. Raison, C. L., Capuron, L. & Miller, A. H. Cytokines Sing the Blues: Inflammation and the Pathogenesis of Depression. *Trends Immunol.* **27**, 24–31 (2006).

128. Olff, M. Stress, depression and immunity: the role of defense and coping styles. *Psychiatry Res.* **85**, 7–15 (1999).

129. Bao, A. M. & Swaab, D. F. The human hypothalamus in mood disorders: The HPA axis in the center. *IBRO Reports* **6**, 45–53 (2019).

130. Finnell, J. E. & Wood, S. K. Neuroinflammation at the interface of depression and cardiovascular disease: Evidence from rodent models of social stress. *Neurobiology of Stress* (2016). doi:10.1016/j.ynstr.2016.04.001

131. Hurley, L. L. & Tizabi, Y. Neuroinflammation, neurodegeneration, and depression. *Neurotoxicity Research* (2013). doi:10.1007/s12640-012-9348-1

132. Sofer, T. *et al.* A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet. Epidemiol.* (2019). doi:10.1002/gepi.22188

APPENDIX A

QUALITYLAB PIPELINE METHOD

Methods

We extracted data on all lab tests collected in the routine clinical care of 1,521,125 VUMC patients, amounting to 275,991,157 observations across 11,061 lab tests (Figure 1). Of these lab tests, 5,028 were reported in non-numeric values and 1,618 had only been administered to one patient, leaving 4,415 quantitative lab tests for further cleaning. Some lab tests had observations recorded in different units (e.g., Selenium reported in both mcg/L and ug/L), thus we restricted to lab tests for which at least 70% of the observations were measured in the same unit and required that each lab have at least 100 patients and at least 1,000 numeric observations, for a total of 939 labs retained for further analysis.

For each of these 939 labs, we applied lab-specific quality control filters (Figure 2). First, we filtered infinite and non-numeric values, as well as observations outside of 4 standard deviations from the overall sample mean, indicative of biologically implausible values due to technical or recording errors, monogenic disorders, or extreme environmental influence. We calculated the median lab value for each patient and extracted the patient's age at median lab value. For patients in whom we had to calculate the median lab value (e.g., those with an even number of observations), we defined the age at median lab value as the mid-point of the patient's ages at the two lab values used to calculate the median lab value.

We applied QualityLab to a dataset constructed from pediatric and adult observations, in both sexes, and in patients of all races (Figure 2).

The QualityLab pipeline also provides user with the option to stratify data (Figure 1b), by age at observation, sex, and EHR-recorded race, for a total of 72 different data subsets. The QualityLab pipeline generates summary statistics and plots for each strata (e.g., mean, maximum, and minimum of the median lab value; Table 1, Figure 3), and returns two versions of the data for downstream analyses. The first is a table of median lab values and age at median lab value for each individual. The second is an inverse normal quantile transformation (INT) of the median lab value data, to account for skewness and non-normality[90,132]. Importantly, the choice of quality control thresholds is completely in the control of the user. The choices made here reflect the goals of this study which focus on the central tendencies of large populations. However, the outlier thresholds and normalization methods employed here would not be appropriate in a study of rare, potentially pathogenic, variation where large genetic effects and extreme phenotypes may be expected.

Results

A total of 94,474 BioVU patients with clean lab data, of whom 66,732 were also of European genetic ancestry were included in the PGS LabWAS analyses (Figure 1). These 66,732 patients had data on 939 labs, containing 30,421,498 observations. The median number of unique lab tests per patient was 44, and the median number of lab observations per patient was 201. Slightly more than half of the BioVU patients in the sample were female (55.6%), and the average median age across the EHR was 52.0 years. In the African ancestry sample, 12,383 patients had data on 925 labs, containing 5,367,062 observations. More than half the patients were female (61.6%) and the average median age was 38.5 years. The median number of

unique lab tests per patient was 41, and the median number of lab observations per patient was 150 (Table 2).

Figure 1. Selection of VUMC patients and datasets for different analyses. VUMC patients were selected in parallel for lab test cleaning and for genotyping.

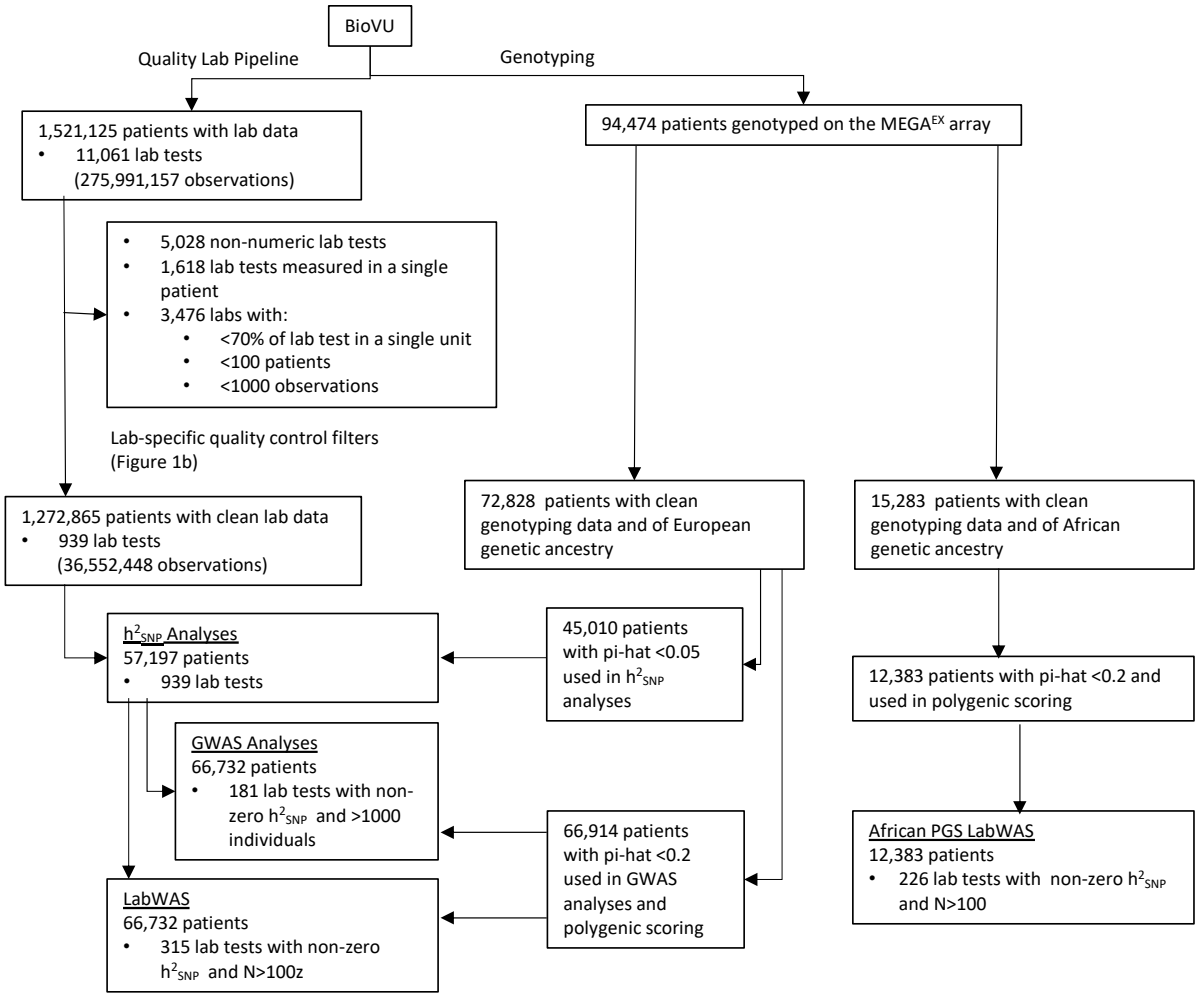


Figure 2. Lab-specific quality control filters and subsetting were applied to the 939 lab tests in the 94,474 patients with clean lab data. Parallelograms denote input and output datasets. Options highlighted in green were selected for the proof-of-principle analyses of lipids.

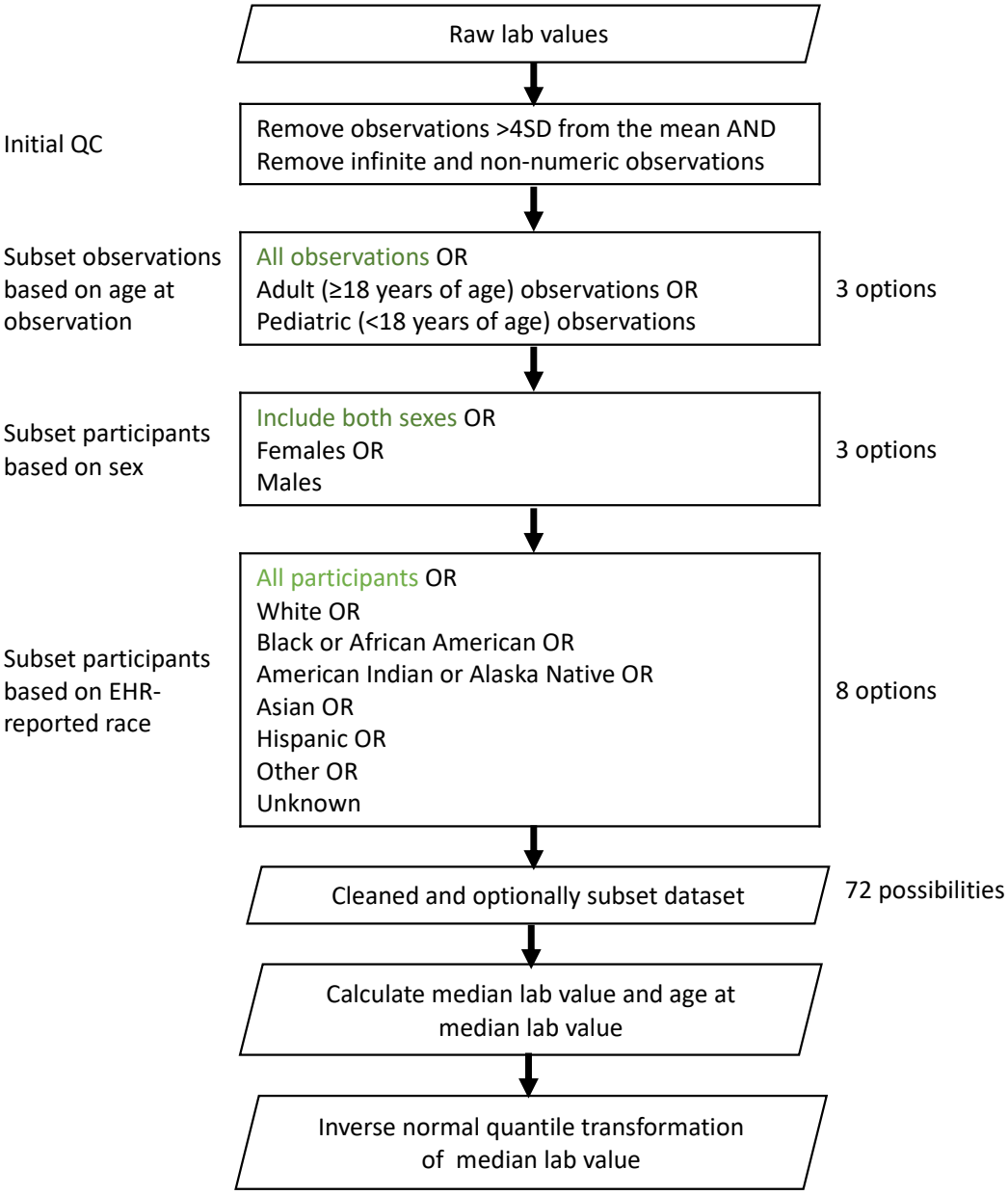
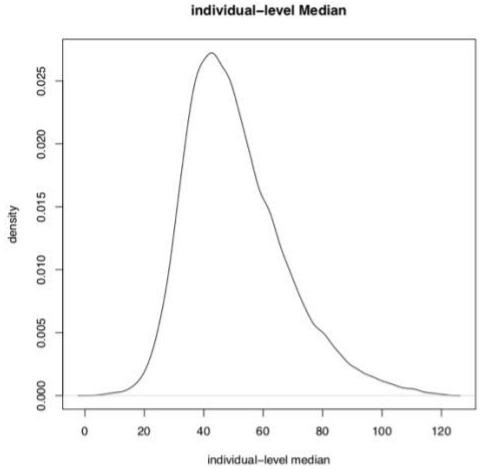
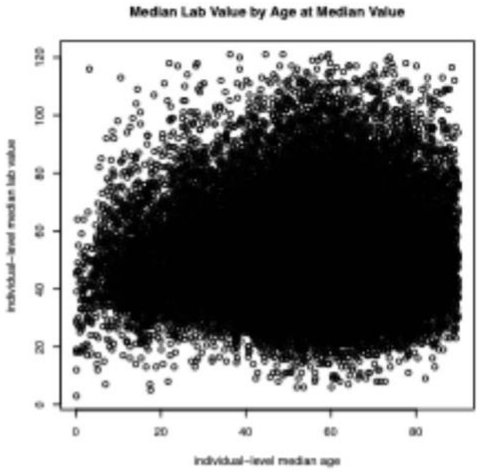
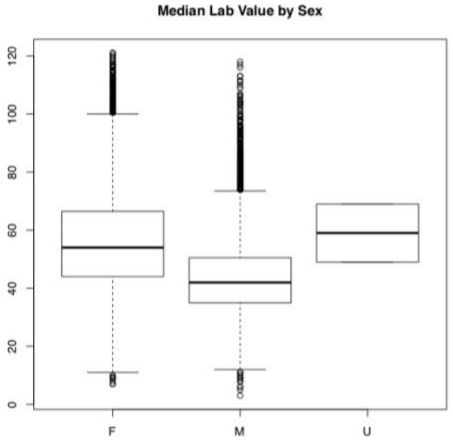
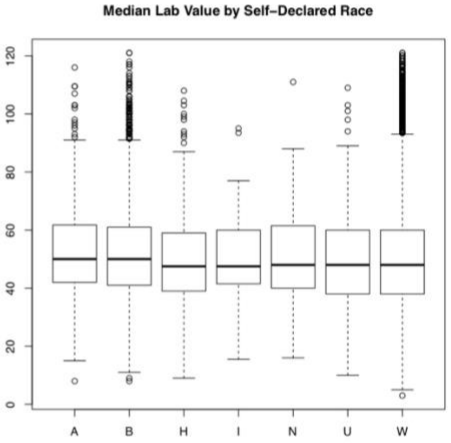


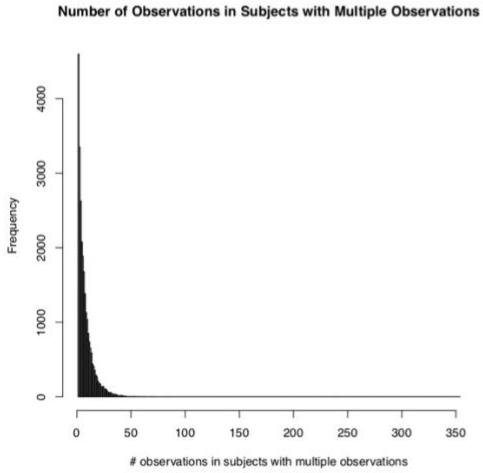
Figure 3. Data visualizations are generated by default by the QualityLab pipeline. Pictured are the visualizations for HDL, measured in mg/dL, in 70,639 patients with clean HDL lab values identified by the QualityLab pipeline.

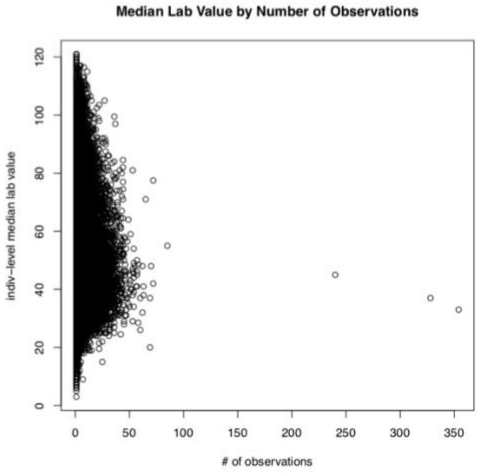


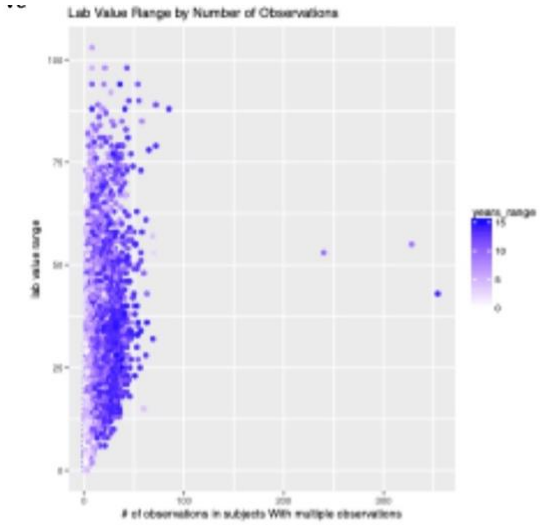


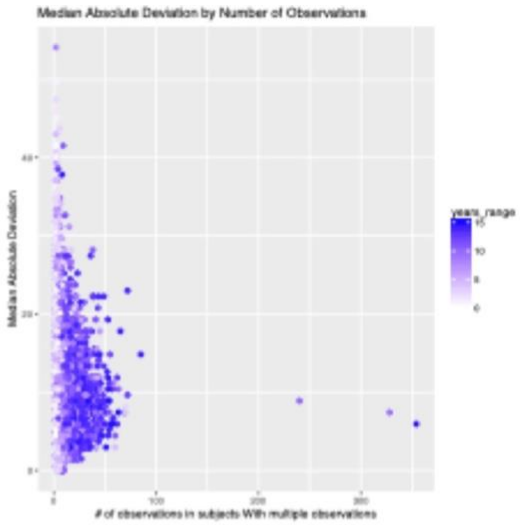


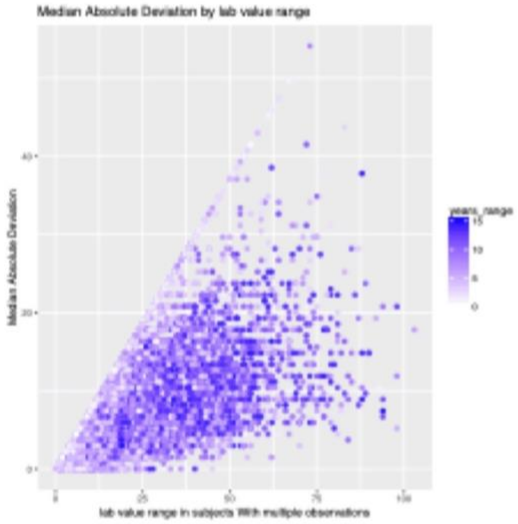


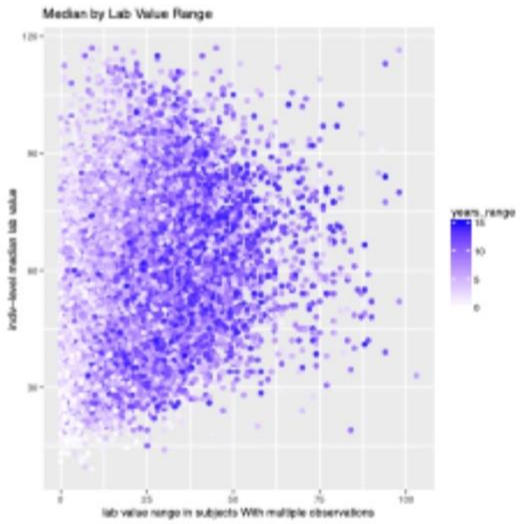





Lab Value Range by Number of Observations
lab value range
# of observations in subjects With multiple observations
years_range


Median Absolute Deviation by Number of Observations
Median Absolute Deviation
# of observations in subjects With multiple observations
years_range


Median Absolute Deviation by lab value range
Median Absolute Deviation
lab value range in subjects With multiple observations
years_range


Median by Lab Value Range
indiv-level median lab value
lab value range in subjects With multiple observations
years_range

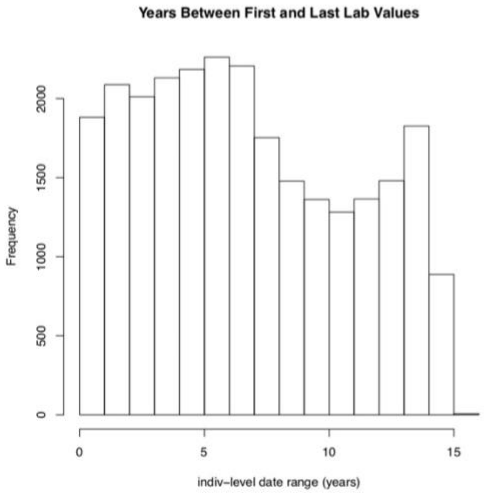

Years Between First and Last Lab Values
Frequency
0
500
1000
1500
2000
0
5
10
15
indiv-level date range (years)

Table 1. Example from lipids of summary statistics calculated by QualityLab.

Statistic	HDL	LDL	TG
N pre-filter	1,092,730	975,828	1,180,987
N post-filter	1,092,724	975,823	1,180,971
sample count all	309,856	288,545	321,942
sample count adult	279,896	259,821	285,763
sample count ped	32,835	31,385	39,117
sample count male	145,649	134,813	152,621
sample count female	164,195	153,725	169,310
sample count adult male	131,073	120,902	134,636
sample count adult female	148,811	138,912	151,116
sample count ped male	15,797	15,033	19,251
sample count ped female	17,038	16,352	19,866
min median age	0.003	0.003	0.003
max median age	90	89.99	90
mean n obs per indiv	3.52	3.37	3.65
sd n obs per indiv	5.02	4.45	5.55
max n obs per indiv	496	340	494
mean n multi obs per indiv	5.94	5.67	6.15
sd n multi obs per indiv	6.13	5.32	6.85
mean date range (years)	5.66	5.63	5.41
sd date range (years)	4.49	4.37	4.48
max date range (years)	17.48	17.48	17.48
mean value range	15.04	43.57	107.41
sd value range	11.5	33.61	110.47
min value range	0	0	0
max value range	108	247	806
mean mad	5.71	16.18	38.27
sd mad	4.65	13.22	40.93
min mad	0	0	0
max mad	74.87	154.19	524.1
mean of medians	50.95	102.77	134.14
median of medians	48.5	101	111
sd of medians	16.42	32.83	87.18
min of medians	-8	-25	0
max of medians	121	253	826
mean of means	51.03	103.1	136.56
median of mean	48.71	101	113
sd of mean	16.38	32.43	87.98
median quantile 1	21	35	34.5
median quantile 5	29	53	48
median quantile 95	82	159.5	300
median quantile 99	99	190.668	471

Table 2. Characteristics of VUMC individuals used in validation analyses of QualityLab.

Characteristic	**European Ancestry**	**African Ancestry**
Women, N(%)	36,940 (55.6%)	7,550 (61.1%)
Median age in years across the EHR, mean (sd)	52 (22.3)	39 (21)
<18 years of age at last visit, N (%)	7,395 (11.1%)	1,777 (14.4%)
CAD cases, N(%)	10,015 (15.1%)	744 (6.0%)
CAD controls, N(%)	49,702 (74.9%)	10,635 (8.6%)
Length of EHR in years, median (min-max)	7.9 (0-28.3)	6.8 (0-28.3)
Total number of unique laboratory tests	939	925
Total number of laboratory observations	30,421,498	5,367,062
Number of unique laboratory tests per patient, median (min-max)	44 (1-250)	41 (1-240)
Number of laboratory observations per patient, median (min-max)	201 (1-14,163)	150 (1-15,471)